ACCSCIENCE PUBLISHING

# Artificial Intelligence in Health

ISSN: 3041-0894 (Print)
ISSN: 3029-2387 (Online)
Volume 3 · Issue 1
January 2026

ACCSCIENCE PUBLISHING

# Artificial Intelligence in Health

**Print ISSN: 3041-0894**
**Online ISSN: 3029-2387**

***Artificial Intelligence in Health*** aims to provide a freely accessible multidisciplinary and comprehensive platform for researchers, scientists, and AI in health and medicine sciences practitioners to publish and exchange cutting-edge advancements, insights, technological development and innovations at the intersection of artificial intelligence (AI) and health. The journal seeks to explore the transformative potential of AI in improving and understanding health and medicine research outcomes, enhancing clinical decision-making, optimizing resource allocation, and addressing various challenges in the multidisciplinary field of health.


## About the Publisher

AccScience Publishing is a publishing company based in Singapore. We publish a range of high-quality, open-access, peer-reviewed journals and books from a broad spectrum of disciplines.

Contact Us

**Managing Editor**
aih.office@accscience.sg

AccScience Publishing
9 Raffles Place, Republic Plaza 1 #06-00 Singapore 048619.

Volume 3 • Issue 1 • January 2026
ISSN 3041-0894 (print) ISSN 3029-2387 (online)

# ARTIFICIAL INTELLIGENCE IN HEALTH



ACCSCIENCE PUBLISHING
Access Science Without Barriers

*Artificial Intelligence in Health*

ISSN: 3041-0894 (print)
ISSN: 3029-2387 (online)

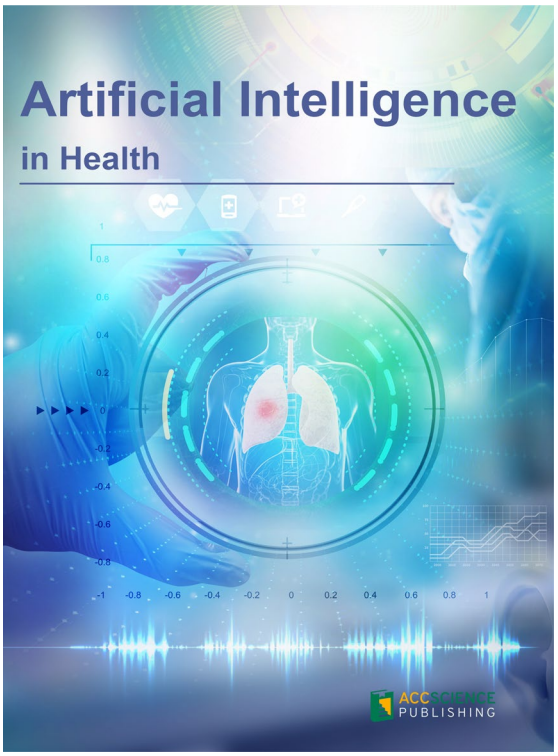
**Editorial and Production Credits**

Publisher: AccScience Publishing
Managing Editor: Freda Wang
Production Editor: Sharmila Velapasamy
Article Layout and Typeset: Sinjore Technologies (India)

For all advertising queries, contact aih.office@accscience.sg.

**Supplementary file**

Supplementary files of articles can be obtained at https://accscience.com/journal/AIH/3/1.




**Disclaimer**

AccScience Publishing is not liable to the statements, perspectives, and opinions contained in the publications. The appearance of advertisements in the journal shall not be construed as a warranty, endorsement, or approval of the products or services advertised and/or the safety thereof. AccScience Publishing disclaims responsibility for any injury to persons or property resulting from any ideas or products referred to in the publications or advertisements. AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# Artificial Intelligence in Health

Editorial Board



*Editorial Board Members as of January 14, 2026

ACCSCIENCE PUBLISHING

Volume 3 | Issue 1 | January 2026 | Page 1 – 176 |

**Artificial Intelligence in Health**

https://accscience.com/journal/AIH

# CONTENTS

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

# Advances in three-dimensional bioprinting and artificial intelligence for enhanced tumor modeling: Current progress and future perspectives

**Jia Weng[1,2], Bincan Deng[3], Xia Huang[4], and Xuan Xue[1,2]***

[1]Department of Chemistry and Materials Science, School of Science, Xi'an Jiaotong–Liverpool University, Suzhou, Jiangsu, China

[2]Department of Chemistry, School of Physical Sciences, University of Liverpool, Liverpool, United Kingdom

[3]Department of Foundational Mathematics, School of Mathematics and Physics, Xi'an Jiaotong–Liverpool University, Suzhou, Jiangsu, China

[4]Department of Biological Sciences and Bioinformatics, School of Science, Xi'an Jiaotong–Liverpool University, Suzhou, Jiangsu, China

## Abstract

Over the past decade, the global increase in cancer prevalence and cancer-related mortality has fueled extensive research to enhance the effectiveness of cancer treatments. Such efforts include the fabrication of lab-grown tissues and organs for transplantation, and the development of *in vitro* models for cancer drug testing and screening. Notably, three-dimensional (3D) tissue models offer advantages over two-dimensional cultures and have benefited from recent advancements in cutting-edge techniques like 3D printing, enabling the reconstruction of various tumor models *in vitro*. In this review, we focus on recent progress in *in vitro* 3D tumor models, with particular emphasis on the roles of 3D bioprinting and artificial intelligence. Furthermore, we provide future perspectives on employing bioprinting to develop tumor models that accurately mimic the complexity and heterogeneity of real tumor microenvironments.

***Keywords:*** Tumor model; Three-dimensional bioprinting; Bioink; Artificial intelligence-powered clinical diagnostic aids

***Corresponding author:**
Xuan Xue
(xuan.xue@xjtlu.edu.cn)

**Citation:** Weng J, Deng B, Huang X, Xue X. Advances in three-dimensional bioprinting and artificial intelligence for enhanced tumor modeling: Current progress and future perspectives. *Artif Intell Health*. 2026;3(1):1-17. doi: 10.36922/AIH025230052

**Received:** June 5, 2025

**Revised:** July 18, 2025

**Accepted:** July 29, 2025

**Published online:** August 11, 2025



## 1. Introduction

As one of the deadliest diseases in the world, cancer is a significant challenge in the medical field. In 2020, the International Agency for Research on Cancer reported that cancer is one of the top causes of death globally.[1] The American Cancer Society's latest estimates for 2024 reveal the magnitude of the challenge, with an expectation of 2,001,140 new cancer cases and 611,720 deaths in 2024.[2] Hence, it is crucial to improve cancer prevention, screening, scientific research, and innovation, particularly for malignant digestive tract tumors. However, developing anticancer drugs is a complex and challenging process,[3] and the outcome of late-stage clinical trials is often uncertain.

Therefore, improving cancer prevention and treatment outcomes requires the development of models that can accurately simulate tumor complexity and heterogeneity. Such models are expected to provide us with a deeper understanding of cancer, support drug development and treatment, and ultimately contribute to more success in the fight against cancer.

## 2. Recent progress in three-dimensional tumor models

In living organisms, cells reside within a three-dimensional (3D) environment, interacting with neighboring cells and the extracellular matrix (ECM). This milieu regulates vital life processes such as proliferation, differentiation, migration, receptor expression regulation, gene transcription and translation, and programmed cell apoptosis. Despite this, two-dimensional (2D) cell and animal models[4] are unable to simulate complex *in vivo* processes in the laboratory. While 2D cultures are easy to control, they do not accurately represent the 3D growth environment and cellular diversity found[5] *in vivo*, which can substantially alter cell behavior. Animal models, though commonly used as experimental substitutes for human studies due to shared physiological and pathological features, involve lengthy and costly experimental procedures. Moreover, the inherent biological differences limit the direct translatability of findings to human conditions. In addition, animal testing raises ethical concerns and hinders the development of new drugs. Nowadays, researchers are increasingly using 3D models to more accurately simulate human cellular and tumor behavior, addressing limitations in existing models. Such models create 3D structures and microenvironments that simulate real-life conditions, enhancing their reliability for tumor biology studies.[6,7] Using 3D cell culture, researchers can choose from various cell types and scaffold materials that provide physical and biochemical support tailored to experimental needs.[8] 3D culture models offer benefits such as accurately simulating different types of tumors at various stages, serving as powerful tools for cancer research and drug screening.[9] Figure 1 represents the progression in terms of increasing complexity to recapitulate the tumor microenvironment (TME) from 2D cell cultures to bioprinted tumor constructs.

These cell models can be broadly categorized into: (i) co-culture systems that combine multiple cell types, (ii) spheroids and organoids formed through cell self-assembly, and (iii) scaffold-supported structures that utilize ECM-mimetic biopolymeric scaffolds to recreate structural and biochemical cues of the native microenvironment. Table 1 compares the standard tumor models, including animal models and 2D versus 3D cultures, across parameters such as modeling ease, survival rates, build

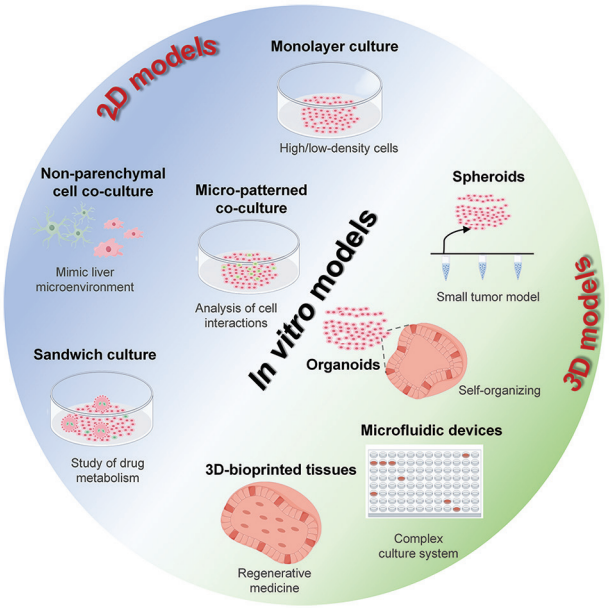



**Figure 1.** Various cancer models are used in *in vitro* constructs. The figure illustrates the evolution of cell culture models from simple two-dimensional (2D) models to complex three-dimensional (3D) models. Traditional 2D monolayer culture, monolayer co-culture, floating membrane growing cells, and sandwiched monolayer cells are the most common 2D tumor models used in research and drug screening. Cancer cells grown in 3D sphere culture, organoid culture, cancer/matrix cells grown in microfluidic devices, and advanced bioprinted structures are among the 3D cancer models available. Image created by the authors.

time, cost-effectiveness, and biological relevance. Malhão *et al.*[10] successfully inoculated and cultured four human breast cell lines (MCF7, MCF12A, MDA-MB-231, and SKBR3) in multicellular aggregates using a non-stick, low-adhesion culture plate. Watters *et al.*[11] established an organotypic model of the TME of ovarian cancer by co-culturing human mesothelial cells, fibroblasts, and ECM (collagen or fibronectin) and introducing them into ovarian cancer cells. In 2009, Sato *et al.*[12] pioneered experiments with organoid cultures. They cultured small intestinal tissue in a matrix with certain growth factors and unexpectedly discovered approximately six Lgr5 stem cells at the base of the small intestine. Two years after that, a study by Spence *et al.*[13] showed that pluripotent stem cells or embryonic stem cells can be used to create human gut-like organs by differentiating into endoderm and hindgut lineages under defined conditions. From existing models, it is evident that different research teams have made breakthroughs in constructing organoids and TME models using various cell types, and demonstrated the potential of stem cell differentiation and microenvironment simulation in regenerative medicine and tumor research. Designing specific niches for each cell type is critical to avoid the failure of human cancer treatments.

## 3. Three-dimensional bioprinting for tumor modeling

### 3.1. Three-dimensional bioprinting

Three-dimensional bioprinting is a revolutionary technology that has emerged as a result of the rapid advancement of 3D additive manufacturing technology and its integration with cells, growth factors, and biomaterials. This technology aims to create biomedical components that closely mimic the properties of natural tissues,[14] opening up unprecedented possibilities in the medical field. 3D bioprinting technology uses computer-aided design to combine cutting-edge technologies from several fields, such as mechanical engineering, materials science, cell biology, and biochemical support. The method typically uses computers, 3D modeling software, polymer materials, and 3D printers. The workflow of 3D bioprinting is shown in Figure 2. These diverse technological factors come together to determine the specific capabilities of 3D bioprinting, showcasing its significant potential and wide-ranging development opportunities.

At present, no universal 3D-bioprinting technique exists that can satisfy all the requirements for ideal tissue fabrication or accommodate every tissue type. In fact, 3D bioprinting technologies have diversified into distinct categories based on their forming principles and printing materials, such as inkjet, laser direct writing, extrusion, and light-curing printing. These technologies have unique advantages and features, contributing to the broader application of 3D bioprinting in the biomedical field. A detailed comparison of these technologies based on critical indicators such as print resolution, print speed,

**Table 1. Comparative analysis of commonly used tumor models**

| Tumor model | Modeling difficulty | Cell survival | Build time | Costs | High-throughput drug screening |
|---|---|---|---|---|---|
| Two-dimensional cell model | Easy | Moderate | Short | Low | Suitable |
| Animal model | Difficult | Moderate | Long | High | Unsuitable |
| Spheroid | Easy | Low | Short | Moderate | Suitable |
| Organoid | Easy | Moderate | Short | Moderate | Moderate |
| Three-dimensional bioprinting | Moderate | Low | Short | High | Moderate |
| Microfluidic chip | Moderate | High | Short | High | Suitable |

Notes: Modeling difficulty refers to the technical complexity in constructing the model, cell survival refers to the viability of cells during/after modeling, build time refers to the approximate time required to establish the model, cost is the relative financial cost, and high-throughput screening refers to the compatibility with automated, large-scale drug testing platforms.

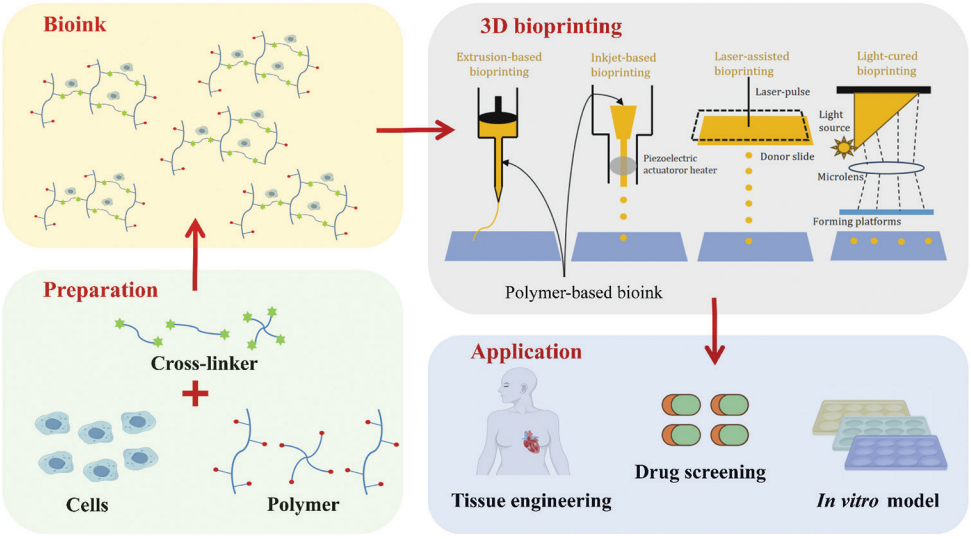



**Figure 2.** A common workflow of three-dimensional (3D) bioprinting processes. Modeling and designing 3D printable structural objects using computer-aided design software. Before 3D bioprinting, designs could be easily corrected and changed. After preparation, the cells are mixed with the material and the crosslinker, and the bioink is configured. The most suitable printer has to be selected for printing according to different properties and applications. Bioprinted scaffolds can be applied to tissue engineering, drug screening, and *in vitro* culture. The components of cells and the human body were drawn using FigDraw.

cell viability, and material usage is presented in Table 2. Table 2 shows that different 3D bioprinting technologies exhibit significant differences in various metrics. For example, inkjet printing may have better resolution and higher speed compared to other methods, while laser direct writing printing may be superior for cell viability and material efficiency. Extruded and light-cured printing also have unique characteristics and application scenarios. These differences allow researchers to choose the most appropriate 3D bioprinting technology for their needs, promoting advances in biomedical research and innovations in clinical treatments. Moreover, ongoing technological progress continues to yield more efficient, accurate, and reliable 3D bioprinting technologies, which may significantly contribute to human health.

**Table 2. Attributes of three-dimensional bioprinting technologies**

| Causality printing method | Resolution | Printing speed | Cell viability (%) | Material usage |
|---|---|---|---|---|
| Inkjet-based bioprinting | High | Fast | ≥85 | High |
| Laser direct writing bioprinting | High | Middle | ≥95 | Low |
| Extrusion-based bioprinting | Middle | Slow | 40–80 | High |
| Light-cure bioprinting | High | Slow | ≥85 | Medium |

Similar to office printers, inkjet-based 3D bioprinting uses piezoelectric or thermally driven printheads to precisely eject bioink from an ink cartridge into tiny droplets. These tiny droplets are deposited in layers onto a substrate to construct complex 3D structures of living tissues with fine spatial control[15] (Figure 3A). Inkjet printing technology is well-regarded for its high-resolution capability.[16] Park *et al.*[17] demonstrated precise localization of alveolar cells in a cell culture matrix using inkjet printing technology. A 10-micron-thick model of a three-layer alveolar barrier was created to mimic the structure, morphology, and function of actual lung tissue. In addition, inkjet 3D bioprinters are typically equipped with multiple printheads to print different bioinks simultaneously, which accelerates the printing process.[18-23] However, inkjet printing has limitations due to the low driving pressure of the printheads, which hampers the handling of highly viscous materials and concentrated bioinks. In addition, thermally driven printheads can generate heat when

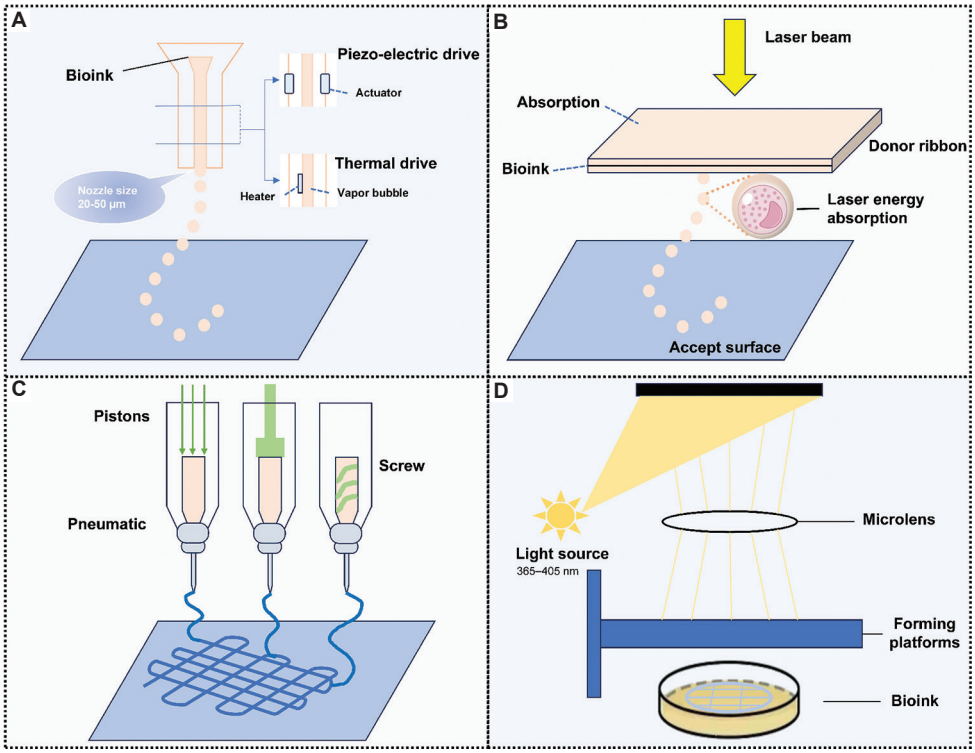



**Figure 3.** Three-dimensional bioprinting technologies. (A) Droplet-based inkjet bioprinting driven by piezoelectric or thermal actuation. (B) Laser-assisted bioprinting uses a laser beam to transfer bioink from a donor ribbon to an acceptor surface through laser energy absorption. (C) Extrusion-based bioprinting, driven by pneumatic, piston, or screw mechanisms, is suitable for high-viscosity bioinks. (D) Light-assisted bioprinting (e.g., digital light processing) that utilizes patterned light (365–405 nm) through microlens arrays to selectively solidify photosensitive bioinks. Image created by the authors.

printing, which may harm cells, limiting the widespread use of inkjet 3D bioprinting.

Laser direct writing technology was first used to make metal stencils.[24-42] However, in the early 21st century, Odde and Renn[43] first introduced live cell printing using laser direct writing 3D bioprinting technology, which greatly contributed to its rapid development. In this process, bioink is uniformly deposited on a layer that absorbs it. A high-energy laser beam is then used to penetrate the glass substrate and sinter or cure the biomaterial layer by layer in a controlled manner, allowing it to be precisely deposited onto a forming platform[44] (Figure 3B). It is worth noting that laser direct writing technology uses a nozzle-less inkjet printing method. This feature avoids direct contact between the bioink and the processing device, significantly reducing cell mechanical damage. As a result, the cells can maintain a high activity level during the printing process. In addition, laser writing is suitable for printing thick biomaterials, increasing its potential uses.[45]

Extruded 3D bioprinting technology, the most widely used method, is particularly well-suited for printing high-density cells or thick biomaterials,[46] functioning in a manner similar to fusion printing technology.[47] The bioink is extruded in a controlled manner through precise control of the air pressure, piston, or screw. The nozzle's ability to move in the X-Y-Z direction ensures that the bioink is deposited accurately, creating the desired intricate pattern (Figure 3C). This technology has the unique ability to continuously extrude and form continuous fiber-like structures. This feature allows 3D bioprinting[39,42,48-56] technology to effectively print polymers with different viscosities and varying cell concentrations, making it suitable for a wide range of applications. This technology can be used to create structurally robust and morphologically diverse biological tissue models.

In the 1980s, light-curing technology emerged as an advanced process in 3D printing, primarily using photosensitive resins as the principal printing materials.[57] With ongoing technological progress, the potential advantages of applying light-curing technology in bioprinting have begun to be explored innovatively. There are two types of light-curing printing technology: stereolithography and digital light processing (DLP)[29,32,57-65] (Figure 3D). Stereolithography is a special 3D printing method that uses light to selectively solidify bioinks and create precise structures. The material is cured layer by layer, gradually accumulating and stacking to form the desired 3D scaffold.[66] DLP[67] uses ultraviolet or visible light through the digital micromirror device to form 2D shapes, constructing them into 3D structures. Although both stereolithography and DLP share similar principles, they differ in their technical implementation. Compared to other bioprinting methods, light-curing devices are simpler and easier to control. However, the use of ultraviolet light and photoinitiators still poses challenges, as they can potentially damage cells. As research continues and technology advances, light-curing technology is anticipated to have more applications in bioprinting.

Three-dimensional bioprinting technologies have demonstrated distinct advantages in constructing tumor models, and the choice of printing modalities should be closely aligned with the biological characteristics of specific tumor types. Extrusion-based bioprinting, due to its high cell-loading capacity (>$10^7$ cells/mL) and mechanical stability, is particularly suitable for fabricating matrix-dense tumors such as cholangiocarcinoma and pancreatic cancer. Using this approach, Mao *et al.*[68] successfully recapitulated the fibrotic microenvironment of patient-derived cholangiocarcinoma xenograft, achieving a gemcitabine resistance prediction accuracy of 89% ($p$<0.01). This success was primarily attributed to the precise simulation of the ECM barrier effect.[68]

For tumors that require accurate vascular network simulation, such as gliomas and metastatic breast cancer, the microscale resolution (~20 μm) of light-assisted bioprinting (e.g., DLP) is essential. Peng *et al.*[69] constructed a vascularized glioma model using DLP technology, in which human umbilical vein endothelial cell-patterned vasculature increased cancer cell migration distance by 3.2-fold, significantly enhancing the model's value for studying blood–brain barrier penetration.[69]

In high-throughput drug screening scenarios, inkjet-based bioprinting facilitates the rapid production of tumor microarrays (>200 models/h), thereby accelerating treatment strategy optimization. For example, Chen *et al.*[70] used inkjet bioprinting to construct a hepatocellular carcinoma model that successfully demonstrated the synergistic effect of sorafenib and radiotherapy, increasing the accuracy of personalized treatment prediction to 82%.[70]

Notably, the functional integration of key components of the TME, such as immune cell infiltration and hypoxic gradients, is becoming a central focus in tumor bioprinting. Cui *et al.*[71] embedded chimeric antigen receptor-T cells into a glioblastoma model, resulting in a 67% increase in T-cell infiltration depth compared to conventional models, offering a more precise platform for immunotherapy research.[71]

These advancements highlight a paradigm shift in bioprinting technology, from generalized tissue fabrication to tumor-specific customized models, with early translational potential demonstrated in patient-derived

organoid- or patient-derived xenograft-based drug sensitivity testing.

### 3.2. Bioink

Bioink, the core material for 3D bioprinting, is a complex system that includes scaffolding materials, cells, and a variety of biological and chemical factors, as well as crosslinking agents to ensure a smooth printing process. The term was first introduced in a 2003 article on organ printing[72] alongside the concept of biopaper. Initially, bioinks were used in 3D cultures to support cellular components in hydrogels. With advances in bioprinting technology, the incorporation of cellular components such as nuclei and cell clusters has become increasingly important in bioink formulations. Unlike traditional 3D printing materials, bioinks used in 3D bioprinting are typically designed to be biocompatible and often bioactive, with mechanical properties tailored to the target tissue, ranging from flexibility for soft tissue to rigidity for bone applications.[73] Bioinks support cell viability and function, maintain structural fidelity during the printing process, and promote the formation of functional biological constructs. However, an ideal bioink that meets all biological, mechanical, and printability requirements has yet to be fully realized. Researchers must comprehensively consider various factors when developing bioinks, such as the performance of 3D bioprinters (including printing speed, extrusion pressure, and printing temperature), biocompatibility support for cellular activity and growth, degradability, and cell adhesion. Collectively, these constitute the complex challenges of bioink development and serve as the driving force in the field.[74] Hospodiuk *et al.*[75] conducted the first comprehensive review comparing the properties, advantages, and disadvantages of various bioink materials, where they discussed the current limitations of bioink materials and provided insights into future directions, offering a comprehensive view of the present and potential future of bioinks.

Natural hydrogels, derived from plant or animal origins, are highly hydrophilic materials composed of 3D polymer networks crosslinked through physical or chemical bonds and contain a high proportion of water. The unique structure and composition of natural hydrogel make it suitable for a wide range of applications. Type I collagen, renowned for its biocompatibility, has become a widely used material in 3D bioprinting, leading to new possibilities in biomedical engineering.[74] Gelatin is a natural substance made from proteins found in animal skin and bones[76] and is essential in pharmaceutical manufacturing and various industrial applications.[77] Fibrin hydrogel is activated by thrombin to initiate the fibrinogen process, leading to the bonding of fibrin molecules and the formation of a stable fibrin clot.[78]

These commonly used natural hydrogels are highly appreciated for their superior bioactivity. They mimic the ECM microenvironment of body tissues, providing ideal conditions for the growth of printed cells and facilitating cellular communication. In 3D bioprinting, the fluidity of bioinks is critical, as they must solidify and retain their shape rapidly after printing. Therefore, the malleability of bioinks is a key element in successful printing. However, natural hydrogels have limitations under physiological conditions, such as reduced stability and inferior mechanical properties (e.g., strength and toughness) compared to synthetic hydrogels. Therefore, natural hydrogels may not meet the requirements for certain applications. Over the years, research into synthetic hydrogels has increased.[79] Synthetic hydrogels have significantly improved the stability of bioinks by combining physical, chemical, mechanical, and physiological support and incorporating other components. In addition, the mechanical properties of synthetic hydrogels can be adjusted to ensure suitability for 3D printing, providing a broader perspective for 3D bioprinting.[80]

Gelatin methacryloyl has a high mechanical strength and low swelling rate, making it suitable for blending with other hydrogels to enhance cell survival. These properties position Gelatin methacryloyl as a promising biomaterial for tissue engineering.[81,82] Polyethylene glycol-based hydrogels, which have been approved by the Food and Drug Administration for use in the biomedical field, can provide temporary structural support in the fabrication of complex 3D tissue-engineered structures.[74] However, synthetic hydrogels generally exhibit lower intrinsic biocompatibility compared to natural hydrogels, despite being more cost-effective and offering greater tunability in mechanical and chemical properties.[83] Therefore, the choice between natural and synthetic hydrogels can be guided by several factors, including application need, performance criteria, and cost considerations. As shown in Table 3, these considerations can facilitate making well-informed decisions to meet the varying needs of different applications.

Three-dimensional bioprinting technology is widely used in clinical medicine for the fabrication of bone and cartilage,[85] skin,[86] heart, and muscle tissue.[87] While bioprinted tissues and organs can be used to treat diseases, significant challenges in tissue engineering remain. There is an urgent need for bioinks that not only meet essential criteria – such as biocompatibility, appropriate degradation rates, and mechanical strength – but are also cost-effective.

To construct 3D tumor models suitable for different purposes, researchers need to review a wide range of existing studies and identify solutions, such as

**Table 3. Comparative summary of common natural and synthetic bioinks used in three-dimensional bioprinting[75,84]**

| Classification | Bioink | Crosslinking method (time) | Concentration | Formability | Feature resolution | Cell viability (%) | Cellular support capacity |
|---|---|---|---|---|---|---|---|
| Natural hyaluronic | Collagen type I | Thermal (one to several minutes) | 0.025–0.5% | Poor | 200–1,000 μm | 33–95 | High |
| | Gelatin | Thermal (5 min) | 7–20% | Moderate | 350–1,000 μm | 85–90 | High |
| | Fibrin | Enzymatic (seconds to 6 min) | 10–60 mg/mL or 20–50 U/mL | Good | 144–750 μm | ~74 | Moderate |
| | Agarose | Thermal (minutes to 2 h ) | 0.3% | Good | 250 μm | 90–98.8 | High |
| | Chitosan | pH-mediated (2 h) | 3% | Poor | 400–500 μm | – | Moderate |
| | Alginate | Ionic (seconds) | 0.1–8% | Moderate | 400–600 μm | 90.8–95 | Low |
| | Hyaluronic acid | Photo/pH-mediated (3–10 min) | 1.5% | Good | 200–760 μm | – | High |
| Synthetic hyaluronic | Gelatin methacryloyl | Photopolymerization (10 s–10 min) | 5–20% | Good | 150–750 μm | 63.2–97 | High |
| | Polyethylene glycol | Photopolymerization (up to 19 min) | 10–20% | Moderate | 168–550 μm | 89–90 | Moderate |
| | Pluronic® F-12 | Thermal (minutes) | 25–30% | Good | 150–600 μm | 60–91.3 | Low |

incorporating growth factors, adjusting material ratios, or using different hydrogels, to improve cell survival and phenotypes. Cutting-edge methods such as 3D bioprinting and microfluidics have been used to create effective tumor models for cultivation, with quality evaluation based on cell viability, morphology, proliferation, and differentiation.

Chen *et al.*[88] used 3D bioprinting to develop a colorectal cancer model by constructing a biological scaffold and co-culturing HCT116 human colorectal cancer cells with tumor-associated endothelial cells. The 3D scaffold effectively supported the cells and maintained physiological processes such as cell adhesion, proliferation, stemness retention, and vascular conservation. In the model, the activated stromal stem cells expressed multiple tumor-associated factors and dense ECM. The tumor tissues exhibited transcriptomic features closely resembling those found in real tumors. Sbirkov *et al.*[89] constructed a cost-effective 3D-printed Caco-2 human colon cancer model with a histological appearance that is similar to adenoid tissue. The model's RNA expression profiles were characterized by enhanced cell adhesion, hypoxia-related signatures, upregulation of genes in the epidermal growth factor receptor/Kirsten rat sarcoma viral oncogene homolog pathway, and downregulation of genes involved in cell cycle regulation. Chemotherapeutic drug testing experiments showed that the overall drug resistance of tumor cells in the 3D-printed model increased compared to 2D cultures, closely resembling the drug responsiveness of tumors *in vivo*. The researchers suggested that the platform can be expanded to include primary colorectal cancer samples, making it a promising platform for novel, individualized drug screening.

## 4. Artificial intelligence (AI) in three-dimensional bioprinting

With the advancement of novel productive forces, 3D bioprinting is encountering increasingly stringent demands. These demands include enhanced biocompatibility and mechanical properties in material selection[90] and the need to attain a high degree of precision and stability during the printing process.[91] In addition, real-time monitoring and adjustment of the printing process are necessary to ensure the quality of the final product. These challenges have driven scientists to explore novel approaches for optimizing 3D bioprinting technology, with the integration of AI opening new opportunities in the field.[92,93]

As a crucial component of modern science and technology, AI can learn and analyze vast amounts of data, extracting patterns to make predictions and optimizations.[94] This capability makes AI widely applicable in 3D bioprinting. AI technology enables intelligent optimization of material formulations,[95] real-time adjustment of printing parameters, and precise control over the final product quality,[96] thereby significantly enhancing the efficiency and quality of 3D bioprinting.

This section aims to explore the specific applications of AI technology in extrusion 3D bioprinting, with a focus on three key areas: bioink formulations, printing parameter optimization, and quality control. First, this section will discuss the role of AI in optimizing bioink formulations, then explore its application in printing parameter optimization, and finally analyze its specific use in defect detection and quality assessment (Figure 4). These discussions aim to provide valuable insights and lessons for the field of 3D bioprinting.

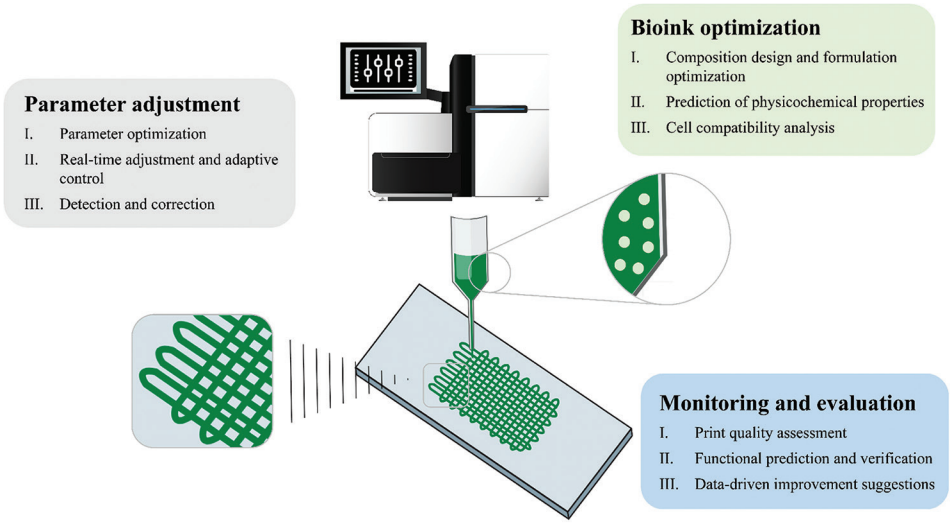



**Figure 4.** Application of artificial intelligence in three-dimensional bioprinting. Image created by the authors.

### 4.1. Bioink formulation optimization

In 3D bioprinting, the multi-dimensional nature of characteristic variables complicates the identification of suitable printing conditions. As the foundation of bioinks, the physicochemical properties of biomaterials determine the feasibility of meeting specific printing requirements.[97] However, changes in biomaterial composition and formulation can lead to highly nonlinear variations in these properties. Therefore, given the challenges posed by such complex variables, AI has emerged as a powerful tool for rapid selection and optimization of printing parameters.[98,99]

Chen *et al.*[100] assessed the printability of 210 ink formulations derived from six biomaterials and developed a printability prediction model using a machine learning (ML) algorithm with over 80% accuracy. Hashemi *et al.*[101] utilized a Bayesian optimization algorithm to create a chitosan-gelatin-agarose bioink, which exhibited good cell morphology and viability, along with optimal rheological properties, degradability, and hydrophilicity. These studies demonstrate that ML algorithms can effectively predict print suitability and optimize materials and formulations for specific printing needs. However, this approach is limited, as it only supports formulation screening under specific printing parameters and does not serve as a general modeling method for evaluating biomaterials.

### 4.2. Printing parameter optimization

In the actual printing process, both bioink and printing parameters jointly determine printability. Therefore, to achieve good printability, it is essential to optimize both simultaneously. Ruberu *et al.*[102] created an optimization algorithm using ink formulation, pressure, printing speed, platform temperature, and ink reservoir temperature as input features, with printability scored quantitatively as the objective. They employed a Bayesian optimization algorithm to iteratively refine the ink formulation and printing parameters, significantly reducing the number of experiments needed to find the optimal values. However, this method still overlooks the material itself. Rheological properties, particularly in hydrogels, such as viscosity, viscoelastic shear moduli, elastic recovery, and shear stress, significantly influence printability.[103] Thus, incorporating the physicochemical properties of materials (e.g., mechanical and rheological properties) into modeling could enable multiscale printability prediction. Lee *et al.*[104] discovered that a high elastic modulus improves shape fidelity and permits extrusion at lower critical yield stresses with the aid of ML. Through multivariate regression analyses, they derived various formulations of naturally derived bioinks that maintain high shape fidelity. This innovative approach to optimizing ink formulations from the perspective of material rheology demonstrates a comprehensive printability assessment by integrating the rheological properties of biomaterials (e.g., shear storage, loss modulus, viscosity, and shear-thinning properties)[95] into the printability evaluation model. By studying the correlation between the rheological properties of materials and printing processes, we can better understand the theoretical basis affecting bioink printability. This allows for a more accurate evaluation of factors such as extrudability, shape fidelity, and cell viability during printing. Establishing a printability evaluation model based on material property interpretability ensures that

the ML model is not just a black box but a scientifically grounded, generalized evaluation model.

### 4.3. Quality control

Throughout the 3D bioprinting process, AI intervention in automated defect detection and quality assessment is a promising approach for establishing better evaluation criteria with quantitative indexes. Conventional intelligent algorithms can effectively handle basic quality assessment tasks. For instance, the structural similarity method used by Fastowicz and Okarma[105] employs a Monte Carlo approach to randomly select regions for comparison and additional region matching, achieving reliable accuracy in classifying high- and low-quality printed samples. However, such algorithms are insufficient for more complex evaluation characteristics and control requirements. Instead, integrating AI with control systems allows real-time use of evaluation results during the 3D bioprinting process. For example, Paraskevoudis *et al.*[106] assessed the quality of fused filament fabrication 3D-printed objects during printing through AI-based computer vision. Based on the assessment, the printing process can be terminated, or the parameters related to detected defects can be adjusted. This approach is also applicable to real-time extrudability evaluation in 3D bioprinting. In addition, Jin *et al.*[92] developed an anomaly detection system using layer-by-layer sensor images and ML algorithms to distinguish and categorize defects in transparent hydrogel-based bioprinting materials, enabling real-time autonomous correction of process parameters. These methods effectively evaluate and regulate the 3D bioprinting process, leading to efficient printing decisions and high-quality tissue construction in specific target environments. Furthermore, the concept of large models also introduces generalization possibilities for defect detection and real-time regulation in 3D bioprinting. Brion and Pattinson[107] created a large and diverse dataset of extruded 3D prints based on images automatically labeled according to deviations from optimal printing parameters. They employed neural networks and control loops for real-time detection and rapid correction of various errors, demonstrating effectiveness across different 2D and 3D geometries, materials, printers, tool paths, and extrusion methods. Consequently, AI-driven computer vision techniques are increasingly capable of achieving high-quality print result detection, with deep learning networks playing a central role in defect detection and modeling the evaluation of 3D bioprinting outcomes.

## 5. AI in three-dimensional tumor models

AI technology is progressively transforming the cultivation of 3D tumor models by enhancing image recognition, feature extraction, and data analysis capabilities. These advances have significantly improved the efficiency of model evaluation and the accuracy of drug efficacy prediction. AI not only optimizes culture conditions and reduces reliance on animal experiments, but also enhances the precision of TME modeling, thereby supporting personalized drug screening.[108] For instance, Chen *et al.*[109] developed the spheroid monitoring and AI-based recognition technique, which employs convolutional neural networks to automatically identify the boundaries formed during 3D tumor spheroid culture and quantify their invasive characteristics, thereby improving the assessment of dynamic behaviors during cultivation.[109] Mali *et al.*[110] proposed an end-to-end deep learning pipeline that integrates a U-Net model with a convoluted neural network-based regression network to automatically detect and classify spheroid morphology and cell viability, achieving a prediction accuracy of up to 98%. This provides an effective tool for high-throughput drug screening.[110] These studies offer substantial practical value for optimizing and advancing 3D tumor model cultivation, laying a strong foundation for the multiscale application of AI in this field.

However, it is important to note that current applications of AI in 3D tumor models often lack continuity and integration. Future research should focus on the comprehensive evaluation of specific models such as tumor spheroids, organoids, and matrix-embedded models. This evaluation should be based on a set of indicators, including morphological structure, cell viability and proliferation capacity, cellular heterogeneity, and stemness. Furthermore, the application of multi-objective optimization algorithms may lead to the development of robust AI-based strategies for evaluating and improving 3D tumor models.

## 6. Three-dimensional bioprinting market

Bioprinting, as a revolutionary technology in the fields of regenerative medicine and tissue engineering, has made significant progress in recent years and demonstrated immense application potential. Technologically, bioprinting is rapidly evolving, with new materials and printing technologies continually emerging, making it increasingly feasible to construct more complex and functional tissues and organs. The current advances in bioprinting technology are primarily reflected in several areas. First, there has been a significant increase in the diversity of bioinks, including natural polymers, synthetic polymers, and composite materials, which more accurately mimic the characteristics of ECM *in vivo*, improving cell viability and tissue functionality.[75] Second, multi-cell type and multi-material composite tissue printing have gradually advanced, better replicating the biological complexity

of tissues, for example, the successful construction of vascularized tissues through simultaneous printing of stem cells, endothelial cells, and supporting cells.[111] In addition, researchers have constructed functional tissues, such as skin, cartilage, and small organ-like structures – by precisely controlling microenvironmental factors (e.g., mechanical strength, porosity, degradability, and the release of bioactive molecules) – which exhibit higher biocompatibility and functionality.[112] Moreover, integration of bioprinting technology with emerging technologies such as AI, nanotechnology, and microfluidics can further enhance printing precision and efficiency, accelerating the development of personalized medicine.[113]

## 7. Future perspectives

Three-dimensional cell culture technology has demonstrated significant advantages in modern biomedical research, as it can more realistically simulate the complex growth environment of cells *in vivo* than in traditional 2D cultures. In addition, 3D cell culture technology can partly replace animal experiments, reducing the need for animals and ethical concerns. This method boosts the production of cytokines, antibodies, and other vital biomolecules and improves the overall efficiency of cell culture. 3D cell culture provides a powerful tool for in-depth research on the mechanisms of tumor occurrence and development for screening drugs. Cutting-edge technologies for constructing and cultivating 3D models have garnered the attention of researchers to improve the quality of the existing models. In this context, 3D bioprinting has shown several advantages, allowing cells to grow in a 3D scaffold with controlled fine structures to improve the microenvironment for tumors. However, the existing 3D bioprinting techniques have limitations in fully replicating the ECM structures and functions. Specifically, the current challenges include: (i) difficulties in precisely depositing cells and biomaterials to simulate complex structures, leading to models that lack the physiological relevance of actual tumors, and (ii) the inability of bioinks to accurately mimic the biomechanical and biochemical properties of tumor ECM, which hinders the maintenance of cancer cell functions and interactions within the printed tissue. To address these challenges, pioneering efforts on the printing side have explored approaches such as combined coaxial[114] and multi-material printing,[115] enabling the construction of complex, multi-layered tissues that more accurately replicate physiological features in tumor models. AI is anticipated to make significant contributions to these research fields in the near future. Recently, the AI-guided bioink design has facilitated the development of engineered and composite bioinks[116,117] that fulfill specific mechanical and chemical requirements, enabling more effective fabrication of tumor models that better mimic physiological functions and interactions.

Over the past decade, bioprinting has undergone rapid development, driven by multiple factors. The increase in medical demand is one of the main drivers, especially in the fields of organ transplantation and personalized medicine. The persistent global shortage of organs makes bioprinting particularly valuable in fabricating artificial organ and tissue substitutes.[118] In addition, increased research and development investment from governments and research institutions globally and significant capital inflows have accelerated technological development and market growth.[119] Despite its vast potential, bioprinting's market growth still faces regulatory and ethical challenges. For instance, obtaining clinical application approval and certification for printed complex organs or tissues may involve lengthy approval processes. Social and ethical issues must also be considered, particularly in applications involving human cells and gene editing.[120] Beyond medical applications, bioprinting holds significant potential in non-medical fields such as drug development, cosmetic testing, and environmental science. Applications such as 3D-printed liver organoids for drug testing and artificial skin models for cosmetic testing further expand the market for bioprinting.[121]

Looking ahead, the prospects for bioprinting technology are highly promising, with continued rapid development anticipated, driven by ongoing technological innovation, growing market demand, and evolving regulatory frameworks. In terms of technological innovation, future optimizations of bioprinting will focus on achieving higher printing precision and fabricating more complex tissue structures. It will leverage new bioinks, nanomaterials, and microfluidic technologies to achieve more precise control over cells and tissues, which will more accurately simulate complex biological microenvironments.[87] The demand for personalized medicine will further promote the development of bioprinting applications, enabling more targeted treatment plans by combining patient cell and genetic data.[74] In addition, the development of smart bioink materials will significantly enhance the functionality of 3D-bioprinted tissues, such as dynamically regulating cell behavior and tissue function by gradually releasing drugs or growth factors in response to specific physiological or pathological stimuli.[122]

From a market standpoint, the application areas of bioprinting are expected to diversify further. It is anticipated that bioprinting will achieve breakthroughs in areas such as tissue and organ transplantation, new drug development, and toxicity testing, further reducing organ shortages for transplants, optimizing drug development processes, and

lowering research and development costs and time.[123] As technology advances and costs decrease, the bioprinting market is projected to maintain a strong compound annual growth rate over the next decade, especially in emerging economies like China and India, where growth is expected to be faster.[124] However, future market expansion must still address multiple challenges, including the gradual improvement of regulatory standards and the resolution of ethical issues.[125] Further development of bioprinting will depend on strengthening interdisciplinary research and building cross-industry collaborations. Integrating knowledge and methods from multiple disciplines will drive technological advancement and market expansion.[126]

As 3D bioprinting technology evolves, AI integration enables more efficient and precise tasks, such as optimizing bioink formulations, adjusting printing parameters, and enhancing quality control.[95,100-107] Although AI applications in 3D bioprinting have made significant progress, challenges remain, including high dependence on large datasets and limited model interpretability in practical use.[127,128]

To overcome these challenges, future research should focus on developing algorithmic models guided by the physicochemical properties of the printed materials. These models can better cope with the complexity and diversity of different materials and printing conditions, even with limited data. In addition, further exploration of multimodal learning, which integrates diverse data sources such as rheological properties and biological performance data, could improve accuracy and reliability. Shifting toward larger, more accessible datasets and embracing open science principles will foster global research collaboration, further advancing AI-driven 3D bioprinting technologies.[129,130] It is expected that AI is set to become a central force driving innovation in 3D bioprinting, and this incorporation will be beneficial for tumor tissue engineering and unlocking new opportunities in biomedical engineering and regenerative medicine. In parallel, AI is also reshaping the cultivation and evaluation of 3D tumor models; however, systematic research is still needed to establish standardized frameworks for model assessment, enhance consistency across applications, and fully realize the potential of AI-driven optimization in this field.

In summary, 3D bioprinting technology aims to enhance printing accuracy and construct complex tissues, advancing personalized medicine. Market applications will diversify, especially in tissue transplantation and drug development, but regulatory and ethical challenges must be addressed. AI will drive technological innovation, enabling bioprinting to create new opportunities in the fields of oncology, tissue engineering, and regenerative medicine.

## Acknowledgments

None.

## Funding

This research was supported by the Postgraduate Research Scholarship (PGRS FOS2211JM02), the Research Development Funding (RDF-22-02-002) provided by Xi'an Jiaotong-Liverpool University (Suzhou, China), the Suzhou Industrial Park High Quality Innovation Platform of Functional Molecular Materials and Devices (YZCXPT2023105), and the XJTLU Advanced Materials Research Center (AMRC).

## Conflict of interest

The authors declare that they have no conflicts of interest.

## Author contributions

*Conceptualization:* Xuan Xue
*Visualization:* Jia Weng, Bincan Deng
*Writing–original draft:* Jia Weng, Bincan Deng
*Writing–review & editing:* Xuan Xue, Xia Huang

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

The datasets used and analyzed during the present study are available from the corresponding author on reasonable request.

## References

1. Sung H, Ferlay J, Siegel RL, *et al*. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209-249.

   doi: 10.3322/caac.21660

2. Siegel RL, Giaquinto AN, Jemal A, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. 2024;74(1):12-49.

   doi: 10.3322/caac.21820

3. Dolgin E. Cancer drug approvals and setbacks in 2024. *Nat Cancer*. 2024;5(12):1756-1758.

   doi: 10.1038/s43018-024-00873-3

4. Carvalho V, Maia I, Souza A, *et al*. *In vitro* biomodels in stenotic arteries to perform blood analogues flow visualizations and measurements: A review. *Open Biomed Eng J*. 2020;14(1):87-102.

doi: 10.2174/1874120702014010087

5. Fontoura JC, Viezzer C, dos Santos FG, *et al*. Comparison of 2D and 3D cell culture models for cell growth, gene expression and drug resistance. *Mater Sci Eng C Mater Biol Appl*. 2020;107:110264.

   doi: 10.1016/j.msec.2019.110264

6. Jubelin C, Muñoz-Garcia J, Griscom L, *et al*. Three-dimensional *in vitro* culture models in oncology research. *Cell Biosci*. 2022;12(1):155.

   doi: 10.1186/s13578-022-00887-3

7. Weng J, Li S, Weng J, *et al*. Bioinspired 3D hydrogel scaffold to mimic tumor microenvironment for investigating into the anoikis resistance mechanisms in colorectal cancer. *Mater Today Bio*. 2025;33:102061.

   doi: 10.1016/j.mtbio.2025.102061

8. Abuwatfa WH, Pitt WG, Husseini GA. Scaffold-based 3D cell culture models in cancer research. *J Biomed Sci*. 2024;31(1):7.

   doi: 10.1186/s12929-024-00994-y

9. Li W, Zhou Z, Zhou X, *et al*. 3D biomimetic models to reconstitute tumor microenvironment *in vitro*: Spheroids, organoids, and tumor-on-a-chip. *Adv Healthc Mater*. 2023;12(18):e202202609.

   doi: 10.1002/adhm.202202609

10. Malhão F, Macedo A, Ramos A, Rocha E. Morphometrical, morphological, and immunocytochemical characterization of a tool for cytotoxicity research: 3D cultures of breast cell lines grown in ultra-low attachment plates. *Toxics*. 2022;10(8):415.

    doi: 10.3390/toxics10080415

11. Watters KM, Bajwa P, Kenny HA. Organotypic 3D models of the ovarian cancer tumor microenvironment. *Cancers* (*Basel*). 2018;10(8):265.

    doi: 10.3390/cancers10080265

12. Sato T, Vries RG, Snippert HJ, *et al*. Single Lgr5 stem cells build crypt-villus structures *in vitro* without a mesenchymal niche. *Nature*. 2009;459(7244):262-265.

    doi: 10.1038/nature07935

13. Spence JR, Mayhew CN, Rankin SA, *et al*. Directed differentiation of human pluripotent stem cells into intestinal tissue *in vitro*. *Nature*. 2010;470(7332):105-109.

    doi: 10.1038/nature09691

14. Zhang Z, Zhou X, Fang Y, Xiong Z, Zhang T. AI-driven 3D bioprinting for regenerative medicine: From bench to bedside. *Bioact Mater*. 2025;45:201-230.}

    doi: 10.1016/j.bioactmat.2024.11.021

15. Bedell ML, Navara AM, Du Y, Zhang S, Mikos AG. Polymeric systems for bioprinting. *Chem Rev*. 2020;120(19):10744-10792.

    doi: 10.1021/acs.chemrev.9b00834

16. Ng WL, Shkolnikov V. Jetting-based bioprinting: process, dispense physics, and applications. *Bio Design Manuf*. 2024;7(5):771-799.

    doi: 10.1007/s42242-024-00285-3

17. Park JA, Lee Y, Jung S. Inkjet-based bioprinting for tissue engineering. *Organoid*. 2023;3:e12.

    doi: 10.51335/organoid.2023.3.e12

18. Peng Y, Zhao H, Hu S, *et al*. Exploring the impact of osteoprotegerin on osteoclast and precursor fusion: Mechanisms and modulation by ATP. *Differentiation*. 2024;138:100789.

    doi: 10.1016/j.diff.2024.100789

19. Zhao DK, Xu HQ, Yin J, Yang HY. Inkjet 3D bioprinting for tissue engineering and pharmaceutics. *J Zhejiang Univ Sci*. 2022;23(12):955-973.

    doi: 10.1631/2023.a2200569

20. Huang X, Ng WL, Yeong WY. Predicting the number of printed cells during inkjet-based bioprinting process based on droplet velocity profile using machine learning approaches. *J Intell Manuf*. 2023;35(5):2349-2364.

    doi: 10.1007/s10845-023-02167-4

21. Huang C, Lin H, Yang J. (Xiaoming). A robust method for increasing Fc glycan high mannose level of recombinant antibodies. *Biotechnol Bioeng*. 2015;112(6):1200-1209.

    doi: 10.1002/bit.25534

22. Xu K, Ko SH, Chen J. Advances in wearable and implantable bioelectronics for precision medicine. *Biodes Manuf*. 2024;7(4):383-387.

    doi: 10.1007/s42242-024-00302-5

23. Zeng C, Hua S, Zhou J, *et al*. Oral microalgae-based biosystem to enhance irreversible electroporation immunotherapy in hepatocellular carcinoma. *Adv Sci* (*Weinh*). 2025;12(15):e2409381.

    doi: 10.1002/advs.202570101

24. Glass AM, Patel JS, Goodby JW, Olson DH, Geary JM. Pyroelectric detection with smectic liquid crystals. *J Appl Phys*. 1986;60(8):2778-2782.

    doi: 10.1063/1.337111

25. Lee CKW, Pan Y, Yang R, Kim M, Li MG. Laser-Induced Transfer of Functional Materials. *Top Curr Chem (Cham)*. 2023;381(4):1-18.

    doi: 10.1007/s41061-023-00429-6

26. Liu Y, Ding Y, Yang L, Sun R, Zhang T, Yang X. Research and progress of laser cladding on engineering alloys: A review. *J Manuf Process*. 2021;66:341-363.

    doi: 10.1016/j.jmapro.2021.03.061

27. Alayavalli K, Bourell D. Fabrication of modified graphite bipolar plates by indirect selective laser sintering (SLS) for direct methanol fuel cells. *Rapid Prototyp J*. 2010;16(4):275-278.

doi: 10.1108/13552541080000464

28. Smith AV. How to use SNLO nonlinear optics software to select nonlinear crystals and model their performance. In: *SPIE Proceedings*. Vol. 4972. SPIE; 2003. p. 50.

doi: 10.1117/12.472831

29. Caughman JBO, Baylor LR, Guillorn MA, Merkulov VI, Lowndes DH, Allard LF. Growth of vertically aligned carbon nanofibers by low-pressure inductively coupled plasma-enhanced chemical vapor deposition. *Appl Phys Lett*. 2003;83(6):1207-1209.

doi: 10.1063/1.1597981

30. Zhou W, Yu Y, Bai S, Hu A. Laser direct writing of waterproof sensors inside flexible substrates for wearable electronics. *Opt Laser Technol*. 2021;135:106694.

doi: 10.1016/j.optlastec.2020.106694

31. Hsieh JF, Lin PD. Application of homogenous transformation matrix to measurement of cam profiles on coordinate measuring machines. *Int J Mach Tools Manuf*. 2007;47(10):1593-1606.

doi: 10.1016/j.ijmachtools.2006.11.001

32. Blasco E, Müller J, Müller P, *et al*. Fabrication of conductive 3D gold-containing microstructures via direct laser writing. *Adv Mater*. 2016;28(18):3592-3595.

doi: 10.1002/adma.201506126

33. Jo Y, Park HJ, Kim Y, *et al*. Form-factor free 3D copper circuits by surface-conformal direct printing and laser writing. *Adv Funct Mater.* 2020;30(45):2004659.

doi: 10.1002/adfm.202004659

34. Li RZ, Peng R, Kihm KD, *et al*. High-rate in-plane micro-supercapacitors scribed onto photo paper using in situ femtolaser-reduced graphene oxide/Au nanoparticle microelectrodes. *Energy Environ Sci*. 2016;9(4):1458-1467.

doi: 10.1039/c5ee03637b

35. Zhou X, Guo W, Zhu Y, Peng P. The laser writing of highly conductive and anti-oxidative copper structures in liquid. *Nanoscale*. 2020;12(2):563-571.

doi: 10.1039/c9nr07248a

36. Mahmood MA, Popescu AC, Mihailescu IN. Metal matrix composites synthesized by laser-melting deposition: A review. *Materials* (*Basel*). 2020;13(11):2593.

doi: 10.3390/ma13112593

37. MacKenzie M, Chi H, Varma M, Pal P, Kaar A, Paterson L. Femtosecond laser fabrication of silver nanostructures on glass for surface enhanced Raman spectroscopy. *Sci Rep.* 2019;9(1):17058.

doi: 10.1038/s41598-019-53328-6

38. Armon N, Greenberg E, Edri E, Nagler-Avramovitz O, Elias Y, Shpaisman H. Laser-based printing: From liquids to microstructures. *Adv Funct Mater*. 2021;31(13):2008547.

doi: 10.1002/adfm.202008547

39. Zhou W, Bai S, Ma Y, *et al*. Laser-direct writing of silver metal electrodes on transparent flexible substrates with high-bonding strength. *ACS Appl Mater Interfaces*. 2016;8(37):24887-24892.

doi: 10.1021/acsami.6b07696

40. Pinkerton AJ. Laser direct metal deposition: Theory and applications in manufacturing and maintenance. In: *Advances in Laser Materials Processing*. Netherlands: Elsevier; 2010. p. 461-491.

doi: 10.1533/9781845699819.6.461

41. Lu WE, Zhang YL, Zheng ML, *et al*. Femtosecond direct laser writing of gold nanostructures by ionic liquid assisted multiphoton photoreduction. *Opt Express*. 2013;3(10):1660.

doi: 10.1364/ome.3.001660

42. Gąsiorowski L, Chai C, Rozanski A, *et al*. Regeneration in the absence of canonical neoblasts in an early branching flatworm. *Nat Commun*. 2025;16(1):1232.

doi: 10.1038/s41467-024-54716-x

43. Odde DJ, Renn MJ. Laser-guided direct writing of living cells. *Biotechnol Bioeng*. 2000;67(3):312-318.

doi: 10.1038/s41467-024-54716-x

44. Schiele NR, Corr DT, Huang Y, Raof NA, Xie Y, Chrisey DB. Laser-based direct-write techniques for cell printing. *Biofabrication*. 2010;2(3):032001.

doi: 10.1088/1758-5082/2/3/032001

45. Daly AC, Prendergast ME, Hughes AJ, Burdick JA. Bioprinting for the biologist. *Cell*. 2021;184(1):18-32.

doi: 10.1016/j.cell.2020.12.002

46. Sun W, Starly B, Daly AC, *et al*. The bioprinting roadmap. *Biofabrication*. 2020;12(2):022002.

doi: 10.1088/1758-5090/ab5158

47. Zein I, Hutmacher DW, Tan KC, Teoh SH. Fused deposition modeling of novel scaffold architectures for tissue engineering applications. *Biomaterials*. 2002;23(4):1169-1185.

doi: 10.1016/s0142-9612(01)00232-0

48. Yan B, Ouyang Q, Zhao Z, *et al*. Potent killing of HBV-related hepatocellular carcinoma by a chimeric protein of anti-HBsAg single-chain antibody and truncated Bid. *Biomaterials*. 2013;34(20):4880-4889.

doi: 10.1016/j.biomaterials.2013.03.046

49. Delaneau O, Marchini J, 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun*. 2014;5(1):3934.

    doi: 10.1038/ncomms4934

50. Markstedt K, Mantas A, Tournier I, Martínez Ávila H, Hägg D, Gatenholm P. 3D bioprinting human chondrocytes with nanocellulose-alginate bioink for cartilage tissue engineering applications. *Biomacromolecules*. 2015;16(5):1489-1496.

    doi: 10.1021/acs.biomac.5b00188

51. Kalluru P, Vankayala R, Chiang CS, Hwang KC. Nano-graphene oxide-mediated *in vivo* fluorescence imaging and bimodal photodynamic and photothermal destruction of tumors. *Biomaterials*. 2016;95:1-10.

    doi: 10.1016/j.biomaterials.2016.04.006

52. Kolesky DB, Truby RL, Gladman AS, Busbee TA, Homan KA, Lewis JA. 3D bioprinting of vascularized, heterogeneous cell-laden tissue constructs. *Adv Mater*. 2014;26(19):3124-3130.

    doi: 10.1002/adma.201305506

53. Kirillova A, Maxson R, Stoychev G, Gomillion CT, Ionov L. 4D biofabrication using shape-morphing hydrogels. *Adv Mater*. 2017;29(46):1703443.

    doi: 10.1002/adma.201703443

54. Marques CF, Diogo GS, Pina S, Oliveira JM, Silva TH, Reis RL. Collagen-based bioinks for hard tissue engineering applications: A comprehensive review. *J Mater Sci Mater Med*. 2019;30(3):32.

    doi: 10.1007/s10856-019-6234-x

55. Nawroth JC, Barrile R, Conegliano D, van Riet S, Hiemstra PS, Villenave R. Stem cell-based lung-on-chips: The best of both worlds? *Adv Drug Deliv Rev*. 2019;140:12-32.

    doi: 10.1016/j.addr.2018.07.005

56. Ng WL, Shkolnikov V. Optimizing cell deposition for inkjet-based bioprinting. *Int J Bioprint*. 2024;10(2):2135.

    doi: 10.36922/ijb.2135

57. Ligon SC, Liska R, Stampfl J, Gurr M, Mülhaupt R. Polymers for 3D printing and customized additive manufacturing. *Chem Rev.* 2017;117(15):10212-10290.

    doi: 10.1021/acs.chemrev.7b00074

58. AbouelNour Y, Gupta N. *In-situ* monitoring of sub-surface and internal defects in additive manufacturing: A review. *Mater Des*. 2022;222:111063.

    doi: 10.1016/j.matdes.2022.111063

59. Tomono D, Nishimura T. Near-infrared two-wavelength shift-and-add. In: *SPIE Proceedings*. Vol. 3019. SPIE; 1997. p. 9.

    doi: 10.1117/12.275176

60. Han D, Yang C, Fang NX, Lee H. Rapid multi-material 3D printing with projection micro-stereolithography using dynamic fluidic control. *Additive Manufacturing*. 2019;27:606-615.

    doi: 10.1016/j.addma.2019.03.031

61. Das S, Beaman JJ, Wohlert M, Bourell DL. Freeform fabrication of high performance titanium components via SLS/HIP. *MRS Proceedings*. 1998;542:45-50.

    doi: 10.1557/proc-542-45

62. Stampfl J, Baudis S, Heller C, *et al*. Photopolymers with tunable mechanical properties processed by laser-based high-resolution stereolithography. *J Micromech Microeng*. 2008;18(12):125014.

    doi: 10.1088/0960-1317/18/12/125014

63. Lim KS, Levato R, Costa PF, *et al*. Bio-resin for high resolution lithography-based biofabrication of complex cell-laden constructs. *Biofabrication*. 2018;10(3):034101.

    doi: 10.1088/1758-5090/aac00c

64. Kafle A, Luis E, Silwal R, Pan HM, Shrestha PL, Bastola AK. 3D/4D printing of polymers: Fused deposition modelling (FDM), selective laser sintering (SLS), and stereolithography (SLA). *Polymers* (*Basel*). 2021;13(18):3101.

    doi: 10.3390/polym13183101

65. Jin G, Shin SH, Shim JS, Lee KW, Kim JE. Accuracy of 3D printed models and implant-analog positions according to the implant-analog-holder offset, inner structure, and printing layer thickness: An *in-vitro* study. *Int J Dent*. 2022;125:104268.

    doi: 10.1016/j.jdent.2022.104268

66. Kumar AS, Venkatesalu S, Dilliyappan S, *et al*. Microfluidics as diagnostic tools. *Clin Chim Acta*. 2024;556:117841.

    doi: 10.1016/j.cca.2024.117841

67. Skoog SA, Goering PL, Narayan RJ. Stereolithography in tissue engineering. *J Mater Sci Mater Med.* 2013;25(3):845-856.

    doi: 10.1007/s10856-013-5107-y

68. Mao S, Pang Y, Liu T, *et al*. Bioprinting of in vitro tumor models for personalized cancer treatment: A review. *Biofabrication*. 2020;12(4):042001.

    doi: 10.1088/1758-5090/ab97c0

69. Peng X, Xia X, Xu X, *et al*. Ultrafast self-gelling powder mediates robust wet adhesion to promote healing of gastrointestinal perforations. *Sci Adv*. 2021;7(23):eabe8739.

    doi: 10.1126/sciadv.abe8739

70. Chen J, Sun J, Wang Q, *et al*. Systemic deficiency of PTEN accelerates breast cancer growth and metastasis. *Front Oncol*. 2022;12:825484.

    doi: 10.3389/fonc.2022.825484

71. Cui G, Dong K, Zhou JY, *et al*. Spatiotemporal transcriptomic atlas reveals the dynamic characteristics and key regulators of planarian regeneration. *Nat Commun*. 2023;14(1):3205.

doi: 10.1038/s41467-023-39016-0

72. Mironov V. Printing technology to produce living tissue. *Expert Opin Biol Ther.* 2003;3(5):701-704.

doi: 10.1517/14712598.3.5.701

73. Lee VK, Dai G. Printing of three-dimensional tissue analogs for regenerative medicine. *Ann Biomed Eng.* 2016;45(1):115-131.

doi: 10.1007/s10439-016-1613-7

74. Mandrycky C, Wang Z, Kim K, Kim DH. 3D bioprinting for engineering complex tissues. *Biotechnol Adv.* 2016;34(4):422-434.

doi: 10.1016/j.biotechadv.2015.12.011

75. Hospodiuk M, Dey M, Sosnoski D, Ozbolat IT. The bioink: A comprehensive review on bioprintable materials. *Biotechnol Adv*. 2017;35(2):217-239.

doi: 10.1016/j.biotechadv.2016.12.006

76. Ferreira AM, Gentile P, Chiono V, Ciardelli G. Collagen for bone tissue regeneration. *Acta Biomater.* 2012;8(9):3191-3200.

doi: 10.1016/j.actbio.2012.06.014

77. Gómez-Guillén MC, Giménez B, López-Caballero ME, Montero MP. Functional and bioactive properties of collagen and gelatin from alternative sources: A review. *Food Hydrocolloids*. 2011;25(8):1813-1827.

doi: 10.1016/j.foodhyd.2011.02.007

78. Sun TL, Kurokawa T, Kuroda S, *et al*. Physical hydrogels composed of polyampholytes demonstrate high toughness and viscoelasticity. *Nat Mater*. 2013;12(10):932-937.

doi: 10.1038/nmat3713

79. Slaughter BV, Khurshid SS, Fisher OZ, Khademhosseini A, Peppas NA. Hydrogels in regenerative medicine. *Adv Mater*. 2009;21(32-33):3307-3329.

doi: 10.1002/adma.200802106

80. Seliktar D. Designing cell-compatible hydrogels for biomedical applications. *Science*. 2012;336(6085):1124-1128.

doi: 10.1126/science.1214804

81. Benton JA, DeForest CA, Vivekanandan V, Anseth KS. Photocrosslinking of gelatin macromers to synthesize porous hydrogels that promote valvular interstitial cell function. *Tissue Eng Part A*. 2009;15(11):3221-3230.

doi: 10.1089/ten.tea.2008.0545

82. Nichol JW, Koshy ST, Bae H, Hwang CM, Yamanlar S, Khademhosseini A. Cell-laden microengineered gelatin methacrylate hydrogels. *Biomaterials*. 2010;31(21):5536-5544.

doi: 10.1016/j.biomaterials.2010.03.064

83. Akimoto AM, Hasuike E, Tada H, *et al*. Design of tetra-arm PEG-crosslinked thermoresponsive hydrogel for 3D cell culture. *Anal Sci*. 2016;32(11):1203-1205.

doi: 10.2116/analsci.32.1203

84. Wolf MT, Daly KA, Brennan-Pierce EP, *et al.* A hydrogel derived from decellularized dermal extracellular matrix. *Biomaterials*. 2012;33(29):7028-7038.

doi: 10.1016/j.biomaterials.2012.06.051

85. Samandari M, Quint J, Rodríguez-delaRosa A, Sinha I, Pourquié O, Tamayol A. Bioinks and bioprinting strategies for skeletal muscle tissue engineering. *Adv Mater*. 2022;34(12):e2105883.

doi: 10.1002/adma.202105883

86. Lasaosa FL, Zhou Y, Song J, *et al*. Nature-inspired scarless healing: guiding biomaterials design for advanced therapies. *Tissue Eng Part B Rev*. 2024;30(3):371-384.

doi: 10.1089/ten.teb.2023.0224

87. Noor N, Shapira A, Edri R, Gal I, Wertheim L, Dvir T. 3D printing of personalized thick and perfusable cardiac patches and hearts. *Adv Sci* (*Weinh*). 2019;6(11):1900344.

doi: 10.1002/advs.201900344

88. Chen H, Cheng Y, Wang X, *et al*. 3D printed in vitro tumor tissue model of colorectal cancer. *Theranostics*. 2020;10(26):12127-12143.

doi: 10.7150/thno.52450

89. Sbirkov Y, Molander D, Milet C, *et al.* A colorectal cancer 3D bioprinting workflow as a platform for disease modeling and chemotherapeutic screening. *Front Bioeng Biotechnol*. 2021;9:755563.

doi: 10.3389/fbioe.2021.755563

90. Chen X, Yue Z, Winberg PC, Lou YR, Beirne S, Wallace GG. 3D bioprinting dermal-like structures using species-specific ulvan. *Biomater Sci.* 2021;9(7):2424-2438.

doi: 10.1039/d0bm01784a

91. Derakhshanfar S, Mbeleck R, Xu K, Zhang X, Zhong W, Xing M. 3D bioprinting for biomedical devices and tissue engineering: A review of recent trends and advances. *Bioact Mater*. 2018;3(2):144-156.

doi: 10.1016/j.bioactmat.2017.11.008

92. Jin Z, Zhang Z, Shao X, Gu GX. Monitoring anomalies in 3D bioprinting with deep neural networks. *ACS Biomater Sci Eng*. 2023;9(7):3945-3952.

doi: 10.1021/acsbiomaterials.0c01761

93. Zhu Z, Ng DWH, Park HS, McAlpine MC. 3D-printed multifunctional materials enabled by artificial-intelligence-assisted fabrication technologies. *Nat Rev Mater*. 2020;6(1):27-47.

doi: 10.1038/s41578-020-00235-2

94. Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. *NPJ Comput Mater*. 2023;9(1):24.

doi: 10.1038/s41524-023-01000-z

95. Freeman S, Calabro S, Williams R, Jin S, Ye K. Bioink formulation and machine learning-empowered bioprinting optimization. *Front Bioeng Biotechnol*. 2022;10:913579.

doi: 10.3389/fbioe.2022.913579

96. Bonatti AF, Vozzi G, Chua CK, Maria CA. Deep learning quality control loop of the extrusion-based bioprinting process. *Int J Bioprint*. 2022;8(4):620.

doi: 10.18063/ijb.v8i4.620

97. Theus AS, Ning L, Hwang B, *et al*. Bioprintability: Physiomechanical and biological requirements of materials for 3D bioprinting processes. *Polymers* (*Basel*). 2020;12(10):2262.

doi: 10.3390/polym12102262

98. Vahed R, Zareie Rajani HR, Milani AS. Can a black-box AI replace costly DMA testing?-A case study on prediction and optimization of dynamic mechanical properties of 3D printed acrylonitrile butadiene styrene. *Materials* (*Basel*). 2022;15(8):2855.

doi: 10.3390/ma15082855

99. Deng B, Chen S, Lasaosa FL, *et al*. Predicting rheological properties of HAMA/GelMA hybrid hydrogels via machine learning. *J Mech Behav Biomed Mater*. 2025;168:107005.

doi: 10.1016/j.jmbbm.2025.107005

100. Chen H, Liu Y, Balabani S, Hirayama R, Huang J. Machine learning in predicting printable biomaterial formulations for direct ink writing. *Research* (*Wash D C*). 2023;6:0197.

doi: 10.34133/research.0197

101. Hashemi A, Ezati M, Zumberg I, *et al.* Characterization and optimization of a biomaterial ink aided by machine learning-assisted parameter suggestion. *Mater Today Commun*. 2024;40:109777.

doi: 10.1016/j.mtcomm.2024.109777

102. Ruberu K, Senadeera M, Rana S, *et al*. Coupling machine learning with 3D bioprinting to fast track optimisation of extrusion printing. *Appl Mater Today*. 2021;22:100914.

doi: 10.1016/j.apmt.2020.100914

103. Schwab A, Levato R, D'Este M, Piluso S, Eglin D, Malda J. Printability and shape fidelity of bioinks in 3D bioprinting. *Chem Rev*. 2020;120(19):11028-11055.

doi: 10.1021/acs.chemrev.0c00084

104. Lee J, Oh SJ, An SH, Kim WD, Kim SH. Machine learning-based design strategy for 3D printable bioink: Elastic modulus and yield stress determine printability. *Biofabrication*. 2020;12(3):035018.

doi: 10.1088/1758-5090/ab8707

105. Fastowicz J, Okarma K. Fast quality assessment of 3D printed surfaces based on structural similarity of image regions. In: *Conference: 2018 International Interdisciplinary PhD Workshop (IIPhDW). IEEE*; 2018.

doi: 10.1109/iiphdw.2018.8388399

106. Paraskevoudis K, Karayannis P, Koumoulos EP. Real-time 3D printing remote defect detection (stringing) with computer vision and artificial intelligence. *Processes*. 2020;8(11):1464.

doi: 10.3390/pr8111464

107. Brion DAJ, Pattinson SW. Generalisable 3D printing error detection and correction via multi-head neural networks. *Nat Commun*. 2022;13(1):4654.

doi: 10.1038/s41467-022-31985-y

108. Momoli C, Costa B, Lenti L, *et al*. The evolution of anticancer 3D *in vitro* models: The potential role of machine learning and AI in the next generation of animal-free experiments. *Cancers* (*Basel*). 2025;17(4):700.

doi: 10.3390/cancers17040700

109. Chen Z, Ma N, Sun X, *et al*. Automated evaluation of tumor spheroid behavior in 3D culture using deep learning-based recognition. *Biomaterials*. 2021;272:120770.

doi: 10.1016/j.biomaterials.2021.120770

110. Mali AK, Murugappan S, Prasad JR, Tofail SAM, Thorat ND. A deep learning pipeline for morphological and viability assessment of 3D cancer cell spheroids. *Biol Methods Protoc*. 2025;10(1):bpaf030.

doi: 10.1093/biomethods/bpaf030

111. Kang HW, Lee SJ, Ko IK, Kengla C, Yoo JJ, Atala A. A 3D bioprinting system to produce human-scale tissue constructs with structural integrity. *Nat Biotechnol*. 2016;34(3):312-319.

doi: 10.1038/nbt.3413

112. Kolesky DB, Homan KA, Skylar-Scott MA, Lewis JA. Three-dimensional bioprinting of thick vascularized tissues. *Proc Natl Acad Sci U S A*. 2016;113(12):3179-3184.

doi: 10.1073/pnas.1521342113

113. Lee A, Hudson AR, Shiwarski DJ, *et al*. 3D bioprinting of collagen to rebuild components of the human heart. *Science*. 2019;365(6452):482-487.

doi: 10.1126/science.aav9051

114. Zhang YS, Arneri A, Bersini S, Shin SR, *et al*. Bioprinting 3D microfibrous scaffolds for engineering endothelialized myocardium and heart-on-a-chip. *Biomaterials*. 2016;110:45-59.

doi: 10.1016/j.biomaterials.2016.09.003

115. Hinton TJ, Jallerat Q, Palchesko RN, *et al*. Three-dimensional printing of complex biological structures by freeform reversible embedding of suspended hydrogels. *Sci Adv*. 2015;1(9):e1500758.

doi: 10.1126/sciadv.1500758

116. Zhang YS, Yue K, Aleman J, *et al*. 3D bioprinting for tissue and organ fabrication. *Ann Biomed Eng*. 2017;45(1):148-163.

doi: 10.1007/s10439-016-1612-8

117. Chae S, Ha DH, Lee H. 3D bioprinting strategy for engineering vascularized tissue models. *Int J Bioprint*. 2023;9(5):748.

doi: 10.18063/ijb.748

118. Badylak SF, Weiss DJ, Caplan A, Macchiarini P. Engineered whole organs and complex tissues. *Lancet*. 2012;379(9819):943-952.

doi: 10.1016/S0140-6736(12)60073-7

119. Langer R, Vacanti JP. Tissue engineering. *Science*. 1993;260(5110):920-926.

doi: 10.1126/science.8493529

120. Kirillova A, Bushev S, Abubakirov A, Sukikh G. Bioethical and legal issues in 3D bioprinting. *Int J Bioprint*. 2020;6(3):272.

doi: 10.18063/ijb.v6i3.272

121. Datta P, Cabrera LY, Ozbolat IT. Ethical challenges with 3D bioprinted tissues and organs. *Trends Biotechnol*. 2023;41(1):6-9.

doi: 10.1016/j.tibtech.2022.08.012

122. Jian H, Wang M, Wang S, Wang A, Bai S. 3D bioprinting for cell culture and tissue fabrication. *Biodes Manuf*. 2018;1(1):45-61.

doi: 10.1007/s42242-018-0006-1

123. Groll J, Boland T, Blunk T, *et al*. Biofabrication: Reappraising the definition of an evolving field. *Biofabrication*. 2016;8(1):013001.

doi: 10.1088/1758-5090/8/1/013001

124. Williams DJ, Sebastine IM. Tissue engineering and regenerative medicine: Manufacturing challenges. *IEE Proc Nanobiotechnol*. 2005;152(6):207-210.

doi: 10.1049/ip-nbt:20050001

125. Moroni L, Boland T, Burdick JA, *et al*. Biofabrication: A guide to technology and terminology. *Trends Biotechnol*. 2018;36(4):384-402.

doi: 10.1016/j.tibtech.2017.10.015

126. Ozbolat IT, Moncal KK, Gudapati H. Evaluation of bioprinter technologies. *Addit Manuf*. 2017;13:179-200.

doi: 10.1016/j.addma.2016.10.003

127. Papadimitroulas P, Brocki L, Christopher Chung N, *et al*. Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. *Phys Med*. 2021;83:108-121.

doi: 10.1016/j.ejmp.2021.03.009

128. An J, Chua CK, Mironov V. Application of machine learning in 3D bioprinting: Focus on development of big data and digital twin. *Int J Bioprint*. 2021;7(1):342.

doi: 10.18063/ijb.v7i1.342

129. Rodríguez-Salvador M, Rio-Belver RM, Garechana-Anacabe G. Scientometric and patentometric analyses to determine the knowledge landscape in innovative technologies: The case of 3D bioprinting. *PLoS One*. 2017;12(6):e0180375.

doi: 10.1371/journal.pone.0180375

130. García-García LA, Rodríguez-Salvador M. Disclosing main authors and organisations collaborations in bioprinting through network maps analysis. *J Biomed Semantics*. 2020;11(1):3.

doi: 10.1186/s13326-020-0219-z

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

# Transforming pharmaceutical quality assurance and validation through artificial intelligence

**Vaibhav Adhao*, Jaya Ambhore*, and Shreyash Chaudhari**

Department of Quality Assurance, Dr. Rajendra Gode College of Pharmacy, Malkapur, Maharashtra, India

## Abstract

The evolution of artificial intelligence (AI) in the pharmaceutical industry spans from its early applications in automating administrative tasks to its pivotal role in drug discovery, personalized medicine, and safety enhancement. AI contributes significantly to data analysis, real-time process monitoring, defect detection, predictive maintenance, and compliance assurance, thereby enhancing efficiency, accuracy, and regulatory adherence. This review assesses the transformative functions of AI integration in revolutionizing quality assurance and validation across the pharmaceutical industry and highlights the contribution of AI in advancing quality frameworks, core values, and smart manufacturing. Moreover, the role of AI in enhancing validation processes and the critical importance of data and algorithms are discussed. As AI continues to reshape the pharmaceutical industry, it emphasizes the synergy between technological innovation and quality enhancement.

***Keywords:*** Artificial intelligence; Quality assurance; Validation; Pharmaceutical industry; Software development; Predictive maintenance; Compliance

***Corresponding authors:**
Vaibhav Adhao
(adhao.vaibhav@gmail.com)
Jaya Ambhore
(ambhorejp02@gmail.com)

**Citation:** Adhao V, Ambhore J, Chaudhari S. Transforming pharmaceutical quality assurance and validation through artificial intelligence. *Artif Intell Health*. 2026;3(1):18-28.
doi: 10.36922/AIH025160032

**Received:** April 17, 2025

**1st revised:** July 21, 2025

**2nd revised:** July 29, 2025

**Accepted:** August 1, 2025

**Published online:** August 13, 2025



## 1. Introduction

Artificial intelligence (AI) is a branch of computer science that focuses on creating smart computer programs to solve diverse problems. AI has been applied in different sectors, namely, business, healthcare, and engineering.[1] The primary objective of AI is to resolve significant information processing issues and transparently communicate them. AI also involves developing and using specialized software to analyze data, make inferences, and extract deeper insights, including tasks, such as object identification, pattern recognition, and related item classification.[2]

AI has gained significant traction in the pharmaceutical industry in recent years. By performing tasks traditionally carried out by humans, scientists have utilized AI to develop new medications and expedite the process, significantly revolutionizing the industry.[3]

### 1.1. Quality assurance (QA) and validation in various industries

Most industries, including pharmaceuticals, manufacturing, healthcare, and software development, conduct QA and validation to ensure that products, processes, and systems meet established standards, regulations, and consumer requirements. The emergence of AI technology in recent years has created new opportunities to modernize

these processes using data-driven insights, automation, and intelligent decision-making.

#### *1.1.1. Pharmaceutical industry*

In the pharmaceutical industry, QA and validation play a key role in maintaining the safety, effectiveness, and consistency of pharmaceutical products. AI-driven predictive modeling can help in identifying potential hazards and optimizing industrial processes.[4] AI systems can uncover trends in large datasets from clinical trials and research that human experts may overlook,[5] resulting in quicker identification of potential side effects and enhanced patient safety. In addition, AI-powered robotic systems can automate equipment certification, reducing human error and improving productivity while meeting strict regulatory standards.[6]

The integration of AI in the pharmaceutical industry has enhanced QA and validation processes.[7] Conventionally, pharmaceutical QA and validation have relied on manual documentation, empirical observations, and rule-based controls. While these conventional practices have ensured compliance and consistency, they often come with limitations, such as human error, slow data processing, and reactive quality control mechanisms.[8] AI, through machine learning (ML), natural language processing (NLP), and pattern recognition, is revolutionizing QA and validation by shifting the paradigm from reactive to predictive quality management. It enhances operational efficiency, decision-making accuracy, and real-time compliance with global regulatory standards. The role of AI is not merely to automate tasks but to provide intelligent insights, risk prediction, and adaptive process control that enhance product quality and patient safety.[9]

In pharmaceutical manufacturing, QA ensures that products meet the desired standards and comply with regulations. With increasing complexity in drug development, especially in biologics and personalized medicines, maintaining quality across the product lifecycle is more challenging than ever. AI-driven systems can address these complexities by enabling predictive quality analytics. For instance, ML algorithms can analyze historical batch data, environmental monitoring trends, and deviations to identify patterns that may precede quality issues. By recognizing early warning signs, AI allows QA teams to intervene proactively, reducing the risk of batch failures and product recalls. This predictive capability is particularly beneficial in continuous manufacturing settings, where real-time monitoring and quick corrective actions are critical.

In addition, AI plays a significant role in validating manufacturing processes. Process validation traditionally involves extensive experimentation, documentation, and statistical analysis to ensure consistent product quality.[10] AI enhances this process through advanced modeling techniques, such as multivariate analysis and digital twins. These tools simulate various process scenarios using vast datasets, allowing validation teams to understand the impact of critical process parameters (CPPs) and critical quality attributes without the need for multiple physical trials.[11] Moreover, AI algorithms can optimize the design of experiments, reduce the number of validation runs required, and ensure better robustness and reproducibility. By using AI, pharmaceutical companies can accelerate validation timelines, lower costs, and maintain higher assurance levels of process consistency.[12]

Another critical application of AI in pharmaceutical QA lies in analytical method validation. Conventionally, validating an analytical method involves assessing various parameters, such as accuracy, precision, specificity, linearity, and robustness. AI can enhance this process by using neural networks and regression models to predict the behavior of analytical methods under different conditions.[13] For instance, in high-performance liquid chromatography method development, AI models can predict optimal mobile phase compositions and flow rates, improving method robustness and reproducibility. Furthermore, AI tools can automate peak detection, integration, and outlier analysis in chromatography data, thereby reducing analyst bias and increasing throughput.[14]

Data integrity is a cornerstone of pharmaceutical QA, and AI contributes significantly to enhancing it. Regulatory authorities, such as the Food and Drug Administration (FDA) and European Medicines Agency, emphasize the importance of ALCOA+ principles (Attributable, Legible, Contemporaneous, Original, Accurate, Complete, Consistent, Enduring, and Available).[15] AI-based document management and audit trail systems can automatically flag anomalies, unauthorized changes, and inconsistencies in data logs.[16] NLP algorithms can parse large volumes of laboratory and production data to detect compliance gaps, suggest corrective actions, and ensure traceability.[17] AI can also monitor electronic batch records in real time, identify discrepancies during manufacturing, and alert QA teams instantly, thereby preventing the propagation of errors and ensuring batch integrity.[4,17]

In cleaning validation, AI-driven image recognition and sensor data analysis can detect residual contaminants on equipment surfaces with high precision. By combining AI with vision systems and real-time data from sensors, companies can implement continuous and automated cleaning verification.[18] This significantly reduces the reliance on swab and rinse sampling, improves turnaround

times, and enhances overall equipment reliability. In addition, AI can help determine the optimal duration, temperature, and solvent concentration of cleaning cycles, thereby conserving resources while ensuring regulatory compliance.[19]

The role of AI extends to environmental monitoring and facility validation, particularly in sterile and aseptic manufacturing environments.[20] AI-powered environmental monitoring systems can process data from hundreds of sensors, effectively monitoring particulate matter, microbial loads, temperature, and humidity in cleanrooms. ML models analyze this data to detect anomalies, predict excursions, and assess the impact of environmental variations on product quality.[21] For example, predictive models can alert operators to possible microbial contamination risks before they manifest, allowing proactive interventions. Moreover, AI-based systems help optimize heating, ventilation and air conditioning systems by predicting load variations and adjusting filtration and air changes in real time, ensuring continued compliance with ISO and good manufacturing practices standards.[22]

AI also significantly contributes to Computer System Validation (CSV) and supports compliance with regulations, such as 21 CFR Part 11 and EU Annex 11.[23] Traditional CSV relies heavily on manual documentation to verify that computerized systems operate as intended and uphold data integrity.[24] AI enhances this process by automating risk assessments, validating system performance through simulated use cases, and generating validation documentation dynamically. AI tools can assess system logs and user behavior to identify abnormal activities that may suggest data integrity violations.[25] Moreover, AI enables continuous validation by monitoring software performance in real time and updating validation statuses automatically when system upgrades or configuration changes occur.[26]

In the context of clinical trials and contract research, AI supports QA by improving protocol compliance and data validation. AI tools can automatically monitor electronic case report forms, detect inconsistencies, and alert monitors about deviations in real time. Predictive models can also identify patients at higher risk of dropout, protocol non-adherence, or adverse events, allowing for timely corrective actions.[27] NLP algorithms analyze clinical notes and adverse event reports to ensure accurate and complete safety reporting, a crucial component of QA in clinical research. This is particularly important given the increasing complexity of decentralized and virtual clinical trials, where data are generated from multiple sources and devices.[28]

Furthermore, AI enhances pharmacovigilance, which is a critical QA activity in the post-marketing phase. AI systems can analyze vast pharmacovigilance databases and real-world evidence to detect emerging safety signals more rapidly and accurately than traditional methods.[29] By automating case triage, report generation, and signal detection, AI allows pharmacovigilance teams to focus on high-value tasks, including benefit-risk assessment and regulatory communication.[30] ML models can also identify patterns in adverse drug reaction reports and correlate them with demographic or genetic data, leading to a better understanding and mitigation of safety risks.[31]

In regulatory inspections and audits, AI-driven quality management systems offer real-time compliance tracking, automated documentation generation, and audit readiness dashboards. AI tools can mine previous inspection reports, warning letters, and audit findings to generate risk-based audit plans.[32] During inspections, AI-powered chatbots and virtual assistants can retrieve standard operating procedures (SOPs), batch records, and validation protocols on demand, improving responsiveness and transparency. This not only reduces the stress and burden on QA personnel but also demonstrates a state of control and preparedness to regulators.[33]

The impact of AI on QA and validation is especially profound in the biopharmaceutical and personalized medicine sectors, where variability is inherent and processes are highly sensitive.[34] AI enables adaptive process control using real-time feedback from bioreactors and inline sensors. AI can learn from small datasets typical of personalized therapies and optimize each batch individually, ensuring consistent quality even in low-volume, high-variability scenarios. In cell and gene therapies, AI helps validate vector production, transfection efficiency, and sterility testing through predictive analytics and automated image recognition.[35]

Despite the numerous advantages, integrating AI into QA and validation faces multiple challenges. Regulatory frameworks are still evolving, and there is uncertainty about how AI-based decisions and predictions will be evaluated during audits.[36] There are concerns around transparency (the "black box" nature of some algorithms), data privacy, and cybersecurity. Pharmaceutical companies must ensure that AI systems are trained on quality-controlled, relevant datasets and that model performance is continually monitored. Validation of AI models themselves becomes a new dimension in QA, requiring documentation of algorithm design, training data provenance, performance metrics, and change management protocols.[37] Training and change management are also essential, as QA professionals require skills in data science, algorithm validation, and digital tools to effectively interact with AI systems.[38] Cross-functional collaboration between QA, information

technology, data scientists, and regulatory affairs is vital to ensure successful AI integration. Organizations must also establish governance structures to oversee ethical AI use, compliance, and continuous improvement.[39,40]

#### *1.1.2. Manufacturing*

In manufacturing, QA and validation ensure that products meet the required standards and maintain consistency across different batches. AI can analyze real-time sensor data from manufacturing lines to identify any deviations from ideal conditions and potential abnormalities.[41] This predictive capability allows manufacturers to detect potential quality issues early, minimizing downtime and waste. Intelligent algorithms can also optimize production settings to maximize efficiency and product quality. In addition, AI-powered virtual simulations aid in validating manufacturing processes before actual implementation, leading to cost savings and reduced time-to-market.[42]

#### *1.1.3. Healthcare*

In the healthcare industry, QA and validation remain vital for maintaining patient safety and ensuring accurate diagnosis and treatment plans. AI-enabled medical image analysis can enhance the precision of disease identification, facilitating early intervention. AI-driven decision support systems analyze patient data and reference extensive medical knowledge databases to assist healthcare providers in selecting optimal treatment options.[43] Automated validation of electronic health records and compliance with regulatory standards can streamline QA processes in healthcare, minimizing errors that could jeopardize patient safety.[44]

#### *1.1.4. Software development*

In software development, QA and validation are essential to ensure that code meets functional, security, and performance standards. AI can accelerate the testing process by creating and executing test cases, identifying flaws, and observing software performance in different scenarios.[45] ML algorithms can help focus QA efforts on the most critical areas by learning from past data to detect potential flaws. AI-powered code analysis can enhance continuous integration and delivery pipelines, leading to early bug detection and efficient problem resolution.[46]

### 1.2. Parameters

Table 1 presents several parameters to consider when integrating AI for pharmaceutical QA and validation.

**Table 1. Parameters to consider when integrating artificial intelligence (AI) for quality assurance (QA) and validation**

| Parameter | Description |
|---|---|
| Cost | Implementing and maintaining AI systems can be costly; Pharmaceutical companies must carefully assess the cost-effectiveness before investing in AI. |
| Security | AI systems can be susceptible to cyberattacks. Pharmaceutical companies need robust security measures to safeguard their data and systems. |
| Ethics | The ethical use of AI in healthcare raises important considerations. Pharmaceutical companies must ensure the responsible and moral use of their AI systems. |
| Scalability | AI systems should be scalable to meet the increasing demands of pharmaceutical companies. |
| Interoperability | AI systems need to work seamlessly with other systems used in the pharmaceutical industry. |
| Data privacy | Pharmaceutical companies must adhere to data privacy regulations when utilizing AI. |
| Return on investment (ROI) | Pharmaceutical companies need to be able to gauge the ROI of their AI investments. |
| Effect on jobs | AI automation may affect certain jobs in the pharmaceutical industry. Companies need strategies to mitigate the impact on employees. |
| Transparency and explainability | Pharmaceutical companies must be able to clarify how their AI systems function and make decisions. This is vital for ensuring the safety and effectiveness of AI-powered solutions.[2] |
| Bias | AI algorithms can exhibit bias, potentially leading to inaccurate or unfair outcomes. Pharmaceutical companies must acknowledge this risk and take measures to address it. |
| Integration with existing quality management systems | AI systems should seamlessly integrate with the present quality management systems utilized in the pharmaceutical industry. |
| User-friendliness | AI systems should be intuitive and easy for employees to use. |
| Validation and verification | AI systems need to undergo validation and verification to ensure they meet the requirements for their intended use. |
| Continuous improvement | Pharmaceutical companies need a process for continually enhancing their AI systems. |
| Collaboration | Pharmaceutical companies should engage in collaboration with regulators, academia, and other industry stakeholders to advance the use of AI in pharmaceutical QA/validation. |

## 2. History of AI in the pharmaceutical industry

The incorporation of AI in the pharmaceutical industry represents an important transformation in addressing complex challenges through the fusion of science and technology. Over the past few decades, AI has increasingly become intertwined with the pharmaceutical sector, reshaping various aspects of drug research, development, and healthcare. Initially, AI was primarily employed in pharmaceutical companies to streamline administrative tasks through the automation of repetitive processes and data management. However, as AI technologies advanced, their ability to analyze and comprehend large datasets became more widely recognized. Consequently, algorithms were developed to swiftly sift through massive amounts of chemical data to identify possible drug candidates. AI has since expanded its role to assist in predicting medication interactions, optimizing clinical trials, and even personalizing patient treatment plans. The ongoing collaboration between AI and pharmaceutical companies illustrates a gradual transition from simple automation to comprehensive data-driven insights, offering state-of-the-art solutions to some of the most intricate challenges in the industry.[47-51]

## 3. Role of AI in QA

AI is increasingly utilized in several aspects of pharmaceutical production and quality control (Table 2).

## 4. Application of AI in QA

AI has been applied in various areas in the pharmaceutical industry (Figure 1), and some of these are described as follows:[52]

(i) Automated testing: AI-driven test automation tools accelerate feedback loops and boost the quality of software by streamlining the testing process
(ii) Defect detection: AI algorithms can identify flaws and abnormalities in production processes, reducing waste and raising product quality
(iii) Predictive maintenance: Real-time monitoring of equipment, reduced downtime, and improved maintenance plans are possible through AI-powered predictive analytics
(iv) Image and video analysis: AI systems can analyze visual data to find flaws in goods or production methods, ensuring high-quality output
(v) NLP for compliance: In highly regulated businesses, NLP algorithms help ensure compliance by assisting in the interpretation and analysis of regulatory papers and guidelines.

## 5. AI in QA and productivity

AI is transforming quality management in the pharmaceutical industry. AI-powered tools can rapidly process large volumes of data, enabling businesses to make informed decisions by identifying trends in real time. AI

**Table 2. Artificial intelligence (AI) in quality assurance (QA)**

| Role | Description |
|---|---|
| Automated data analysis and pattern recognition | Using AI with performance metrics helps keep track of product and location trends, allowing early identification of potential issues and proactive intervention before they escalate.[52] |
| Process monitoring and control | AI plays a significant role in monitoring and controlling advanced manufacturing processes, optimizing process design, and driving continuous improvement.[53] |
| Defect detection and visual inspection | AI and computer vision technologies are used to detect flaws and irregularities in pharmaceutical items and packaging, improving quality control.[52] |
| Predictive maintenance and equipment monitoring | AI is used to foresee equipment breakdowns and maintenance requirements, saving downtime and ensuring continuous production operations.[52] |
| Risk assessment and compliance | AI-driven risk assessment models are used in the pharmaceutical supply chain to ensure compliance with legal requirements.[53] |
| Adverse event monitoring and pharmacovigilance | AI is used to evaluate real-world data and spot trends associated with negative medication responses, enabling prompt and efficient safety measures.[52] |
| Personalized medicine and drug development | AI helps with customized medicine by evaluating patient data to determine the best course of action and easing medication development.[54] |
| Quality control and batch release | AI-based solutions help with quality control procedures and ensure that each batch of pharmaceutical products satisfies necessary quality standards before release.[55] |
| Supply chain management and demand prediction | AI enhances supply chain efficiency by forecasting demand, monitoring inventory levels, and optimizing logistics to ensure timely product delivery.[56] |
| Validation and regulatory considerations | AI's validity and associated legal considerations in pharmaceutical QA are thoroughly documented to ensure alignment with applicable regulatory standards.[53] |

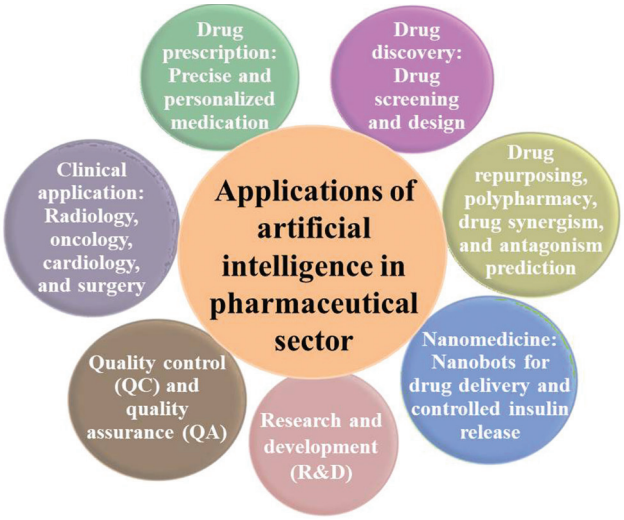



**Figure 1.** Applications of artificial intelligence in the pharmaceutical sector

is significantly improving quality control by automating tasks traditionally performed by humans. Through the automation of testing and inspection processes, AI enhances accuracy, reduces errors, and contributes to greater customer satisfaction. These systems are highly adaptable—easily integrated into existing infrastructure and scalable to meet evolving business demands. By leveraging deep learning, AI systems can continuously learn from data, improving performance over time without the need for manual reprogramming. This represents a major shift from conventional rule-based systems, which relied on static inspection parameters. In effect, AI acts as an intelligent assistant, continuously monitoring products and services to identify defects or inefficiencies. Overall, AI is making quality control smarter, more efficient, and more reliable across a wide range of industries.[57]

The concept of "Industry 4.0" encompasses a micro-industry development strategy that aims to create various business models through personalized design and marketing. It focuses on improving collaboration between engineering and logistics by integrating activities along the entire value chain, from supply to demand. This integration facilitates communication between customers and suppliers, leading to accelerated quality analysis, improvement, and design enhancement. The strategy also emphasizes openness and effectiveness in resolving common problems through collaboration between customers and suppliers in logistics. While smaller businesses may lack the potential to fully embrace Industry 4.0, certain mature corporations are aligning themselves with this objective. Production management in Industry 4.0 adheres to the MESA/ISA-95 standard, and small and medium-sized businesses can gradually achieve Industry 4.0 objectives by enhancing the quality of information within their internal processes.[52] The integration of AI in quality management is revolutionizing traditional business processes by enhancing efficiency, accuracy, and decision-making capabilities. AI automates quality control processes, enabling organizations to streamline inspections and testing procedures. It also facilitates predictive analytics for QA, allowing for the collection of real-time data, monitoring of quality parameters, and identification of anomalies. AI-driven technologies are reshaping quality control by automating testing and inspection processes, making them more scalable, manageable, and efficient. Acting as a highly capable assistant, AI can detect defects in products and services across various industries before they escalate into serious issues. By analyzing historical data, AI systems can predict potential quality concerns, enabling proactive intervention and reducing the likelihood of failure. This predictive capability leads to greater consistency in product quality and improved customer satisfaction. Furthermore, AI supports post-production quality monitoring—such as verifying packaging integrity or identifying contaminants—ensuring that products meet standards throughout the entire lifecycle. Overall, AI is driving a shift toward smarter, end-to-end QA, reinforcing high standards from production to final delivery.[56]

### 5.1. AI-enhanced validation

AI plays a pivotal role in enhancing validation processes across multiple dimensions of pharmaceutical manufacturing and QA. From ensuring data integrity and algorithm reliability to enabling virtual simulations and high-precision image analysis, AI technologies are streamlining validation workflows, reducing human error, and improving overall compliance and product quality. AI algorithms can cross-reference and validate large datasets to ensure the correctness and integrity of the data. Nonetheless, AI should be extensively validated to ensure reliable and secure outcomes in the pharmaceutical sector. In addition, AI-powered simulations improve validation procedures by enabling virtual testing of goods and systems before practical deployment.[58] Besides that, AI-powered image analysis is a crucial aspect of pharmaceutical QA, involving the use of advanced image recognition algorithms to detect defects, verify labelling accuracy, and assess the physical attributes of pharmaceutical products. AI algorithms excel at detecting even the smallest defects in images and sensor data. They play a crucial role in ensuring that pharmaceutical products are manufactured accurately and comply with stringent regulatory standards. By quickly and precisely identifying flaws in tablets and other products, AI enhances inspection speed and

accuracy. Unlike humans, AI systems do not experience fatigue or overlook details, making them reliable across diverse inspection scenarios. Employing AI for quality checks helps maintain high product standards, accelerates production processes, and ultimately safeguards patient health.[59]

### 5.2. Continuous process validation: Integrating AI into pharmaceutical manufacturing

The incorporation of AI technologies into the validation processes of pharmaceutical manufacturing enables real-time monitoring and control of production, ensuring that operations remain within validated parameters. Continuous process validation powered by AI offers significant benefits, including consistent product quality, increased efficiency, and enhanced regulatory compliance. AI-powered image analysis plays a crucial role in pharmaceutical QA by employing advanced image recognition algorithms to detect defects, verify labeling accuracy, and assess the physical attributes of products. These algorithms can analyze vast amounts of images and sensor data to identify subtle issues that might be overlooked by humans. By automating tests and inspections, AI tools improve quality control, easily integrate with existing systems, and efficiently manage large workloads, ensuring smooth operations. This proactive approach helps identify problems early, thereby maintaining high product standards and increasing customer satisfaction. Overall, AI is transforming traditional business processes by enhancing efficiency, accuracy, and decision-making in quality management, ultimately leading to improved product quality and customer satisfaction.[60]

### 5.3. Real-time monitoring and analysis: The role of AI in pharmaceutical processes

In pharmaceutical manufacturing, AI technology is used to closely monitor critical factors and detect potential issues early, streamlining the entire process and ensuring that medicines produced meet the highest quality standards. AI functions as a highly intelligent assistant that oversees everything, from verifying optimal temperatures to identifying problems before they escalate. This not only improves efficiency but also ensures that the medicines are safe and of high quality. For example, AI can analyze vast amounts of data, such as images and sensor readings, to detect even the smallest errors that humans might overlook. Moreover, AI performs these tasks rapidly, allowing for quick corrective actions that minimize downtime and reduce waste, ensuring continuous production of high-quality medicines.[52]

## 6. Future trends

The integration of AI in pharmaceutical QA and validation is rapidly transforming the landscape of drug manufacturing, with future trends pointing toward a more intelligent, proactive, and efficient quality ecosystem.[61] As regulatory demands and the complexity of pharmaceutical processes increase, AI is emerging not only as a supporting tool but also as a key innovation in ensuring product quality and regulatory compliance.[62] Notably, the widespread adoption of predictive analytics—particularly those based on ML—leverages historical and real-time manufacturing data to identify patterns and correlations that may be overlooked by humans.[63] These insights enable predictive identification of deviations, equipment malfunctions, or quality failures before they occur, significantly reducing batch rejections, recalls, and compliance risks. This transition from reactive to predictive quality management represents a major paradigm shift in QA strategy.[64]

Another potential application of AI is the implementation of real-time release testing, also supported by process analytical technology. With AI algorithms monitoring CPPs and quality attributes in real-time, pharmaceutical companies can ensure continuous product quality throughout production, rather than relying solely on final product testing. This capability not only accelerates product release but also improves product consistency and compliance. Similarly, the advancement of continuous process verification leverages AI tools to continuously verify that processes remain within validated parameters by analyzing vast streams of operational data in real-time. This approach provides continuous assurance of process performance and control, aligning with regulatory expectations and International Council for Harmonisation (ICH) Q8–Q11 guidelines.[65]

Digital twins—virtual, AI-powered replicas of physical manufacturing environments—enable pharmaceutical companies to simulate process changes, conduct risk assessments, and optimize manufacturing conditions without disrupting actual operations. These tools enhance the effectiveness of quality by design approaches by facilitating hypothetical scenario testing, sensitivity analysis, and process optimization in a virtual environment.[66] Equally transformative is the growing use of natural language generation and NLP in automating validation documentation, such as SOPs, protocols, risk assessments, deviations, and CAPA reports. These tools can draft, update, and audit documents using regulatory-compliant language, reducing human error, workload, and approval time.[67]

As AI becomes deeply embedded in QA operations, data integrity and compliance monitoring will also evolve.

AI-driven anomaly detection algorithms will enhance data reliability by identifying unusual patterns in electronic records, audit trails, or equipment logs, thereby supporting adherence to ALCOA+ principles.[68] Furthermore, regulatory intelligence systems, powered by AI, will continuously scan global regulatory databases and inspection reports to identify changes in requirements and automatically flag areas within the organization that may require updates or corrective actions. These systems can facilitate dynamic risk assessment and adaptive compliance strategies, ensuring a state of ongoing regulatory readiness.[69]

In addition, AI will increasingly support personalized training and intelligent auditing. Employees can receive AI-curated learning paths based on their roles, prior performance, and audit outcomes, improving quality culture and knowledge retention. Similarly, internal audits can be enhanced by AI tools that can analyze large volumes of quality data and generate insights for continuous improvement. AI integration with blockchain technology will further improve traceability across the supply chain, ensuring transparent, tamper-proof records of raw material sourcing, manufacturing conditions, and product distribution.[70]

Finally, with AI's expanding footprint, ensuring its ethical use and regulatory harmonization will become increasingly crucial. Global regulatory agencies, such as the FDA, EMA, and ICH, are expected to provide clear guidance on AI validation, accountability, and transparency within good practice environments. As these frameworks develop, AI will not only drive QA and validation but also redefine regulatory strategies, making the entire pharmaceutical lifecycle more robust, compliant, and patient-centric.[71]

## 7. Conclusion

The integration of AI into QA and validation processes across various industries has brought about transformative advancements, enhancing efficiency, precision, and compliance. The historical progression from manual production to the era of Industry 4.0 illustrates a journey toward intelligent automation and data-driven decision-making, with AI at its forefront.

The role of AI in QA and validation is multifaceted. It enables computerized data analysis, pattern recognition, and predictive modeling, leading to more informed decisions and proactive risk mitigation. Real-time process monitoring and control ensure consistent product quality, while defect detection through AI-driven visual inspection minimizes errors and waste. Predictive maintenance powered by AI optimizes equipment performance and reduces downtime, bolstering operational efficiency. In the pharmaceutical sector, AI facilitates drug discovery, adverse event monitoring, and personalized medicine, collectively enhancing patient outcomes.

The integration of AI has also emphasized the importance of quality principles and core values. Key concepts, such as continuous enhancement, integration, practical implementation, and sustainable progress form the basis of effective AI-driven QA and validation strategies. Aligning AI initiatives with these established values supports a holistic approach to quality management, ensuring that the full potential of AI is effectively realized.

As industries continue to embrace AI for QA and validation, it is imperative to maintain a balance between technological advancement and human expertise. Rigorous algorithm and simulation validation, along with the integration of AI-powered systems, create a synergistic relationship that ensures both accuracy and safety.

Taken together, the application of AI in QA and validation represents a remarkable harmonization of innovation and tradition, driving industries toward higher standards of quality, productivity, and reliability. The ongoing collaboration between AI and quality professionals promises a future where data-driven insights and intelligent automation continue to elevate the standards of excellence across diverse sectors.

## Acknowledgments

None.

## Funding

None.

## Conflict of interest

The authors declare they have no competing interests.

## Author contributions

*Conceptualization:* Vaibhav Adhao
*Visualization:* Vaibhav Adhao
*Writing – original draft:* All authors
*Writing – review & editing:* All authors

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

Not applicable.

## References

1. Dasta JF. Application of artificial intelligence to pharmacy and medicine. *Hosp Pharm*. 1992;27:312-315.
2. Duch W, Swaminathan K, Meller J. Artificial intelligence approaches for rational drug design and discovery. *Curr Pharm Des*. 2007;13(14):1497-1508.
   doi: 10.2174/138161207780765954
3. Makridakis S. The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*. 2017;90:46-60.
4. Saha GC, Eni LN, Saha H, *et al*. Artificial intelligence in pharmaceutical manufacturing: Enhancing quality control and decision making. *Riv Ital Filosofia Anal Junior*. 2023;14(2):116-126.
5. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. *Trends Pharmacol Sci*. 2019;40(8):577-591.
   doi: 10.1016/j.tips.2019.05.005
6. Waqar M, Bhatti I, Khan AH. AI-powered automation: Revolutionizing industrial processes and enhancing operational efficiency. *Rev Intel Artif Med*. 2024;15(1):1151-1175.
7. Mosisa B, Malay K, Fufa F. Transformative role of artificial intelligence in the pharmaceutical sector. *J Angiother*. 2024;8(9):1-7.
8. Huang J, O'Connor T, Ahmed K, *et al*. AIChE PD2M advanced process control workshop-moving APC forward in the pharmaceutical industry. *J Adv Manuf Process*. 2021;3(1):e10071.
9. Cherekar R. The future of AI quality assurance: Emerging trends, challenges, and the need for automated testing frameworks. *Int J Emerg Trends Comput Sci Inform Technol*. 2021;2(1):19-27.
10. Mohammad AS, Devidi S, Fatima N, *et al*. An overview of validation and basic concepts of process validation: Quality assurance view point. *Asian J Pharm Technol*. 2016;6(3):169-176.
11. Borchert D, Zahel T, Thomassen YE, Herwig C, Suarez-Zuluaga DA. Quantitative CPP evaluation from risk assessment using integrated process modeling. *Bioengineering*. 2019;6(4):114.
    doi: 10.3390/bioengineering6040114
12. Rahman SN, Katari O, Pawde DM, *et al*. Application of design of experiments® approach-driven artificial intelligence and machine learning for systematic optimization of reverse phase high performance liquid chromatography method to analyze simultaneously two drugs (cyclosporin A and etodolac) in solution, human plasma, nanocapsules, and emulsions. *AAPS PharmSciTech*. 2021;22(4):155.
    doi: 10.1208/s12249-021-02026-6
13. Mundhra S, Kadiri SK, Tiwari P. Harnessing AI and machine learning in pharmaceutical quality assurance. *J Pharm Qual Assur Qual Control*. 2024;6:19-29.
14. Pawar A. Recent innovations in high-performance liquid chromatography (HPLC): Method development and validation strategies. *J Drug Deliv Biother*. 2024;1(1):55-61.
15. Gokulakrishnan D, Venkataraman S. Ensuring Data Integrity: Best Practices and Strategies in Pharmaceutical Industry. Intelligent Pharmacy. 2024. [In press].
16. Samuel A. Enhancing financial fraud detection with AI and cloud-based big data analytics: Security implications. *World J Adv Eng Technol Sci*. 2023;9(2):417-434.
17. Vaghela MC, Rathi S, Shirole RL, Verma J, Shaheen, Panigrahi S, *et al*. Leveraging AI and machine learning in six-sigma documentation for pharmaceutical quality assurance. *Chin J Appl Physiol*. 2024;40:e20240005.
18. Nandhakumar D, Kumar AM, Pavithra S. Advancements in AI-powered robotic cleaning systems: Autonomous path planning, predictive maintenance, and cleanliness assessment frameworks. In: *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE; 2025. p. 1077-1082.
19. Thakur A, Kumar A. Innovative technologies for the removal of pollutants in the chemical industries. *In: Innovative and Hybrid Technologies for Wastewater Treatment and Recycling 2024*. United States: CRC Press. p. 167-195.
20. McCall J, Barnard N, Gadient K, *et al*. Environmental monitoring for closed robotic workcells used in aseptic processing: data to support advanced environmental monitoring strategies. *AAPS PharmSciTech*. 2022;23(6):215.
    doi: 10.1208/s12249-022-02360-3
21. Popescu SM, Mansoor S, Wani OA, *et al*. Artificial intelligence and IoT driven technologies for environmental pollution monitoring and management. *Front Environ Sci*. 2024;12:1336088.
22. Sadrizadeh S. Leveraging artificial intelligence in indoor air quality management: A review of current status, opportunities, and future challenges. *REHVA European HVAC J*. 2024;61(1):35-37.
23. Raja JR, Kella A, Narayanasamy D. The essential guide to computer system validation in the pharmaceutical industry. *Cureus*. 2024;16(8):e67555.
24. Shekhar S. An in-depth analysis of intelligent data migration strategies from oracle relational databases to hadoop ecosystems: Opportunities and challenges. *Internafional J Appl Mach Learn Computafional Intell*. 2020;10(2):1-24.
25. Devineni SK, Kathiriya S, Shende A. Machine learning-powered anomaly detection: Enhancing data security and integrity. *J Artif Intell Cloud Comput*. 2023;2:1-9.
26. Baqar M, Khanda R. *The Future of Software Testing: AI-Powered Test Case Generation and Validation. arXiv preprint arXiv:2409.05808*; 2024.

27. Ahire YS, Patil JH, Chordiya HN, Deore RA, Bairagi VA. Advanced applications of artificial intelligence in pharmacovigilance: Current trends and future perspectives. *J Pharm Res*. 2024;23(1):23-33.

28. Wong A, Plasek JM, Montecalvo SP, Zhou L. Natural language processing and its implications for the future of medication safety: A narrative review of recent advances and challenges. *Pharmacotherapy*. 2018;38(8):822-841.

29. Chakraborty A, Venkatraman JV. Pharmacovigilance through phased clinical trials, post-marketing surveillance and ongoing life cycle safety. In: *The Quintessence of Basic and Clinical Research and Scientific Publishing*. Singapore: Springer Nature Singapore; 2023. p. 427-442.

30. Shukla D, Bhatt S, Gupta D, Verma S. Role of artifical intelligence in pharmacovigilance. *J Drug Discov Health Sci*. 2024;1(4):230-238.

31. Kim HR, Sung M, Park JA, *et al*. Analyzing adverse drug reaction using statistical and machine learning methods: A systematic review. *Medicine*. 2022;101(25):e29387.

32. Verma P, S. Sangle P. Role of digital transformation in inspection and certification. In: *Handbook of Quality System, Accreditation and Conformity Assessment*. Singapore: Springer Nature Singapore; 2023. p. 1-29.

33. Bhagat D, Dorle S. *The Power of Intelligent. Hyperautomation in Business and Society*. Hershey: IGI Global Scientific; 2024. p. 27.

34. Pathak SS, Gawai A, Biyani KR. Quality assurance in the age of personalized medicine: Challenges and opportunities. *Asian J Pharm Res Dev.* 2024;12(2):179-186.

35. Patel P. Impact of AI on Manufacturing and Quality Assurance in Medical Device and Pharmaceuticals Industry. *Int J Innovative Technol Exploring Eng*. 2024;13(8):9-21.

36. Krause D. Addressing the Challenges of Auditing and Testing for AI Bias: A Comparative Analysis of Regulatory Frameworks. SSRN. 2024. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5050631

37. Venkata SK. AI in audit: Unlocking deep analytical-based testing. *J Comput Sci Technol Stud*. 2025;7(3):592-601.

38. Wang C, Yang Z, Li ZS, Damian D, Lo D. *Quality Assurance for Artificial Intelligence: A Study of Industrial Concerns, Challenges and Best Practices. arXiv Preprint arXiv:2402.16391*; 2024.

39. Halwani MA, Amirkiaee SY, Evangelopoulos N, Prybutok V. Job qualifications study for data science and big data professions. *Inform Technol People*. 2022;35(2):510-525.

40. Shrivastava S, Patel D, Bhamidipaty A, *et al*. Dqa: Scalable, automated and interactive data quality advisor. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE; 2019. p. 2913-2922.

41. Villegas-Ch W, García-Ortiz J, Sánchez-Viteri S. Towards intelligent monitoring in IoT: AI applications for real-time analysis and prediction. *IEEE Access*. 2024;12:40368-40386.

doi: 10.1109/ACCESS.2024.3376707

42. Archana T, Stephen RK. The future of artificial intelligence in manufacturing industries. In: *Industry Applications of Thrust Manufacturing: Convergence with Real-Time Data and AI*; 2024. New York: IGI Global. p. 98-117.

43. Bagheri M, Bagheritaba M, Alizadeh S, Parizi MS, Matoufinia P, Luo Y. *AI-driven Decision-making in Healthcare Information Systems: A Comprehensive Review*. *Preprints*; 2024.

44. Elhaddad M, Hamam S. AI-driven clinical decision support systems: An ongoing pursuit of potential. *Cureus*. 2024;16(4):e57728.

45. Atoum I, Baklizi MK, Alsmadi I, *et al*. Challenges of software requirements quality assurance and validation: A systematic literature review. *IEEE Access*. 2021;9:137613-137634.

46. Pargaonkar S. Synergizing requirements engineering and quality assurance: A comprehensive exploration in software quality engineering. *Int J Sci Res*. 2023;12(8):2003-2007.

47. Aguilar-Gallardo C, Bonora-Centelles A. integrating artificial intelligence for academic advanced therapy medicinal products: Challenges and opportunities. *Appl Sci*. 2024;14(3):1303.

48. Harrer S, Menard J, Rivers M, *et al*. Artificial intelligence drives the digital transformation of pharma. In: *Artificial Intelligence in Clinical Practice*. United States: Academic Press; 2024. p. 345-372.

49. Kulkov I. The role of artificial intelligence in business transformation: A case of pharmaceutical companies. *Technol Soc*. 2021;66:101629.

50. Vora LK, Gholap AD, Jetha K, Thakur RR, Solanki HK, Chavda VP. Artificial intelligence in pharmaceutical technology and drug delivery design. *Pharmaceutics*. 2023;15(7):1916.

doi: 10.3390/pharmaceutics15071916

51. Raza MA, Aziz S, Noreen M, *et al*. Artificial intelligence (AI) in pharmacy: An overview of innovations. *Innov Pharm*. 2022;13(2):1-8.

doi: 10.24926/iip.v13i2.4839

52. Fisher A. The future is the present: Artificial intelligence in pharmaceutical manufacturing: FDA is anticipating how AI may advance manufacturing and improve supply chain security. *Pharm Technol*. 2023;47(9):32-34.

53. Arden NS, Fisher AC, Tyner K, Lawrence XY, Lee SL, Kopcha M. Industry 4.0 for pharmaceutical manufacturing: Preparing for the smart factories of the future. *Int J Pharm*. 2021;602:120554.

doi: 10.1016/j.ijpharm.2021.120554

54. Rosenberger S. Growth of artificial intelligence in pharma

manufacturing: Lonza describes how artificial intelligence, machine learning, and big data are improving safety, quality, and sustainability-all while lowering costs. *Genet Eng Biotechnol News*. 2022;43(1):34-36.

55. Cerchia C, Lavecchia A. New avenues in artificial-intelligence-assisted drug discovery. *Drug Discov Today*. 2023;28(4):103516.

    doi: 10.1016/j.drudis.2023.103516

56. Owczarek D. The Future of Pharmaceutical Manufacturing Process: Artificial Intelligence; 2021. Available from: https://nexocode.com/blog/posts/ai-in-pharmaceutical-manufacturing/

57. Parker PD, Parker C. Future of Electronic Health Records: A Challenge to Maximize their Utility; 2023. Available from: https://ssrn.com/abstract=4457214 or http://dx.doi.org/10.2139/ssrn.4457214

58. Kalyane D, Sanap G, Paul D, *et al*. Artificial intelligence in the pharmaceutical sector: Current scene and future prospect. In: *The Future of Pharmaceutical Product Development and Research*. United States: Academic Press; 2020. p. 73-107.

59. Chisty NM, Adusumalli HP. Applications of artificial intelligence in quality assurance and assurance of productivity. *ABC J Adv Res*. 2022;11(1):23-32.

60. Mak KK, Wong YH, Pichika MR. Artificial intelligence in drug discovery and development. *Drug Discov Eval Saf Pharmacokinetic Assays*. 2024:1461-1498.

    doi: 10.1007/978-3-031-35529-5_92

61. Blanco-Gonzalez A, Cabezon A, Seco-Gonzalez A, *et al*. The role of AI in drug discovery: Challenges, opportunities, and strategies. *Pharmaceuticals* (*Basel*). 2023;16(6):891.

62. Suriyaamporn P, Pamornpathomkul B, Patrojanasophon P, Ngawhirunpat T, Rojanarata T, Opanasopit P. The artificial intelligence-powered new era in pharmaceutical research and development: A review. *AAPS PharmSciTech*. 2024;25(6):188.

    doi: 10.1208/s12249-024-02901-y

63. Agrawal K, Nargund N. Deep learning in industry 4.0: Transforming manufacturing through data-driven innovation. In: *International Conference on Distributed Computing and Intelligent Technology*. Cham: Springer Nature Switzerland; 2024. p. 222-236.

64. Alsaidalani R, Elmadhoun B. Quality risk management in pharmaceutical manufacturing operations: Case study for sterile product filling and final product handling stage. *Sustainability*. 2022;14(15):9618.

65. Shi Z, Altan S, Banton D, *et al*. Predictive *in-vitro* dissolution for real-time release test (RTRT) for continuous manufacturing process on drug product. In: *Continuous Pharmaceutical Processing and Process Analytical Technology*. United States: CRC Press; 2023. p. 213-270.

66. Mishra V, Thakur S, Patil A, Shukla A. Quality by design (QbD) approaches in current pharmaceutical set-up. *Exp Opin Drug Deliv*. 2018;15(8):737-758.

    doi: 10.1080/17425247.2018.1504768

67. Galvis L, Offermans T, Bertinetto CG, *et al*. Retrospective quality by design r (QbD) for lactose production using historical process data and design of experiments. *Comput Ind*. 2022;141:103696.

68. Sembiring MH, Novagusda FN. Enhancing data security resilience in AI-driven digital transformation: Exploring industry challenges and solutions through ALCOA+ principles. *Acta Inform Med*. 2024;32(1):65-70.

    doi: 10.5455/aim.2024.32.65-70

69. Emeihe EV, Nwankwo EI, Ajegbile MD, Olaboye JA, Maha CC. The impact of artificial intelligence on regulatory compliance in the oil and gas industry. *Int J Life Sci Res Arch*. 2024;7(1):28-39.

70. Tutuncuoglu BT. Beyond the productivity paradox: Unveiling the hidden role of artificial intelligence in enhancing human creativity and innovation; 2024. Available from: https://dx.doi.org/10.2139/ssrn.5246291

71. Kabir M, Rana MR, Debnath A. The role of quality assurance in accelerating pharmaceutical research and development: Strategies for ensuring regulatory compliance and product integrity. *J Angiother*. 2024;8(12):1-1.

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

# Artificial intelligence and biomarker approaches for Parkinson's disease detection

**Gunjan Goswami[1] and Bhanu Prasad[2]***

[1]Independent Researcher, Calgary, Alberta, Canada

[2]Department of Computer and Information Sciences, Florida A&M University, Tallahassee, Florida, United States of America

## Abstract

Parkinson's disease (PD) is a neurological syndrome or condition that occurs due to a deficit of dopamine-producing neurons in the substantia nigra. Diagnosing PD in its early stages is difficult, as its symptoms often resemble those of other neurological diseases. Therefore, recognizing reliable biomarkers is important for discriminating PD from related conditions, monitoring disease progression, and evaluating responses to therapeutic interventions. PD biomarkers are categorized into the following classes: clinical, neuroimaging, biochemical and proteomic, and genetic. Ongoing research aims to discover the most effective PD biomarkers that could help doctors identify PD risk and accelerate early diagnosis. Artificial intelligence (AI) methods, including deep learning and machine learning, have become increasingly significant in recent years due to their ability to evaluate and process large volumes of medical data with high accuracy. Furthermore, these methods have contributed significantly to the early diagnosis and effective treatment of various diseases, such as cancer and neurological conditions such as Alzheimer's disease, PD, and multiple sclerosis. Given that PD affects a large population, the present study aims to review the applications of AI approaches in the early diagnosis of PD and the latest advancements in the field of PD biomarkers. Promising results have been obtained using various AI algorithms, which are helpful not only in identifying the PD stages but also in supporting early diagnosis. However, the implementation of these techniques in clinical practice faces challenges, including data quality and variability, model interpretability, and the need for interdisciplinary collaboration.

***Keywords:*** Artificial intelligence; Parkinson's disease; Machine learning; Biomarkers; Deep learning

***Corresponding author:**
Bhanu Prasad
(bhanu.prasad@famu.edu)

**Citation:** Goswami G, Prasad B. Artificial intelligence and biomarker approaches for Parkinson's disease detection. *Artif Intell Health*. 2026;3(1):29-53. doi: 10.36922/AIH025210048

**Received:** May 20, 2025

**1st revised:** June 27, 2025

**2nd revised:** July 10, 2025

**Accepted:** July 25, 2025

**Published online:** August 26, 2025



## 1. Introduction

Neurodegenerative diseases (NDs) are progressive conditions that damage various parts of the human nervous system, primarily affecting the brain.[1] Alzheimer's disease (AD) and Parkinson's disease (PD) are prominent examples of NDs. While many symptoms of these diseases are treatable, a definitive cure remains elusive, emphasizing the urgent need for more effective treatments and early interventions.

Diagnosing NDs in their early stages is crucial; failure to do so can lead to severe outcomes, including death. To handle this growing crisis, it is vital to prioritize methods for both diagnosis and treatment.

Reliable biomarkers are urgently needed that can, first, identify NDs in their early stages and indicate susceptibility to these conditions. Second, these biomarkers could assist in clinical diagnosis and help define the severity of the disease.

PD is rapidly increasing, damaging the health of millions of people worldwide. According to the World Health Organization,[2] the prevalence of PD has doubled over the past 25 years, with global assessments in 2019 indicating that more than 8.5 million people were suffering from the condition. The rates of disability and death associated with PD are rising faster than those for any other ND. In 2019, PD caused around 329,000 deaths, more than twice the number recorded in 2000.

Statistics[3] show that in the United States of America (USA), approximately one million people are affected by PD, with projections indicating that this figure will rise to 1.2 million by 2030. A similar trend is expected in countries such as India. The rapid increase in PD cases presents significant economic challenges, placing a burden on both the economy and healthcare systems. In addition, PD creates social obstacles within communities.[4] Given the profound economic, social, and personal impacts of PD, it is crucial to optimize strategies for controlling this disease, particularly through early diagnosis.

As noted earlier, biomarkers are essential for diagnosing NDs and can facilitate the early diagnosis of PD. Artificial intelligence (AI), machine learning (ML), and deep learning (DL) methods have seen significant advancements over the past decade. This study provides insights into research linking various types of biomarkers with AI, ML, and DL methods for PD diagnosis. Given that AI encompasses both ML and DL, the term AI is used broadly in the remainder of this study.

### 1.1. Motivation factors and research challenges

Clinicians face several challenges when diagnosing and treating PD, mainly related to precision, cost, and delays.

(i) Precision: identifying the complex clinical signs of PD, especially in its early stages, is challenging due to the similarity between PD and AD symptoms. This makes it difficult for doctors to differentiate between these two NDs. By the time PD is accurately diagnosed, it may be too late to implement effective control measures.
(ii) Delays: diagnosing PD is a lengthy process, involving multiple steps from data gathering to analysis. The time gap between data collection and interpretation leads to delays in reaching a diagnosis and initiating treatment.
(iii) Cost: the high costs associated with diagnostic and treatment procedures can restrict access to healthcare facilities, especially for those in need of early intervention.

Continuous innovation and the development of more efficient methods are important to overcome these challenges, achieving precise results at a lower cost and in a timely manner. The literature highlights the significant potential of AI techniques in addressing these issues. AI methods can quickly process and analyze data, reducing diagnostic delays and enabling early-stage treatment. In addition, AI reduces the need for manual intervention, thereby increasing diagnostic efficiency and lowering overall treatment costs.

### 1.2. Contribution

To the authors' knowledge, this study is the first comprehensive review of AI applications using four major types of biomarkers—clinical, neuroimaging, biochemical and proteomic, and genetic—for the early diagnosis of PD.

The key contributions of this review include:

(i) Systematically exploring the practicalities of AI approaches to enhance the efficiency and precision of PD diagnosis, management, and treatment.
(ii) Analyzing the AI methods and techniques used in diagnosing and treating PD.
(iii) Stratifying different biomarkers based on the available AI approaches.

### 1.3. Organization of the review

The organization of the remainder of the review is as follows: Section 2 discusses the background and various aspects of PD, along with relevant biomarkers and AI approaches. Section 3 elaborates on the research methodology used in this study, followed by findings from the literature related to the application of AI approaches in diagnosing, treating, and managing PD using different types of biomarkers. Section 4 discusses the challenges and future directions for applying AI in the medical field, followed by conclusion in Section 5.

## 2. Background

### 2.1. PD process

PD is triggered by the degeneration of midbrain dopaminergic neurons (mDANs), which are crucial for controlling movement.[5] PD symptoms are categorized as motor and non-motor, and they develop gradually and worsen with age, significantly impacting the quality of life.[6] As PD progresses, it results in mental, physical, social, and emotional challenges. Early and accurate diagnosis is crucial before these symptoms become more severe. Other causes of PD include genetic and environmental factors[7] as well as the presence of Lewy bodies.[8]

### 2.2. PD treatment

PD is a heterogeneous disorder that affects individuals' quality of life. While there is no curative therapy for PD, treatments are available. These treatments focus on improving the patient's quality of life and managing symptoms. Available options for treatment include[8] surgeries, therapies, medications, lifestyle adjustments, and supportive care.

Each PD patient is treated according to the severity of their condition, which is classified into five stages (Stage 1 to Stage 5). Treatment plans are tailored to the individual's specific needs and disease progression. Ongoing research is exploring new therapies, such as stem cell treatments and gene therapy, to improve outcomes and slow disease progression.[8] Recent advancements in PD treatment focus on improving symptom management, developing disease-modifying therapies, and incorporating new technologies based on AI methods. AI can use patient data (e.g., genetics, lifestyle, disease severity) to recommend personalized therapies, improving treatment efficacy and reducing medication side effects.

Another increasingly recognized class of PD biomarkers is known as "gut biomarkers." Although gut biomarkers span traditional biomarker categories, such as biochemical, genetic, microbiome, metabolomic, and even clinical and imaging domains, their focus on the gastrointestinal system and the gut-brain axis makes them a distinct subclass with unique diagnostic and pathogenic significance.

### 2.3. Biomarkers

According to the U.S. Food and Drug Administration and the National Institutes of Health,[21] "A biomarker is not intended to measure how an individual feels, functions, or survives. Instead, it is a defined characteristic that serves as an indicator of normal biological processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic interventions." Different researchers have classified biomarkers into various categories. Some studies[22] have focused on clinical, biochemical, and neuroimaging biomarkers, while Surguchov[23] identified clinical, imaging, pathological, biochemical, and genetic markers as potential PD biomarkers. After reviewing the literature, we identified four major categories: clinical, neuroimaging, biochemical and proteomic, and genetic biomarkers. A summary of these categories is provided in Table 1.

Research on PD biomarkers is rapidly evolving with significant advancements in areas such as genetics, proteomics, neuroimaging, gut microbiome (GM),[24] and exosomal analysis.[16] The integration of AI methods with biomarkers offers greater precision in tracking PD progression and identifying novel targets for early intervention. In the following section, we explore research that utilizes various biomarkers in conjunction with AI-based techniques.

### 2.4. Integration of AI approaches

ML is a branch of AI focused on statistical models and algorithms that can make computers to enhance their performance on a task by learning from data. DL, a specialized subset of ML, uses artificial neural networks (ANNs) with multiple layers (hence "deep") to analyze various forms of data.[25] DL algorithms are robust and efficient, and are hence employed in medical imaging to address diagnostic issues.[26]

## 3. Related work

This section reviews selected research articles that focus on the application of AI in the diagnosis, management,

**Table 1. Parkinson's disease biomarkers and their description**

| Type of biomarker | Description |
|---|---|
| Clinical[9-12] | These biomarkers are related to clinical signs and symptoms that aid in diagnosis and tracking disease progression. Non-motor signs include cognitive decline, sleep disturbances, fatigue, excessive sweating, depression, idiopathic rapid eye movement sleep-behavior disorder, and constipation. Motor signs include tremors, rigidity, akinesia, and postural instability as well as two motor subtypes: postural instability and gait difficulty, and axial symptoms. Bradykinesia and stiffness exhibit the strongest correlation with the degeneration of dopamine-producing neurons in the nigrostriatal system. |
| Neuroimaging[13,14] | These biomarkers use imaging techniques, such as MRI, SPECT, and PET, to detect structural and functional changes in the brain, particularly in regions such as the substantia nigra and dopamine transporter systems. |
| Biochemical[15-17] and proteomic[16,18,19] | Biochemical biomarkers encompass a wide range of molecules and processes such as proteins, metabolites, and cellular changes. Proteomic biomarkers focus specifically on proteins involved in the disease process. These biomarkers are found in cerebrospinal fluid, blood, and saliva. α-syn, dopamine metabolites (HVA, DOPAC), and neuroinflammatory markers are among the most studied. |
| Genetic[16,20] | Genetic biomarkers are relevant to both familial and sporadic forms of PD. Mutations in genes, such as *PRKN*, *LRRK2*, and *GBA*, are recognized as significant genetic contributors to the disease. |

Abbreviations: DOPAC: 3,4-dihydroxyphenylacetic acid; HVA: Homovanillic acid; MRI: Magnetic resonance imaging; PET: Positron emission tomography; SPECT: Single-photon emission computerized tomography; α-syn: α-synuclein.

and treatment of PD. The aim is to provide an overview of recent advancements and practical implementations of AI methods within the field. The articles are analyzed based on the methodologies employed and the results obtained across various biomarker types. This structured comparison allows readers to gain a deeper comprehension of how AI techniques and biomarkers are being integrated in efforts to diagnose, manage, and treat PD effectively.

### 3.1. Methodology

This review aims to examine AI-based approaches for evaluating different types of biomarkers and to provide a comprehensive understanding of their role in PD diagnosis. The research questions guiding this study are presented in Table 2. To conduct this review, we identified, filtered, and assessed relevant research from recent years, focusing on the utilization of AI approaches in diagnosing, managing, and personalizing PD treatment using various types of biomarkers.

### 3.2. Application of AI methods for PD: an overview

The use of AI approaches has considerably improved the accuracy and speed of PD identification and management. These methods can rapidly and precisely analyze large volumes of medical data, including diverse biomarkers, such as brain images and clinical information, by identifying complex patterns that assist in early PD diagnosis.[26] In addition, AI methods are effective in predicting patients' responses to different types of treatments such as deep brain stimulation.[27] Predictive disease progression algorithms can also aid in developing personalized treatment plans. In addition, by modeling the effects of new treatments and refining treatment protocols, these approaches can expedite the development of new drugs and therapeutic strategies, thereby enhancing their effectiveness.[26] In this section, we assess the applicability of AI methods to PD-related studies that utilize various types of biomarkers.

#### *3.2.1. Clinical biomarkers and AI methods*

Clinical biomarkers encompass both motor and non-motor features of PD.[22] Clinical rating scales are utilized to identify these features and assist in early diagnosis. Some commonly used scales include the Hoehn and Yahr scale (H&Y), the Non-Motor Symptoms Scale for PD, and the Unified PD Rating Scale (UPDRS).[22]

As previously mentioned, AI methods are crucial for PD diagnosis. In this section, we explore research that integrates various clinical biomarkers alongside AI-based techniques for the detection, monitoring, and classification of PD patients from healthy individuals.

An automated approach based on a DL method was proposed for PD detection and severity prediction by analyzing gait data.[28] The study involved 166 participants, including 93 PD patients and 73 control subjects. That method achieved 98.7% accuracy for PD gait recognition and 85.3% for predicting PD severity.

Goyal and Rani[29] conducted a comparative analysis of ML, ensemble learning (EL), and DL classifiers to develop classification models for PD detection using 252 voice samples (188 PD and 64 healthy). The random forest (RF) model, with 82.37% accuracy, gave better performance than other base classifiers. The DL model achieved training

**Table 2. Research questions**

| Item | Question | Description |
|---|---|---|
| 1 | What types of biomarkers are used by researchers to diagnose, monitor, and treat PD? | This question aims to identify the categories of biomarkers used in the diagnosis, monitoring, and treatment of PD. |
| 2 | Which AI methods are used to classify PD subjects from healthy subjects? | This question explores which methods—AI, ML, or DL—are employed by researchers to distinguish PD subjects from healthy individuals. |
| 3 | What performance parameters are reported in the research manuscripts? | This question seeks to identify the performance metrics used to evaluate the effectiveness of AI methods in classifying and diagnosing PD subjects compared to healthy individuals. |
| 4 | How do AI methods offer new insights into the PD diagnosis and treatment process? | This question investigates how AI methods contribute to new insights in the diagnosis and treatment of PD. Due to their data-driven nature, these approaches can significantly improve early detection, monitoring, and personalizing treatment. |
| 5 | What strategies can be recommended to improve the efficiency and effectiveness of these approaches in diagnosing and treating PD? | This question aims to identify practical strategies to enhance the performance and impact of AI methods in PD diagnosis and treatment. |
| 6 | How many subjects were evaluated in each study? | This question examines the sample size in each study, as the performance of AI methods can vary based on dataset size. Understanding the number of subjects helps assess the context and reliability of the model's performance. |

Abbreviations: AI: Artificial intelligence; DL: Deep learning; ML: Machine learning; PD: Parkinson's disease.

and testing accuracies of 91.33% and 85.02%, respectively, showing comparable performance to traditional ML classifiers, despite the small dataset. Among EL classifiers, the Light Gradient Boosting Machine model showed the highest accuracy (85.90%), outperforming both ensemble and base models.

Senturk[30] employed ML algorithms to detect PD using voice features from 31 subjects (23 PD and eight healthy). Using support vector machines (SVM) with the recursive feature elimination (RFE) algorithm, the study achieved a diagnostic accuracy of 93.84% with a minimal number of voice features.

Ouhmida *et al.*[31] proposed an approach for PD detection using voice features and ML algorithms on Max Little's University of California, Irvine (UCI) dataset, consisting of 195 samples from 31 subjects (23 PD and eight healthy). The study applied K-nearest neighbors (KNN), SVM, and decision tree (DT) classifiers, combined with two feature subset selection (FSS) techniques: Minimum Redundancy Maximum Relevance (mRMR) and ReliefF. The KNN algorithm outperformed the other, achieving 98.26% area under the curve (AUC), 97.22% sensitivity, 100% specificity, and 97.92% accuracy.

Chintalapudi *et al.*[32] highlighted the contribution of ML methods to improving disease prediction accuracy. Using voice data from 31 subjects (23 PD and eight healthy) in the UCI dataset, they tested three DL models—recurrent neural networks (RNN), multilayer perceptron (MLP), and long short-term memory (LSTM). The LSTM model outperformed the others, achieving 99% accuracy.

Ferreira *et al.*[33] applied the Naïve Bayes (NB) algorithm to spatiotemporal gait parameters collected from 126 participants (63 idiopathic PD [iPD] and 63 healthy), achieving 84.6% classification accuracy, 80.0% recall, and 92.3% precision. For PD stage identification, the RF model yielded an AUC-receiver operating curve (ROC) of 78.6%.

Sigcha *et al.*[34] employed consumer smartwatches equipped with inertial sensors and ML methods to detect and assess bradykinesia in the upper limbs. Thirteen participants (six PD and seven age-matched controls [AMC]) wore smartwatches while performing motor tasks over a minimum period of six weeks. A combination of convolution neural networks (CNN) and RF models produced promising results, achieving 86% accuracy and 94% AUC. This approach offers an unobtrusive and effective method for detecting and evaluating the severity of bradykinesia.

Thakur *et al.*[35] implemented and compared ML algorithms for PD diagnosis using a voice dataset consisting of 252 subjects (188 PD [107 males and 81 females] and 64 healthy [23 males and 41 females]), with 756 instances and 754 features, sourced from www.kaggle.com. The authors utilized SVM, RF, DT, and extra trees (ET) classifiers for PD diagnosis. The best performance was achieved using the ET classifier, with an accuracy of 94.34%, precision of 93.88%, F1-score of 96.84%, and a recall of 100%.

Trabassi *et al.*[36] identified the most accurate supervised ML algorithm for classifying PD subjects from speed-matched healthy individuals using a minimal set of gait features derived from inertial measurement units. Data from 161 subjects (81 PD and 80 healthy) included 22 gait features extracted from trunk acceleration patterns. After applying a three-level FSS, seven gait features were utilized to execute five ML algorithms - DT, SVM, KNN, ANN, and RF. Among these, on test dataset, DT, SVM, and RF showed the highest prediction accuracies exceeding 80%.

Alalayah *et al.*[37] used a dataset comprising 195 voice signals (48 healthy and 147 PD) and applied ML classification algorithms for early PD detection. Their method (RF with t-distributed stochastic neighbor embedding) outperformed existing studies, achieving 97% accuracy, 96.50% precision, 94% recall, and 95% F1-score. In addition, the MLP and principal component analysis (PCA) algorithm achieved 98% accuracy, 97.66% precision, 96% recall, and 96.66% F1-score.

Govindu and Palwe[38] employed four ML methods for the early detection of PD based on voice data from 31 subjects. They found that the RF classifier provided the best PD classification results, with 91.83% accuracy, 95% sensitivity, 86% recall, and an ROC-AUC of 70.1%, using the Multidimensional Voice Program dataset consisting of 195 records with 22 features.

Martinez-Eguiluz *et al.*[39] evaluated nine ML algorithms to differentiate PD subjects from controls using non-motor attributes. The data were sourced from the Biocruces database (96 subjects: 59 PD and 37 healthy) and the Parkinson's Progression Markers Initiative (PPMI) (687 subjects: 490 PD and 197 healthy). Most ML algorithms achieved accuracy exceeding 80%, with SVM and MLP performing best, at 86.3% and 84.7% accuracy, respectively.

Yadav *et al.*[40] introduced an approach to assess the stage and severity of PD using a voice dataset comprising 31 subjects (23 PD and eight healthy), 23 features, and 197 instances, leveraging AI algorithms. The study indicated that the DT classifier achieved 94.87% accuracy and gradient boosting classifier has an AUC of 98.7%.

Goyal *et al.*[41] proposed a Repetitive Pointing Task to evaluate upper-limb bradykinesia utilizing data collected through a 3D motion capture system for nine healthy and 17 PD subjects. Using just six features, their proposed ML model achieved 97.14% accuracy, 97% sensitivity, 95%

specificity, 97% precision, and a 97% F1-score. The authors emphasized that their method is a fast, reliable, and non-invasive for assessing bradykinesia and estimating PD stages.

Faiem *et al.*[42] used a gait dataset from the "PhysioNet" repository, consisting of 93 PD and 73 control subjects, and implemented a perceiver-based multimodal ML framework to predict UPDRS scores for PD patients. Their model achieved a root mean square error of 5.75 ± 4.16, a mean absolute error (MAE) of 2.23 ± 1.31, and a linear correlation coefficient (CC) of 0.93 ± 0.08. These values are better than previous studies in both MAE and CC, highlighting the effectiveness of their multimodal approach.

Palakayala and Kuppusamy[43] demonstrated that PD can be identified with high precision by analyzing non-motor PD features using ML algorithms. To support this, they collected data from 212 participants (106 healthy and 106 PD), who underwent the PD Sleep Test, the Hopkin's Verbal Learning Test, and the Clock Drawing Test from the PPMI database. The study showed that all applied ML classification algorithms—RF, NB, SVM, and logistic regression (LR)—achieved accuracies exceeding 73% ± 8.4% with individual datasets, while an accuracy of 98% ± 0.6% was achieved using a custom hybrid dataset.

Salsone *et al.*[44] investigated the performance of ML models—LR, SVM, RF, and extreme gradient boosting (XGB)—to classify idiopathic REM sleep-behavior disorder (iRBD) patients who exhibited periodic leg movements (PLMS) from those who did not. The study utilized heart rate variability data from 42 consecutive iRBD subjects (19 without PLMS and 23 with PLMS). Their findings demonstrated that the RF model achieved 86% accuracy, 74% specificity, and 96% sensitivity. XGB (accuracy = 78%, specificity = 72%, sensitivity = 83%) and SVM (accuracy = 81%, sensitivity = 83%, specificity = 79%) also performed well. LR exhibited the lowest performance with 71% accuracy.

Wang *et al.*[45] introduced a hybrid signal processing and ML-based gait classification system to detect gait anomalies and assess PD severity levels. Five different ML classifiers were employed for anomaly detection and severity rating as defined in H&Y scale. To verify the system's effectiveness, the "Physionet" gait database, comprising data from 93 individuals with iPD and 73 AMCs, was utilized. Employing a 10-fold cross-validation method, the SVM classifier achieved the highest accuracy, reporting 98.20% for anomaly detection and 96.69% for severity level assessment.

Byeon[46] developed an SVM-based framework to predict depression in PD (DPD) using National Parkinson's Registry data from 223 subjects (130 without depression and 93 with DPD) out of a total of 335. The model incorporated predictors such as health habits, PD symptoms, sociodemographic factors, sleep behavior syndromes, and neuropsychological indicators. Comparing the prediction accuracy of eight different SVM models, the study found that Gaussian Kernel-based Nu-SVM achieved the highest performance, with 96.0% sensitivity, 93.3% specificity, and 95% overall accuracy. In contrast, the polynomial-based C-SVM reached the maximum sensitivity (100%) but had the least specificity (20%) and an overall average accuracy of 70%.

Lee and Ham[47] conducted a review of ML advancements for early depression diagnosis, analyzing 32 original studies out of 120 identified in the Web of Science. They highlighted that various ML methods are suited to different data types. For instance, LR, RF, SVM, and ANN are effective for numeric data, while RF is particularly suitable for genomic data. The reported performance metrics varied widely, with accuracy scores ranging from 60.1% to 100.0% and AUC scores from 64.0% to 96.0%. The study concluded that ML is a valuable tool for the early diagnosis of depression.

Pereira *et al.*[48] explored the relationship between gut bacteria, serum metabolites, and clinical features in 124 subjects (63 PD and 61 healthy). They identified 139 metabolite features that distinguished PD from healthy subjects; however, no associations were found between clinical features within the PD group and metabolic attributes. Using SVM with a radial basis function kernel, they achieved 81% accuracy with gas chromatography-mass spectrometry (GC-MS) data and 72% and 77% accuracy with liquid chromatography-mass spectrometry (LC-MS) in negative and positive ionization modes, respectively.

Li *et al.*[49] presented an AI-based approach for assessing PD grades, aiming to improve standardization and accuracy, avoiding the challenges related to wearable sensors. A dataset of 110 videos was collected from various people with different PD severity levels. Employing MediaPipe, 19 distinct kinematic features were extracted from joint movements, generating real-time kinematic data. Among five different ML algorithms (SVM, Gradient Boosting DTs, KNN, MLP, and RF) tested, KNN delivered the highest overall accuracy of 96.63% and achieved 100% accuracy in distinguishing PD grade 4 from grade 5 subjects.

Lu *et al.*[50] presented a method for extracting time-frequency-based statistical features from dynamic handwriting patterns, focusing on their temporal and frequency characteristics for PD detection. This approach was evaluated using the Cc-PhD dataset (97 subjects:

31 PD, 31 Essential Tremor, and 35 healthy) and the Parkinson's Disease Handwriting Database (PaHaW) dataset (75 subjects: 37 PD and 38 healthy). An Escape Coati Optimization Algorithm was applied to optimize the parameters of the AdaBoost classifier after FSS was performed using the RF algorithm. The method achieved 97.95% and 98.67% accuracy; 98.15% (average) and 97.78% sensitivity; 99.17% (average) and 100% specificity; and AUC scores 98.66% (average) and 98.89% on the Cc-PhD and PaHaW datasets, respectively.

These findings collectively demonstrate that AI methods can effectively utilize clinical biomarkers to accurately differentiate PD subjects from healthy individuals. The results highlight the potential of such methods to serve as valuable tools in clinical settings, supporting early diagnosis, continuous monitoring, and personalized treatment strategies for PD patients.

#### *3.2.2. Neuroimaging biomarkers and AI methods*

In this section, we explore the research conducted so far that utilizes various neuroimaging biomarkers alongside AI methods for detecting, monitoring, and classifying PD subjects from healthy individuals.

Castillo-Barnes *et al.*[51] used ML methods to classify PD and healthy subjects using a balanced set of 386 single-photon emission computerized tomography (SPECT) scans, from 386 subjects (193 PD [127 males and 66 females] and 193 healthy [128 males and 65 females]), from the PPMI. FSS was performed using a Mann-Whitney-Wilcoxon U-test, and classification was carried out using an SVM approach. The authors achieved a balanced accuracy of 97.04% using 10-fold cross-validation, demonstrating the efficacy of the SVM-based framework for distinguishing PD subjects from controls.

Chakraborty *et al.*[52] applied four ML algorithms for the detection of PD based on 3T T1- magnetic resonance imaging (MRI) scans from 906 subjects (203 control, 66 prodromal, and 637 PD). They reported 95.3% accuracy, 97.28% precision, 95.41% recall, and a 94% F1-score using an ANN (specifically an MLP) for PD detection.

In a separate study, Chakraborty *et al.*[53] applied a 3D CNN for PD detection on 3T T1w-MRI scans from 406 subjects (203 PD and 203 healthy) and achieved 95.29% accuracy, 94.3% average recall, 92.7% average precision, 94.30% average specificity, a 93.6% F1-score, and 98% ROC-AUC.

Huang *et al.*[54] evaluated a dataset of functional brain images from 202 subjects (six healthy and 196 PD). They applied various prediction approaches, including multivariate statistical analysis, EL models, and deep CNNs, to predict PD stages. The VGG16 deep CNN model achieved 92.2% training accuracy, 64.9% test accuracy, and 57.6% test F1-score.

Magesh *et al.*[55] employed ML methods for early PD diagnosis using 642 (430 PD and 212 non-PD) SPECT dopamine transporter scans from the PPMI. Using a VGG16 deep CNN, they achieved 95.2% accuracy, 97.5% sensitivity, and 90.9% specificity for classifying PD and non-PD subjects.

Solana-Lavalle and Rosas-Romero[56] conducted MRI-based PD detection by using voxel-based morphometry and seven ML classifiers on a dataset of 480 MRI images (226 PD males, 86 healthy males, 104 PD females, and 64 healthy females) from the PPMI. In male subjects, the NB classifier with 1.5T scanner achieved 99.01% accuracy, 100% precision, 100% specificity, while SVM with 3T scanner achieved 99.35% sensitivity. For female subjects, the logistic classifier with 1.5T scanner achieved 96.97% accuracy, 97.22% precision, and 96.15% specificity, Bayesian network with 1.5T scanner achieved 100% sensitivity, while MLP with 1.5T scanner achieved 96.15% specificity. These results highlight the effectiveness of ML algorithms in detecting PD across genders.

Shu *et al.*[57] examined 144 subjects (72 subjects with PD progression and 72 with stable PD) for predicting disease progression using T1-weighted MRI scans and ML. Their proposed joint model achieved an AUC of 83.6%, compared to 79.5% and 55.0% for the radiomics signature and UPDRS score, respectively. Sensitivity values of 80.5%, 87.5%, and 29.2%, and specificity values of 72.2%, 69.7%, and 86.1% were also reported. For Stage 1 PD, the model achieved 82.7% predictive accuracy, 82.9% sensitivity, and 70.2% specificity; for Stage 2, it achieved 85.4% accuracy, 96.0% sensitivity, and 60.0% specificity.

Veetil *et al.*[58] analyzed 242 MRI samples (150 PD and 92 normal control [NC] subjects) to classify PD and NC subjects using five deep neural network architectures: Xception, DenseNet201, VGG16, VGG19, and ResNet50. The highest accuracy of 92.60% was achieved using VGG19, along with an F1-score of 92.3% for NC and 92.9% for PD. The authors emphasized that AI-based tools are highly effective in supporting early risk assessments and serving as decision-support systems in PD medical imaging.

Guo *et al.*[59] classified early-stage PD using resting-state functional MRI (rs-fMRI) data from 84 subjects (28 in Stage 1 and 56 in Stage 2) from the PPMI. The LSTM model achieved an accuracy of 71.63%, which was 11.56% higher than the CNN and 13.52% higher than the best-performing traditional ML model. These results demonstrated a considerable enhancement in accuracy and robustness compared to other ML classifiers.

Tomer *et al.*[60] compared ML methods for PD detection using T1-weighted MRI scans from 20 individuals, including both PD and NC groups. Of the 968 pre-processed images, only 848 were used for analysis. For feature extraction, the gray level co-occurrence matrix (GLCM) method achieved 90.5% accuracy, surpassing PCA, which yielded 87.5% accuracy.

Vyas *et al.*[61] used 318 brain images from MRI scans and applied both 2D and 3D CNN models for PD early detection. The 3D CNN model demonstrated 88.9% accuracy with an AUC of 86% on the test data, whereas the 2D CNN model showed 72.22% accuracy with an AUC of 50%.

Camacho *et al.*[62] trained a 3D CNN model to detect PD using 2041 T1-weighted MRI scans (1024 PD and 1017 healthy) collected from 13 different studies. Their model achieved 79.3% accuracy, 77.7% sensitivity, 81.3% specificity, 80.2% precision, and an AUC-ROC of 87%.

Erdaş and Sümer[63] developed a fully automated approach utilizing 1130 T1-weighted MRI scans (259 healthy and 871 PD) for detecting and predicting PD severity. Their method employed DL techniques, specifically 2D and 3D CNNs. The 3D CNN model achieved 96.20% accuracy, 95.36% recall, 94.52% F1-score, and 94.07% precision.

Khachnaoui *et al.*[64] proposed a computer-aided diagnosis system for PD using pre-trained CNN models, the bilinear pooling method, and the transfer learning (TL) technique, based on 2720 SPECT images (1360 PD and 1360 healthy) from the PPMI. An accuracy of 98.47% was achieved by using the Bilinear CNN EfficientNet-B0-MobileNet-V2 model. The authors concluded that their method supports accurate PD diagnosis without relying on subjective factors.

Wang *et al.*[65] employed a DL model to analyze quantitative susceptibility maps and T1-weighted images for distinguishing PD patients from healthy subjects. They used two datasets: Dataset 1 with 379 subjects (92 PD and 287 healthy), and Dataset 2 with 155 subjects (83 PD and 72 healthy). In the internal testing sample, the model achieved an AUC of 90.1%, 92.0% accuracy, 83.3% sensitivity, and 94.7% specificity. In the external testing sample, it achieved 84.5% AUC, 78.7% accuracy, 77.1% sensitivity, and 80.6% specificity.

Praneeth *et al.*[66] proposed a technique for classifying PD by using a deep residual CNN combined with the Enhanced Whale Optimization Algorithm to improve classification accuracy. Using 591 diffusion-weighted and T1-weighted MRI scans from the PPMI (412 PD and 179 healthy), they achieved 98.87% accuracy, 97.02% precision, 96.87% sensitivity, and 98.13% specificity in distinguishing PD from healthy subjects.

Ahalya *et al.*[67] proposed an automated PD detection method based on CNN and quantum SVM, using 1000 MRI images, including 60 real-time MRIs (30 healthy and 30 PD). Their hybrid model achieved 87.5% prediction accuracy, 84% recall, 95% precision, and an F1-score of 89%.

Islam *et al.*[68] aimed to analyze PD by applying ML and TL techniques to clinical assessment data and 3D T1-weighted MRI samples. They used two datasets: 1277 clinical records (155 PD and 1122 healthy) and 2500 usable MRI samples from the PPMI. Using the ET classifier, an accuracy of 98.44%, 97.11% precision, 99.02% recall, and a 98.06% F1-score, was attained. In addition, implementing DenseNet169 on the MRI dataset resulted in an optimal accuracy of 85.08%.

Patil and Ford[69] proposed a decorrelated CNN framework to recognize PD using rs-fMRI data. The proposed framework was applied to two datasets: a single-scanner PPMI imbalanced dataset (183 subjects: 164 PD and 19 healthy) and a multi-scanner dataset. After pre-processing, the multi-scanner dataset was formed by combining rs-fMRI data from 215 healthy subjects in the frontotemporal lobar degeneration neuroimaging initiative with those obtained from the PPMI. The model achieved 77.80% accuracy on the multi-scanner dataset, outperforming the single-scanner model.

Redhya and Jayalakshmi[70] proposed an ensembled grid based ML model, for MRI-based PD classification, using 260 MRI images (134 PD and 126 healthy) from the PPMI. The stacking classifier, which combines XGB, SVM, and RF classifiers, achieved 98.41% accuracy in distinguishing PD from healthy subjects.

Zhang *et al.*[71] analyzed T1-weighted MRI images and clinical data from 272 PD subjects in the PPMI, alongside 45 PD subjects from the National Alzheimer's Coordinating Center dataset, to identify depression subtypes in PD subjects using ML methods. By employing PCA and four other unsupervised clustering algorithms, they observed that "Partitioning Around Medoids" outperformed "Gaussian Mixture Model," hierarchical clustering, and K-means with two clusters. The sensitivity, specificity, and AUC in the high-risk testing subtype were 78.6%, 81.5%, and 81%, respectively. The model based on non-high-risk subtypes had an AUC of 85.9%, sensitivity of 65.4%, and specificity of 85.2%.

These findings collectively demonstrate that AI techniques can effectively leverage neuroimaging biomarkers to distinguish PD from healthy subjects with

high accuracy. The results from this research support the potential for such methods to become integral tools in clinical settings, aiding early diagnosis, continuous monitoring, and personalized care for individuals with PD.

#### *3.2.3. Biochemical and proteomic biomarkers and AI methods*

Biochemical biomarkers for PD comprise a broad category of markers that indicate the presence or progression of the disease. These biomarkers include various molecules, such as proteins, hormones, metabolites, lipids, and neurotransmitters. Proteomic biomarkers, a significant subset of biochemical biomarkers, refer specifically to proteins; however, the broader biochemical biomarker category includes other molecules that offer meaningful insights into the disease.[72]

Proteins such as α-synuclein (α-syn), neurofilament light chain, and DJ-1 serve as markers of PD progression. These are analyzed using proteomic techniques such as mass spectrometry and protein assays to understand their roles in the disease. α-syn is particularly crucial to PD pathology due to its misfolding and aggregation, which leads to the formation of Lewy bodies,[17] a hallmark of PD.[73]

It is important to note that while all proteomic biomarkers are biochemical, the reverse is not true. For example, neurotransmitters such as dopamine and its metabolites are widely used as biochemical biomarkers in PD research, but they fall outside the scope of proteomics. These molecules are essential in understanding PD pathophysiology and are typically analyzed using techniques such as liquid chromatography combined with tandem mass spectrometry.[16]

In this section, we examine research studies that have explored the use of biochemical and proteomic biomarkers, alongside AI methods, for detecting, monitoring, and classifying PD subjects from healthy individuals.

Lin *et al.*[74] developed ML algorithms utilizing blood-based biomarkers to identify subjects affected by AD, PD, and frontotemporal dementia (FTD). Plasma samples from 377 subjects were analyzed, including 97 healthy, 76 subjects on the AD spectrum (41 mild cognitive impairment [MCI] and 35 with AD), 173 on the PD spectrum (57 with normal cognition, 29 with MCI, and 87 with PD dementia), and 31 with FTD. Plasma levels of α-syn, amyloid beta (Aβ) 42, total tau, Aβ40, and phosphorylated Tau181 were measured. The developed linear discriminant analysis (LDA) model combined with a RF classifier achieved a 76% accuracy in distinguishing AD, PD, and FTD, and 63% and 83% accuracy in distinguishing disease severity in the PD and AD spectrums, respectively.

Maass *et al.*[75] validated a predictive SVM model for classifying PD and AMC based on cerebrospinal fluid (CSF) bioelement levels. The study included 157 subjects (82 PD, 68 AMC, and seven normal pressure hydrocephalus), achieving an AUC-ROC of 76%, sensitivity of 80%, and specificity of 83% on a new dataset, without incorporating additional features.

Wang *et al.*[76] introduced a DL technique to detect early PD by using pre-motor features, such as iRBD, CSF biomarkers, olfactory loss, and mDAN imaging markers. By comparing their DL model with 12 ML and EL methods on a dataset of 584 subjects (183 healthy and 401 early PD), they found that their framework achieved the highest average accuracy of 96.45%.

Chung *et al.*[77] examined the impact of plasma extracellular vesicles (EV)-borne tau and Aβ1-42 as biomarkers for cognitive deficit in PD, using a dataset of 162 subjects (46 healthy and 116 PD). Subjects were classified according to cognitive function. Using an ANN, their model achieved 91.3% accuracy in detecting cognitive dysfunction in PD patients. Aβ1-42 and plasma EV tau were found to be the most significant factors.

Vacchi *et al.*[78] developed a two-level RF model to distinguish PD from atypical parkinsonisms (AP) using CSF-derived EVs and immune profiling of plasma-derived EVs. The Level 1 "basic" model, applied to 84 subjects (29 PD, 36 healthy, nine multiple system atrophy, and 10 AP-TAU), achieved 92.9% accuracy, 100% sensitivity, and 83.3% specificity in classifying subjects with NDs from healthy individuals. The Level 2 "integrated" model, trained on 54 subjects (48 patients and six healthy), achieved 92.6% accuracy and 96.6% sensitivity in distinguishing PD from healthy subjects. The advanced RF model also performed better in distinguishing AP-Tau, achieving 92.6% accuracy and 70.0% sensitivity compared to the basic model.

Amboni *et al.*[79] aimed to identify important features related to PD with MCI (PD-MCI) using an ML approach. A total of 75 PD subjects (42 without PD-MCI and 33 with PD-MCI) were evaluated through neuropsychological and clinical assessments. Two ML-based models were created: Model 1 combined age, gait, and clinical features, while Model 2 had two variants—Model 2A and Model 2B. Model 2A used mean standardized uptake values of nine brain areas along with the top five features identified in Model 1. Model 2B focused on cortical regions combined with those same top five features. The best performing classifiers in Model 1 were SVM (accuracy = 80.0%, AUC-ROC = 79.2%, sensitivity = 72.7%, specificity = 85.7%) and RF (accuracy = 73.3%, AUC-ROC = 72.2%, sensitivity = 66.7%, specificity = 78.6%). In Model 2A, the

highest performer was SVM (accuracy = 72.2%, specificity = 70.6%, sensitivity = 73.7%, AUC-ROC = 72.1%) and in Model 2B, SVM (accuracy = 75.0%, sensitivity = 73.7%, specificity = 76.5%, AUC-ROC = 75.1%) outperformed other classification algorithms. Overall, Model 1 provided the highest accuracy, and SVM consistently outperformed other classifiers.

Chen *et al.*[80] developed a predictive model for assessing cognitive deterioration in 42 PD by using ML methods. For each participant, three plasma biomarkers and 29 clinical variables were collected, along with neuropsychological test results. ML techniques, including SVM and PCA, were employed to build a cognitive classification model. Using 32 predictive features, the PCA-SVM classifier achieved an accuracy of 92.3% and an AUC of 92.9%. When only 13 carefully selected features were used, the accuracy and AUC both increased to 100%.

Dadu *et al.*[81] applied supervised and unsupervised ML methods to comprehensive and longitudinal clinical data from the PPMI, which included 294 subjects, to predict PD progression and uncover distinct patient subtypes. An independent dataset consisting of 263 clinically well-characterized cases from the PPMI was used to validate the models. The authors made predictions of PD progression over five years following initial diagnosis, achieving average AUCs of 95% ± 2% for fast-progressors, 87% ± 3% for moderate progressors, and 92% for the slow progressors. They also recognized serum neurofilament light as a crucial biomarker for rapid PD progression, along with several other significant indicators of disease progression.

Harvey *et al.*[82] provided an approach to predict cognitive outcomes in newly diagnosed PD subjects, from the PPMI, by developing a multivariate ML model. The dataset included 67 with normal cognition, 39 with PD-MCI, 43 with PD dementia, and 60 with subjective cognitive decline. Four ML methods were evaluated: RF, conditional inference forest (Cforest), SVM, and ElasticNet. For the cognitive impairment model, the following results were obtained: the combined model (clinical and biological data) achieved 86.7% accuracy, 71.9% sensitivity, 93.8% AUC, and 96.1% specificity using Cforest (28 variables). The clinical features model using Cforest (11 variables) yielded 85.5% accuracy, 65.6% sensitivity, 93.0% AUC, and 98.0% specificity. The biofluid model using ElasticNet (four variables) achieved 68.7% accuracy, 62.5% sensitivity, 75.6% AUC, and 72.5% specificity. For PD dementia prediction, the combined model (clinical and biological data) using SVM (10 variables) achieved 81.9% accuracy, 47.1% sensitivity, 86.2% AUC, and 90.9% specificity. The clinical features model using RF (eight variables) achieved 80.7% accuracy, 47.1% sensitivity, 82.8% AUC, and 89.4% specificity. The biofluid model using ElasticNet (five variables) yielded 86.7% accuracy, 47.1% sensitivity, 83.5% AUC, and 97.0% specificity.

Pahuja and Prasad[83] applied DL architectures to detect PD by integrating biological, MRI, and SPECT features from 132 subjects (73 PD and 59 healthy) obtained from the PPMI. Using a CNN model, the highest accuracy achieved was 92.38% and 93.33% in the model-level and feature-level frameworks, respectively.

Yang *et al.*[84] developed an AI model for PD detection and monitoring its progression using nocturnal breathing signals. Evaluation of the model was performed on a dataset of 7671 subjects (757 PD and 6914 control) obtained from public cohorts as well as hospitals in the USA. The AI model achieved an AUC of 90% on held-out test sets and 85% on external test sets. It also demonstrated the ability to predict PD progression and severity, showing a strong correlation with the Movement Disorder Society (MDS)-UPDRS scores.

Allwright *et al.*[85] applied an integrated ML algorithm to the United Kingdom (UK) Biobank dataset and found that neutrophil-to-lymphocyte ratio and elevated serum insulin-like growth factor 1 levels may help predict PD risk. Their analysis of 1753 measured non-genetic variables included 334,062 eligible participants, among whom 2719 developed PD since enrollment. The findings support improved early PD diagnosis and potential therapeutic strategies.

Almgren *et al.*[86] developed and evaluated a multimodal ML model to predict cognitive decline in 213 PD patients from the PPMI. The model incorporated CSF, clinical test scores, brain volumes, and genetic variants. An iterative scheme combining the RReliefF-based feature ranking and support vector regression with 10-fold cross-validation was used to identify optimal predictive features and evaluate the performance of that model. A correlation of 0.44 was observed between actual and predicted Montreal Cognitive Assessment scores. The study also revealed that several predictive features of cognitive impairment in PD, such as tau pathology and CSF Aβ, are commonly associated with AD, suggesting an overlap in cognitive decline mechanisms between PD and AD.

Kelly *et al.*[87] applied five ML approaches—LR, RF, SVM, XGB, and MLP—to identify blood-based biomarkers for PD and AD, utilizing various feature selection methods. After pre-processing, the GSE99039 PD cohort was randomly divided into a training cohort of 303 subjects (162 controls and 141 PD) and a test cohort of 131 subjects (68 PD and 63 controls), initially including 20,183 features. For PD, the RF model achieved an ROC-AUC of 74.3%

while the CNN model achieved 71.5%.

McFall *et al.*[88] identified multi-modal predictors in PD using RF classifier combined with an explainable AI (XAI) method (Tree SHapley Additive exPlanation [Tree SHAP]) and biomarkers from MRI, clinical, etc. They tested 38 predictors across 10 domains to differentiate PD without dementia (PDND) from incipient dementia (PDID). The RF model classified PDID from PDND with an AUC of 84% and a normalized Matthew's correlation coefficient (MCC) of 0.76. Tree SHAP revealed that 10 key features accounted for 62.5% of the model's performance, indicating that dementia risk arises from multiple domains.

Tsukita *et al.*[89] integrated high-throughput CSF proteomics and ML to identify CSF signatures associated with PD using data from 279 non-genetic PD subjects and 141 healthy controls from the PPMI. The Least Absolute Shrinkage and Selection Operator (LASSO) method selected 14 differentially expressed proteins from 23 candidates to construct the PD proteomic score (PD-ProS), which showed strong performance with an AUC of 83%. This was validated in an independent internal validation dataset of 71 non-genetic PD subjects and 35 healthy subjects, achieving an AUC of 81%. In addition, PD-ProS distinguished 258 genetic PD subjects from 365 genetic prodromal subjects and predicted cognitive and motor decline, regardless of genetic status, with significant associations with dementia and H&Y stage IV.

Chen *et al.*[90] noted that CSF biomarkers are more sensitive for identifying prodromal PD than MDS measures. ML algorithms were applied to analyze fingerprint response patterns, enabling both qualitative and quantitative estimation of proteins. The KNN regression algorithm was used to evaluate MDS scores, achieving a mean square error of 38.88.

Dennis and Strafella[91] found that integrating various biomarkers, including neuroimaging and biofluids, can enhance diagnostic accuracy and predict cognitive deterioration in PD, based on a review of 21 studies. Their analysis revealed that MRI and functional MRI achieved accuracy and AUC scores above 80%. In addition, tau and Aβ42 were found effective in identifying PD subjects, with AUC scores and accuracy exceeding 90%.

Hällqvist *et al.*[92] used mass spectrometry-based proteomic phenotyping to find out blood biomarkers that could help detect at-risk individuals to slow PD symptom progression. Blood samples were analyzed from 99 recently diagnosed motor PD subjects, premotor were analyzed with iRBD from two datasets (18 and 54 subjects, longitudinally), and 36 healthy controls. The developed ML model, by analyzing the expression levels of eight proteins, correctly identified all individuals with PD and classified 79% of premotor individuals up to seven years before motor symptoms appeared.

Some studies[15-17] did not employ ML techniques but emphasized the significance of biological biomarkers in investigating PD progression, monitoring, and diagnosis. These studies highlighted the potential of various biological biomarkers to deepen our understanding of PD and improve clinical judgment in managing the disease.

In recent years, research on proteomic and biological biomarkers in PD has expanded significantly, offering new tools for diagnosis, monitoring, and insight into the disease's underlying mechanisms. Moreover, AI methods continue to advance the field by providing powerful means to analyze large-scale datasets and discover novel biomarkers. The combination of these AI technologies have potential for performing early PD detection with higher accuracy. This could enable clinicians to develop customized treatment strategies aimed at slowing disease progression and improving patient outcomes. Future PD research will likely focus on combining biomarkers with ML to broaden our understanding and enhance disease management.

#### *3.2.4. Genetic biomarkers and AI methods*

Genetic biomarkers refer to specific genetic variations that can be associated with the presence, progression, or susceptibility to a disease. In PD, several genetic mutations have been identified that are linked to hereditary forms of the disease, as well as to some sporadic cases. Key genes associated with PD include *SNCA*, *LRRK2*, *PRKN*, *GBA*, *VPS35*,[20] and *PINK1*.[20,93]

Falchetti *et al.*[94] conducted a gene expression meta-analysis of blood transcriptomes from PD and healthy subjects to identify gene signature of PD. Microarray data from four independent cohorts, totaling 711 instances (323 healthy and 388 iPD), were used for analysis. Collinearity recognition algorithms and RFE were employed to derive a 59-gene signature of iPD from the top 100 genes having the highest negative and positive effect sizes. Four sample size-adjusted training sets and nine classification algorithms are used to evaluate this gene signature. Of the 36 models created, 33 demonstrated accuracy greater than the non-information rate. Two models, based on SVM regression, exhibited the highest accuracy in predicting PD and healthy control samples.

Su *et al.*[95] provided insights into the application of ML models by analyzing PD genetic and transcriptomic data. They reviewed studies and emphasized the significant potential of ML in revealing hidden patterns

in PD transcriptomic and genetic data. Their review emphasized that the examined studies have successfully uncovered important knowledge about the pathology and pathogenesis of PD, demonstrating the power of ML in advancing the understanding of the disease.

García-Fonseca *et al.*[96] suggested that ML is an important tool for classifying expression profiles of non-coding RNAs between healthy and PD subjects. Furthermore, these authors explained the importance of ML models in diagnosing NDs by summarizing results from various studies, which demonstrated accuracies ranging from 85% to 95% in ND detection using ML.

Hu *et al.*[97] conducted differential expression analysis (DEA) to identify differentially expressed genes (DEGs) deregulated in both PD and periodontitis using Gene Expression Omnibus (GEO) datasets. Genes associated with inflammatory response were retrieved from the Molecular Signatures Database. K-means clustering was employed for sample clustering, and LASSO model was used to perform FSS. Five genes—*PLAUR, TCIRG1, MANSC1, FMNL1*, and *RNASE6* —were determined as crosstalk biomarkers connecting periodontitis with PD.

Lam *et al.*[98] analyzed data from 1223 UK Biobank subjects to identify clinical and genetic biomarkers associated with NDs, namely, PD, AD, myasthenia gravis, and motor neuron disease. By employing an ML approach with Monte Carlo randomization, they identified biomarkers for predicting these NDs. The study demonstrated that, by training on available clinical markers, the multinomial model predicts NDs with an accuracy of 88.3%.

Makarious *et al.*[99] developed a model using GenoML on multimodal data from the PPMI to predict PD risk. These authors observed that when their final multimodal model was tested on males (65.57% PD and 63.74% control), it achieved 85.56% accuracy and 82.41% balanced accuracy, with 89.31% sensitivity and 75.51% specificity, outperforming the single modality data model. When validated on the PDBP dataset (males: 64.18% PD and 45.25% control), the tuned multimodal model achieved an AUC of 85.03%, with 43.07% specificity and 93.12% sensitivity.

Pantaleo *et al.*[100] used a robust ML approach to classify PD from healthy subjects in 579 samples collected from 390 individuals in the early PD group and 189 AMC individuals in the healthy group, using whole-blood transcriptomics data from the PPMI. Using a nested FSS method based on RF and XGB, they achieved an AUC of 72%. They also discussed the significance of the 493 candidate genes by using functional analysis based on Kyoto Encyclopedia of Genes and Genomes pathways and Gene Ontologies.

Vuidel *et al.*[101] differentiated the following into mDANs: induced pluripotent stem cells (iPSCs) derived from patients with the *LRRK2* G2019S mutation, an isogenic control, and iPSCs that are genetically not related. The authors identified increased levels of serine 129 phosphorylation and α-syn, decreased dendritic complexity, and mitochondrial dysfunction using automated fluorescence microscopy in a 384-well-plate format. ML methods were utilized to classify mDANs based on genotype and to identify drug-treated neurons using image-extracted features. The Z-factor (0.43) of SVM outperformed the Z-factor (0.12) of LDA. Their approach enhanced the applicability of mDANs in PD modeling and in identifying new LRRK2-linked drug targets.

Cai *et al.*[102] utilized SVM and weighted gene co-expression network analysis (WGCNA) for the identification of gene modules and the development of a PD diagnostic model using three GEO datasets. Sixty percent of the combined dataset (38 PD and 29 controls) was used for training, and 40% for testing, along with an external validation dataset (16 PD and nine controls). The developed model showed an AUC above 80% across the training, test, and validation sets, with performance confirmed through Synthetic Minority Over-Sampling Technique analysis. An AUC score of 74% for age features further validated the SVM model's reliability. These results suggest that combining WGCNA with SVM holds promise for biomarker screening and diagnostic model development for PD.

Hajianfar *et al.*[103] aimed to identify two gene mutations in PD by utilizing hybrid ML systems (HMLSs) based on non-imaging and imaging data. From the PPMI, 264 and 129 subjects with identified *LRRK2* and *GBA* mutation status were considered. Each dataset contained 513 features. Multiple HMLSs, consisting of 11 feature extraction or 10 FSS algorithms combined with 21 classifiers, were applied. In addition, Ensemble Voting was used for gene classification. For *LRRK2* and *GBA* mutation status prediction, several HMLSs achieved 98% ± 2% accuracy and 90% ± 8% accuracy, respectively, in five-fold cross-validation data. Additionally, 100% accuracy and 96% accuracy, respectively, were observed in external test data.

Wang *et al.*[104] employed ML and bioinformatics techniques to identify genes related to ferroptosis in PD by analyzing DEGs. A total of 109 PD-related ferroptosis DEGs were identified after combining three cohorts (GSE7621, GSE202665, GSE20146) from the National Center for Biotechnology Information (NCBI) GEO and FerrDb V2 databases. The researchers also identified natural products with anti-PD effects that could be used

for treatment. ML algorithms revealed six hub genes (*IL6, ATG7, TLR4, ADIPOQ, FADS2,* and *PTGS2*) and 29 overlapping genes. In addition, the study screened 263 natural product components and constructed an "Overlapping Genes-Ingredients" network.

Xin *et al.*[105] identified important immune-related hub genes in PD using ML and developed a diagnostic model based on the GEO (GSE8397) database, which includes gene expression data from 15 healthy and 24 PD subjects' substantia nigra (SN) samples. DEGs related to PD were identified using WGCNA and DEA. LASSO and multiple SVM-RFE ML algorithms were used to identify hub genes (*TTD19, DLD, DLK1*, and *IARS*). LR was then employed to develop a PD classification model, and its accuracy was tested in three unrelated cohorts: GSE20292 (18 healthy SN samples and 11 PD SN samples), GSE7621 (nine healthy SN samples and 16 PD SN samples), and GES49036 (eight healthy SN samples and 15 PD SN samples). The AUC scores for *DLK1*, *DLD*, and *TTC19* exceeded 70% in GSE8397 and all three external validation datasets, indicating strong accuracy. In GSE8397 and one external validation cohort, *IARS* showed an AUC greater than 70%, with values ranging from 50% to 70% in the other two datasets, suggesting its research value. The joint diagnostic model, developed with the four immune-related PD hub genes, demonstrated an AUC greater than 90% in GSE8397 and all three external validation datasets.

Zhang *et al.*[106] employed interpretable DL approaches to identify important genes and biomarkers related to PD using gene expression data from a GEO dataset. Their approach yielded promising results, achieving an AUC of 73% and an F1-score of 71%, effectively distinguishing PD subjects and providing valuable insights into relevant biological pathways. Using interpretable DL models, the authors identified important biomarkers (*XK, TUBA4B, TP53,* and *PDK1*) and their associated biological pathways linked to PD. Notably, the *XK* gene showed a strong correlation with PD.

Ameli *et al.*[107] proposed that the integration of Singular Vector Feature Selection and the RF algorithm can be effectively utilized to analyze single-nucleotide polymorphism (SNP) data and identify PD biomarkers. To assess the reproducibility of these biomarkers, they gathered five SNP datasets from the Database of Genotypes and Phenotypes, including dataset IDs phs000394, phs000126, phs000089, phs000089, and phs000048, with sample sizes of 1001, 2082, 1741, 526, and 886, respectively. Their analysis revealed that, on average, 93% of the SNPs identified in one dataset were not repeated in the others. However, when multiple datasets were integrated, the replication gap dropped to 62%. Furthermore, these researchers identified four SNPs directly linked to PD and 50 SNPs indirectly linked to PD in the literature.

Banou *et al.*[108] applied ML algorithms to analyze single-cell RNA-sequencing data related to PD and to explore their association with hyperbaric oxygen therapy (HBOT). The dataset included 4495 cells (2518 from control and 1977 from PD groups), with expression profiles across 18,098 genes. FSS was performed using the XGB. The authors employed 15 ML algorithms, including LR, KNN, NB, DT, RF, gradient boosting machines, SVM, quadratic discriminant analysis, ridge classifier, LDA, extreme gradient boosting machine (LightGBM), CatBoost, AdaBoost, ETs, stochastic gradient descent, and a dummy classifier, to classify cells from PD-affected subjects versus healthy subjects. Using the top 100 genes, LR outperformed the other ML algorithms in terms of accuracy, precision, F1-score, MCC, and Kappa, achieving 99.59%, 99.43%, 99.53%, 99.17%, and 99.16%, respectively. The highest AUC (100%) and recall (100%) were achieved by CatBoost and NB classifiers, respectively. Genes such as *MAP2, WSB1*, and *CAP2*, among others, were found to be highly related to PD and demonstrated notable correlation with HBOT.

Kumar *et al.*[109] employed data-mining techniques to identify novel microRNA (miRNA) biomarkers and subsequently developed an ML model for PD diagnosis based on the identified biomarkers. The training dataset comprised 112 miRNAs (56 PD and 56 non-PD). After filtering, the number of features was reduced from 16,299 to 61. Ten-fold cross-validation tests yielded the following accuracies: 87.50% for RF, 91.07% for the Hoeffding Tree, 91.96% for NB, 90.18% for MLP, and 95.65% for the Sequential Model. As the Sequential Model outperformed the others, its performance was validated using an independent dataset, achieving 93.3% accuracy.

By analyzing transcriptome data of 117 subjects (56 PD and 61 healthy) from the GEO database and validating the findings through reverse transcription-quantitative polymerase chain reaction (RT-qPCR), Peng *et al.*[110] discerned *EAF2* as a significant gene in PD, consistently showing downregulation in PD subjects compared to healthy subjects. DEA, WGCNA, and three ML algorithms (RF, LASSO, and SVM-RFE) were applied to identify critical genes related to PD. The diagnostic performance of *EAF2* showed an AUC of 74.5% in the training dataset, 75.2% in the validation dataset, and 84.2% in blood samples, indicating its association with PD pathology.

Teng *et al.*[111] explored and evaluated critical genetic biomarkers for PD diagnosis. DEA was conducted on the PD datasets obtained from GEO database (GSE20141 – 18 tissue samples, GSE18838 – 28 blood samples, GSE20295

– 93 tissue samples, and GSE6613 – 102 blood samples) consists of both PD and control tissue samples. Using two ML methods, LASSO and SVM, the study identified *GPX2*, *ZNF556*, and *CR1* as genes crucial to PD pathogenesis, suggesting these may serve as potential diagnostic biomarkers. In the validated blood sample dataset, the combined assessment of these three genes outperformed individual gene assessments, achieving an AUC of 70.1%. Samples from peripheral blood mononuclear cells exhibited consistent diagnostic value for each gene, with the combination yielding improved performance with an AUC of 80.1%.

Yan *et al.*[112] performed DEA of the GSE8397 dataset (18 control and 29 PD) from the GEO database and selected 11 key N6-methyladenosine (m6A)-related genes to develop two ML models using SVM and RF algorithms. The RF model achieved an AUC value of 100%, outperforming the SVM model, which achieved an AUC value of 98.3%. In the final stage, the RF model was visualized, and four m6A-related genes (*YTHDC2*, *LRPPRC*, *HNRNPC*, and *IGFBP3*) were identified as major candidates for the "nomogram model," leading to accurate recognition of PD. They also identified two distinct m6A clusters in PD, each characterized by contrasting immune features, by analyzing the information from the 11 m6A-related genes.

Yang *et al.*[113] identified four genes related to aging by training an ML model using whole-blood RNA-sequencing data from 24 subjects (13 healthy and 11 PD). By employing ML algorithms, such as LASSO, SVM, RF, and Ridge regression, along with LASSO regression and Venn diagrams, they identified four genes as significant PD biomarkers. These genes were further assessed using three additional datasets from GEO and RT-qPCR in peripheral blood mononuclear cells from 10 PD and 10 healthy subjects. ROC curve analysis demonstrated that aging-related DEGs could effectively distinguish PD from healthy subjects, with an AUC score exceeding 70%, indicating their potential as PD diagnostic biomarkers.

Yu *et al.*[114] aimed to identify the most relevant gene in each PD locus and uncover novel mechanisms implicated in PD. An XGB ML model was trained using 212 genes (seven well-known genes labeled as positive and 205 genes not associated with PD labeled as negative) from Genome-Wide Association Study loci, utilizing transcriptomic, genomic, and epigenomic data from mDANs and brain tissues. Sixty-three percent of genes were assigned a probability score greater than 75% and were therefore considered related to PD.

Genetic research has revolutionized the understanding of PD, shifting the focus from clinical symptoms to molecular mechanisms. The identification and application of genetic biomarkers hold great promise for early diagnosis, risk prediction, and targeted treatment. Future advancements in genetic screening, personalized medicine, and gene therapy could significantly improve outcomes for individuals living with PD. However, much work remains in translating genetic findings into tangible clinical applications.

#### *3.2.5. Gut biomarkers and AI methods*

Gut biomarkers for PD are mainly classified under both biochemical and markers as well as areas such as microbiome and metabolomic profiling. While not a completely separate category, gut-related biomarkers can also be studied through neuroimaging and clinical observations. Here, we explore research conducted so far that utilizes gut biomarkers in conjunction with AI methods for PD detection, monitoring, and classification of PD subjects from healthy subjects.

Pietrucci *et al.*[115] investigated the role of GM in PD and identified common microbial alterations that could potentially predict PD. The authors applied three ML algorithms—RF, neural network, and SVM—to analyze 846 metagenomic samples (472 PD and 374 healthy). The RF algorithm outperformed the others, achieving an AUC score of 80% ± 1% and 71% accuracy, as compared to AUC score 67% ± 3% for the neural network and 54% ± 8% for the SVM. RF also identified a subset of 22 microbial families capable of distinguishing PD and healthy subjects.

Qian *et al.*[116] created the first GM gene catalog related to PD based on metagenomic sequencing. They collected GM genes from the feces of 40 Chinese PD subjects and their healthy counterparts using shotgun metagenomic sequencing (SMG). By applying the mRMR technique, 25 gene markers were selected from 51,816 genes as potential PD biomarkers. When these 25 biomarkers were used in an SVM classifier, the model achieved: 89.6% AUC, 90% sensitivity, and 75% specificity. The identified genes were further validated using real-time PCR in a separate dataset of 78 PD and 75 healthy subjects, achieving an AUC of 90.5%, 86% sensitivity, and 77% specificity. An AUC of 83.1%, sensitivity of 85%, and specificity of 78% were observed when differentiating 78 PD and 40 multiple system atrophy subjects. Furthermore, 90.1% AUC, 90% sensitivity, and 88% specificity were achieved in differentiating 78 PD from 25 AD subjects.

Lubomski *et al.*[117] developed a PD prediction model to assess GM compositional changes in combination with macronutrient consumption. They conducted a cross-sectional evaluation involving 184 subjects (103 PD and 81 household controls). RF- and SVM-based models were developed to aid in identifying PD. The RF model,

which incorporated taxonomic data at the genus level and the contribution of carbohydrates to total energy intake, exhibited the highest predictive performance, with an AUC score of 74%.

Nie *et al.*[118] investigated the relationship between PD and GM, and developed a PD predictive model by analyzing 2269 16S ribosomal RNA (16S rRNA) specimens (896 healthy and 1373 PD) and 236 SMG specimens (114 healthy and 122 PD). Both 16S rRNA and SMG analyses identified five genera (*Bifidobacterium, Akkermansia, Streptococcus, Desulfovibrio,* and *Lactobacillus*) with increased abundance, and five genera (*Lachnospira, Faecalibacterium, Roseburia, Blautia,* and *Prevotella*) with decreased abundance in PD patients. Moreover, RF models based on 11 genera achieved classification accuracy exceeding 80% in distinguishing PD from healthy subjects. A separate RF model based on six inflammation-related genes outperformed the former, with accuracy >90%. These outcomes highlight the role of inflammation in PD prediction and treatment.

To examine the role of gut dysbiosis in PD progression, Nishiwaki *et al.*[119] developed RF models to predict 2-year PD progression based on GM from 165 PD subjects. The AUC-ROC scores of the GM-based models for H&Y stages 1 and 2 were 79.9% and 70.5%, respectively. In addition, the GM profile predicted MDS-UPDRS III scores' progression in early-stage PD with an AUC-ROC of 72.8%. An increase in mucin-degrading genus Akkermansia and a decrease in short-chain fatty acid-producing genera, *Blautia, Faecalibacterium,* and *Fusicatenibacter,* were associated with faster PD progression.

The objective of the study conducted by Sánchez[120] was to discover potential biomarkers by comparing the GMs of 20 control and 20 PD subjects, identifying candidate taxa, gene families, and pathways that could offer insights into variables important for early PD detection. This was achieved using various metagenomics programs alongside five ML algorithms (DT, RF, NB, SVM, and KNN). The study identified key features, including an overexpression of Myo-chiro and scyllo-inositol degradation pathways and a higher abundance of Lactococcus phage in PD patients.

Boodaghidizaji *et al.*[121] used ML algorithms to analyze the patterns of stool microbiota as well as their response to fiber as a diagnostic tool for lifelong inflammatory diseases. They applied ML algorithms to differentiate between PD, ulcerative colitis, Crohn's disease, HIV, and healthy subjects, with and without fiber treatment, achieving classification accuracy of up to 95%. In addition, ML algorithms achieved accuracy up to 90% when microbiome data were used to predict ulcerative colitis and Crohn's disease.

Dhatrak[122] explored the application of ML predictive technologies to analyze the compositions of GMs and their alterations in PD subjects. The study utilized a dataset of 17 randomly selected fecal samples (nine PD and eight healthy) obtained from the European Nucleotide Archive database under the project PRJEB27564. The performance of two ML algorithms, RF and linear support vector classifier (LSVC), was evaluated using the QIIME 2 classifier and various performance metrics. With RF, the following results were achieved: accuracy = 66%, recall = 66%, precision = 66%, F1-score = 88%, and specificity = 66%. For LSVC, the results were: accuracy = 66%, recall = 100%, precision = 66%, F1-score = 82%, and specificity = 33%.

Li *et al.*[123] thoroughly assessed the performance of GM-based ML classification algorithms across 20 diseases, using 83 case-control cohorts (9708 samples in total) across five main disease groups. Each disease was represented by at least two cohorts. In single-cohort classifiers, high predictive accuracies (~77% AUC) were achieved in within-cohort validation, but lower accuracies were observed in cross-cohort validation, excluding intestinal diseases, where AUC was around 73%. To improve validation scores for non-intestinal diseases, samples from multiple cohorts were used to train combined-cohort classifiers. The study also predicted the sample size needed to reach more than 70% validation accuracies. Additionally, for intestinal diseases, higher validation performance was achieved by classifiers using metagenomic data than 16S amplicon data.

Romano *et al.*[124] conducted a meta-analysis of PD GM studies with 4489 samples from 11 countries across four continents, reporting on the fecal microbiomes of PD subjects and controls using both 16S amplicon sequencing (3165 samples) and SMG (1324 samples). They trained the ML models on various datasets and concluded that GM is associated with both PD diagnosis and its treatment.

Zhang *et al.*[125] proposed an interpretable and accurate neural network approach for PD prediction and biomarker discovery using whole metabolomics datasets without initial FSS. Samples were collected from two cross-sectional studies: the EPIC study (GC-MS: 36 PD and 39 control; capillary electrophoresis-mass spectrometry: 39 PD and 39 control; LC-MS[+]: 39 PD and 39 control; LC-MS[−]: 36 PD and 37 control; composite: 35 PD and 37 control) and the National Health Service study (LC-MS[+]: 80 PD and 56 control; LC-MS[+]: 138 PD and 56 control). The neural network approach demonstrated significantly superior performance in predicting PD from blood plasma metabolomics data, achieving a mean AUC exceeding 99.5%, outperforming five other ML methods. XGB and LR showed similar performance with AUC-ROC scores of 97.0% ± 2.8% and 96.8% ± 3.7%, and AUC-precision-recall

scores of 96.8% ± 3.1% and 96.9% ± 3.7%, respectively. In contrast, RF, LDA, and SVM classifiers showed relatively lower performance, with AUC-ROC and AUC-precision-recall values of 82.9% ± 9.9% and 83.6% ± 9.9% for RF, 64.7% ± 9.3% and 66.1% ± 11.1% for SVM, and 68.1% ± 9.1% and 63.4% ± 11.9% for LDA. Based on the MCC score, the neural network approach outperformed the other classifiers, with a score of 91.8% ± 8.6%, compared to 81.5% ± 13.2% for LR, 78.7% ± 11.9% for XGB, 43.3% ± 19.2% for RF, 27.2% ± 15.2% for LDA, and 21.3% ± 15.5% for SVM.

Li *et al.*[126] developed an AI-guided, gut micro-environment-triggered imaging sensor to accurately and non-invasively identify the PD stages by monitoring α-syn using a DL algorithm. In mouse experiments, PD stages were classified as 0 (early), 1 (middle), and 2 (advanced). Initially, the dataset included samples from 10 normal, nine midterm PD, and 16 advanced PD mice; however, after data augmentation, the dataset expanded to 40 normal, 40 midterm PD, and 60 advanced PD mouse samples. The proposed CNN model, based on AlexNet, outperformed five benchmark ML algorithms (DT, SVM, KNN, LDA, and NB), achieving over 98% testing accuracy.

Zhao *et al.*[127] performed a meta-analysis to examine the role of GM in PD and its diagnostic potential. They integrated six *16S rRNA* gene datasets from five different studies, comprising 456 healthy and 550 PD samples. The analysis identified reduced levels of butyrate-producing taxa (*Faecalibacterium*, *Roseburia*, *Coprococcus_2*) and increased levels of *Akkermansia* and *Bilophila* in PD. Using a network-based approach, the study identified microbial biomarkers for PD and developed a classification model based on RF using 11 key genera, demonstrating strong diagnostic potential. The optimized PD classification model achieved 100% accuracy and 100% AUC on the training dataset, and 80.2% accuracy and 86.4% AUC on the test dataset.

Rojas-Velazquez *et al.*[128] utilized four PD-related datasets from the NCBI, focusing on stool samples, and identified a microbiome signature for diagnosing PD using ML. By employing the Recursive Ensemble Feature Selection algorithm, 84 features were identified from the discovery dataset (PRJEB14674–345 samples: 134 healthy and 211 PD), achieving an accuracy exceeding 80%. The ET classifier demonstrated a diagnostic accuracy with an AUC-ROC of 74% in validating the discovery dataset. During testing, AUC-ROC scores of 64% for PRJEB14674 (345 samples: 134 healthy and 211 PD), 71% for PRJEB27564 (266 samples: 130 healthy and 136 PD), and 62% for PRJNA594156 (300 samples: 103 healthy and 197 PD) were achieved.

Yu *et al.*[129] developed an efficient DL-based prediction method for accurately diagnosing PD by analyzing GM data. The study utilized data from 39 PD subjects and their respective 39 healthy spouses. For FSS, a pre-processing technique called combined ranking using RF scores and PCA contributions was applied, followed by the LSIM (LSTM-penultimate to SVM Input Method) for classifying PD subjects. A soft voting mechanism was then used for final PD prediction. The Parkinson Gut Prediction method demonstrated an AUC of 92%, a mean accuracy of 85%, and an ROC of 92%.

These studies indicate that by leveraging gut biomarkers and AI methods, there is significant potential for improving the detection, monitoring, and management of PD. As research progresses, AI-powered approaches are set to become key tools in clinical practice, thereby enhancing patient care and outcomes in PD.

## 4. Discussion

This section discusses some important parameters used to evaluate the studies related to PD. The aim of this analysis is to explore how these parameters impact various aspects of PD diagnosis and treatment. The parameters under consideration include research objectives, the strengths and weaknesses of AI methods, existing innovations, and their practical applications in the medical field.

### 4.1. Challenges in selecting suitable biomarkers

While assessing different types of biomarkers and AI methods for PD detection, we observed several challenges due to the complex and multifactorial nature of PD. Both biomarkers and AI methods offer distinctive advantages but also have limitations. Some key challenges associated with using different types of biomarkers and AI methods in PD detection include:

(i) Variability and heterogeneity of PD: PD manifests differently across individuals, with varying symptoms, disease progression rates, and responses to treatment. Therefore, no single biomarker can comprehensively represent PD or its progression. As a result, it is challenging to identify reliable and consistent biomarkers that perform well for all affected PD subjects.

(ii) Lack of early detection biomarkers: many existing biomarkers tend to detect PD at later stages of the disease, after significant neuronal damage has already occurred. Early-stage biomarkers are still under investigation but have not yet been clinically established. Biochemical biomarkers, such as CSF, hold potential for early detection but require further validation to ensure reliability and specificity for PD. In addition, biomarkers often suffer from poor

specificity, meaning they may not distinguish PD from other NDs like AD. Similarly, some biomarkers may lack the necessary sensitivity to detect PD in its early stages or in individuals with atypical presentations. Certain fluid-based biomarkers, such as those found in CSF, require invasive procedures such as lumbar punctures, which can be uncomfortable for patients and limit their practical use in routine screening for PD. On the other hand, neuroimaging biomarkers are non-invasive but can be costly, time-consuming, and require specialized equipment and expertise.

(iii) Standardization of PD biomarkers: standardization of many PD biomarkers is still an issue. The features extracted from imaging can vary depending on the type of machine, the scanning method, and how the results are interpreted. Moreover, many biomarkers still lack robust clinical validation in large and diverse cohorts, which is necessary to establish their clinical utility and reliability for PD diagnosis and prognosis. Standardization is important to ensure that the findings from these techniques can be widely applied and trusted across various research and clinical environments.

### 4.2. Challenges in applying AI methodologies

The performance of AI algorithms for PD detection is influenced by several factors, including the quality of data, the choice of model, hyperparameter settings, and data pre-processing. Below are the key factors that significantly affect how well AI models perform in PD detection:

(i) Data quality: accurate and comprehensive labels are critical in supervised learning for training and evaluating AI models. Inaccurate or missing labels can lead to poor model performance. Medical data, particularly from clinical assessments or imaging, may contain noise or outliers that also affect AI algorithm performance and, consequently, the model's ability to generalize. Imbalanced datasets, often suffering from class imbalances, can bias predictions toward the majority class, causing poor detection of PD cases. The presence of missing values in the dataset can also reduce model performance.

(ii) Feature selection: selecting relevant features is important. Irrelevant or redundant features can reduce the model's ability to generalize and increase computational costs. However, high-dimensional datasets can also lead to overfitting.

(iii) Model selection: different AI algorithms are suitable for different data types. For example, SVM performs well with high-dimensional datasets, such as genetic data or neuroimaging features. RF is robust to noise and effective with diverse data types, such as clinical, biomarker, or demographic data. CNNs are well-suited for imaging data due to their ability to learn spatial hierarchies of features. RNNs or LSTM networks are better for time-series data,[130] such as motor symptom progression over time. More complex models may achieve better results but are more prone to overfitting when training data is limited. Simpler models, for example, SVM or RF, may generalize better on smaller or less noisy datasets.

(iv) Hyperparameter tuning: most AI algorithms have hyperparameters (e.g., kernel type in SVM, number of trees, rate of learning, maximum depth) that must be optimized for good performance. Improper tuning can lead to poor model performance, overfitting, or underfitting.

(v) Training data size: AI models, especially DL algorithms, require large datasets to achieve high performance. With small datasets, models tend to memorize training data and fail to generalize to external testing data.

(vi) Cross-validation and testing: cross-validation helps ensure the model generalizes well to unseen external testing data and reduces the likelihood of overfitting to a specific training set. It is also important to evaluate AI models on data representative of real-world conditions.

(vii) Temporal changes and disease progression: models incorporating time-series data or temporal information (e.g., gait, motor fluctuations) often require specialized algorithms such as RNNs, LSTMs, or reinforcement learning. PD models using longitudinal data need to account for variability over time.

(viii) Bias and fairness: AI model capabilities are limited by the quality of the training data. If training datasets for PD detection lack diversity (in age, ethnicity, gender, comorbidities), AI systems may perform poorly for underrepresented groups, leading to biased diagnoses and health disparities.[130] Efforts are needed to ensure AI systems are fair and unbiased, treating all patients equitably. This includes developing diverse and representative datasets and actively testing AI models to assess their performance across different demographic groups.

### 4.3. Ethical and clinical implications

Ienca and Ignatiadis[131] emphasized that while AI holds significant promise for advancing brain research by optimizing and developing effective neurotechnology frameworks, it also raises important ethical and clinical concerns. The impact of AI on scientific validity and neuroethics remains uncertain.

In PD detection, AI systems require access to sensitive data such as medical histories, neuroimaging scans, and

genetic information. Ensuring secure data storage and protecting patient privacy are crucial, with AI systems needing to comply with regulations such as the Health Insurance Portability and Accountability Act and the General Data Protection Regulation. Patients must provide informed consent for the use of their data, and transparency about how their data is handled is essential.[130] Given that AI models are trained using patient data, it is vital that patients fully understand how their data will be utilized. Clear and accessible information should be provided, and patients should be encouraged to ask questions before consenting to participate in AI-based studies or diagnostic tools.

### 4.4. Future directions

While AI has made considerable advances in PD research, future efforts are crucial to overcome existing challenges and fully harness AI's potential in patient care. Key research areas include:

(i) Integration of multimodal data: PD is a multifactorial disease, and a single modality may not provide a comprehensive understanding of its progression. Integrating data from different biomarker modalities, such as neuroimaging, clinical, biochemical, and genetic, could facilitate early diagnosis and improve predictions of disease progression.

(ii) Real-time monitoring and predictive analytics: traditional PD monitoring largely depends on clinical visits. AI, in combination with wearable devices (e.g., smartwatches), can track real-time data, such as motor fluctuations and sleep patterns, to predict disease progression and support personalized treatment adjustments.

(iii) XAI for clinical decision support: developing XAI models can help clinicians interpret AI predictions, improve trust, and support more informed decision-making. For instance, algorithms that highlight key features, such as a combination of motor symptoms, voice analysis, or affected brain regions, can assist clinicians in interpreting results and guiding treatment plans.

(iv) AI in personalized treatment planning: PD affects individuals differently, and treatment responses can vary significantly. AI can analyze patient-specific data (e.g., genetics, lifestyle, disease severity) to recommend customized therapies, potentially improving treatment efficacy and minimizing side effects.

(v) AI in early detection through biomarker discovery: detecting PD in its early stages is challenging, as symptoms often appear after substantial neuronal damage. AI can facilitate the discovery of early biomarkers by analyzing large datasets across various modalities to identify molecular markers that precede clinical symptoms.

(vi) Improved AI for non-motor symptom monitoring: both motor and non-motor symptoms are significantly associated with disease progression. AI tools capable of monitoring non-motor symptoms, such as cognitive changes or depression, through smart home devices or wearable sensors could provide deeper insights into patient status.

(vii) Collaboration between AI experts, clinicians, and patients: a gap often exists between AI model development and its practical application in clinical settings. Without collaboration among AI researchers, clinicians, and patients, models may lack clinical relevance or overlook patient-centered concerns. Close interdisciplinary collaboration is essential to ensure AI models are clinically useful, patient-centric, and applicable in real-world settings.[130]

### 4.5. Limitations of the study

Despite providing a detailed synthesis of the most relevant information on AI methods for PD diagnosis, this review has certain limitations. Although we aimed to address most of the research questions outlined in Section 3.1, some information could not be included due to its absence in the existing literature.

(i) In several studies, the number of subjects affected and unaffected by PD was not clearly stated. This missing information creates uncertainty for other researchers regarding the actual number of individuals affected by PD worldwide.

(ii) We found that the sizes of the training and testing datasets were often not reported, making it difficult for other researchers to replicate the developed models for PD progression.

(iii) Details of performance metrics were also missing in several studies, hindering the ability to evaluate which AI algorithms perform best for PD detection. Consequently, despite some studies reporting impressive classification results, they were excluded from this review due to the lack of necessary performance details.

(iv) In a few studies, inconsistencies were observed between the abstract and results sections, with different values reported for performance metrics. This inconsistency adds to the confusion among researchers.

(v) During our review of the literature, we also noted a lack of collaboration between clinicians and AI researchers, which may limit the clinical relevance and practical application of the proposed AI models.

## 5. Conclusion

The intricate nature of PD arises from complex

interactions between environmental and genetic factors, making it difficult to accurately identify its underlying mechanisms.

Recent advancements in AI have positioned these technologies as powerful tools for analyzing and interpreting medical data. By integrating various types of biomarkers with AI techniques, early diagnosis and enhanced treatment strategies for PD can be facilitated. However, the implementation of these methodologies presents several challenges, including issues related to data quality and diversity, model interpretability, and the need for interdisciplinary collaboration among healthcare professionals and researchers.

Findings from this review indicate that AI can achieve diagnostic accuracies exceeding 90% for PD, while also reducing the time required for diagnosis. Moreover, AI approaches have demonstrated superior prognostic capabilities compared to traditional diagnostic methods. Ultimately, this review provides valuable insights for both medical professionals and researchers into the potential and challenges of leveraging AI techniques to enhance the diagnosis and management of PD.

## Acknowledgments

None.

## Funding

None.

## Conflict of interest

The authors declare that they have no competing interests.

## Author contributions

*Conceptualization:* All authors
*Writing – original draft:* All authors
*Writing – review & editing:* All authors

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

Not applicable.

## Further disclosure

The authors declare that no AI and AI-assisted technologies were used in conducting key aspects of the research, such as generating scientific insights, analyzing or interpreting data, or drawing scientific conclusions.

## References

1. *Neurodegenerative Diseases. Cleveland Clinic*; 2023. Available from: https://my.clevelandclinic.org/health/diseases/24976-neurodegenerative-diseases [Last accessed on 2025 Apr 21].
2. *Launch of WHO's Parkinson Disease Technical Brief*. WHO. Geneva: World Health Organization; 2022.
3. *Statistics. Parkinson's Foundation*. Available from: https://www.parkinson.org/understanding-parkinsons/statistics [Last accessed on 2025 Apr 20].
4. Yang W, Hamilton JL, Kopil C, *et al*. Current and projected future economic burden of Parkinson's disease in the US. *NPJ Parkinsons Dis*. 2020;6:15.

   doi: 10.1038/s41531-020-0117-1
5. Zhou ZD, Yi LX, Wang DQ, Lim TM, Tan EK. Role of dopamine in the pathophysiology of Parkinson's disease. *Transl Neurodegener*. 2023;12(1):44.

   doi: 10.1186/s40035-023-00378-6
6. Zhang T, Yang R, Pan J, Huang S. Parkinson's disease related depression and anxiety: A 22-year bibliometric analysis (2000-2022). *Neuropsychiatr Dis Treat*. 2023:1477-1489.

   doi: 10.2147/NDT.S403002
7. Xiao B, Zhou Z, Chao Y, Tan EK. Pathogenesis of Parkinson's disease. *Neurol Clin*. 2025;43(2):185-207.

   doi: 10.1016/j.ncl.2024.12.003
8. Lee TK, Yankee EL. A review on Parkinson's disease treatment. *Neuroimmunol Neuroinflamm*. 2021;8:222-244.
9. Jankovic J, Tan EK. Parkinson's disease: Etiopathogenesis and treatment. *J Neurol Neurosurg Psychiatry*. 2020;91(8):795-808.

   doi: 10.1136/jnnp-2019-322338
10. Bloem BR, Okun MS, Klein C. Parkinson's disease. *Lancet*. 2021;397(10291):2284-2303.

    doi: 10.1016/S0140-6736(21)00218-X
11. Rodriguez-Sanchez F, Rodriguez-Blazquez C, Bielza C, *et al*. Identifying Parkinson's disease subtypes with motor and non-motor symptoms via model-based multi-partition clustering. *Sci Rep*. 2021;11(1):23645.

    doi: 10.1038/s41598-021-03118-w
12. Ramesh S, Arachchige ASPM. Depletion of dopamine in Parkinson's disease and relevant therapeutic options: A review of the literature. *AIMS Neurosci*. 2023;10(3):200-231.

    doi: 10.3934/Neuroscience.2023017
13. Bidesi NS, Vang Andersen I, Windhorst AD, Shalgunov V, Herth MM. The role of neuroimaging in Parkinson's disease. *J Neurochem*. 2021;159(4):660-689.

doi: 10.1111/jnc.15516

14. Öksüz N, Öztürk Ş, Doğu O. Future prospects in Parkinson's disease diagnosis and treatment. *Noro Psikiyatr Ars*. 2022;59:S36-S41.

doi: 10.29399/npa.28169

15. Jurcau A, Andronie-Cioara FL, Nistor-Cseppento DC, *et al*. The involvement of neuroinflammation in the onset and progression of Parkinson's disease. *Int J Mol Sci*. 2023;24(19):14582.

doi: 10.3390/ijms241914582

16. Ma ZL, Wang ZL, Zhang FY, Liu HX, Mao LH, Yuan L. Biomarkers of Parkinson's disease: From basic research to clinical practice. *Aging Dis*. 2024;15(4):1813-1830.

doi: 10.14336/AD.2023.1005

17. Mazzotta GM, Conte C. Alpha synuclein toxicity and non-motor Parkinson's. *Cells*. 2024;13:1265.

doi: 10.3390/cells13151265

18. Koníčková D, Menšíková K, Tučková L, *et al*. Biomarkers of neurodegenerative diseases: Biology, taxonomy, clinical relevance, and current research status. *Biomedicines*. 2022;10(7):1760.

doi: 10.3390/biomedicines10071760

19. Nila IS, Sumsuzzman DM, Khan ZA, *et al*. Identification of exosomal biomarkers and its optimal isolation and detection method for the diagnosis of Parkinson's disease: A systematic review and meta-analysis. *Ageing Res Rev*. 2022;82:101764.

doi: 10.1016/j.arr.2022.101764

20. *Parkinson's Foundation. Genetics behind Parkinson's. PD GENEration*. Parkinson's Foundation website. Available from: https://www.parkinson.org/advancing-research/our-research/pdgeneration/genetics-behind-pd [Last accessed on 2025 Apr 19].

21. *FDA-NIH Biomarker Working Group*. *Glossary*. Food and Drug Administration; 2025. Available from: https://www.ncbi.nlm.nih.gov/books/NBK338448 [Last accessed on 2025 Apr 20].

22. Available from: https://www.physio-pedia.com/index.php?title=biomarkers_of_parkinson's_disease&oldid=346298 [Last accessed on 2025 Apr 21].

23. Surguchov A. Biomarkers in Parkinson's disease. In: *Neurodegenerative Diseases Biomarkers: Towards Translating Research to Clinical Practice*. Berlin: Springer; 2022. p. 155-180.

24. Triadafilopoulos G. Research in gut PD. In: *The Gut in Parkinson's Disease*. Berlin: Springer Nature Switzerland; 2025. p. 165-167.

25. Rahmani AM, Yousefpoor E, Yousefpoor MS, *et al*. Machine learning (ML) in medicine: Review, applications, and challenges. *Mathematics*. 2021;9:2970.

26. Chakraborty C, Bhattacharya M, Pal S, Lee SS. From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. *Curr Res Biotechnol*. 2024;7:100164.

27. Oliveira AM, Coelho L, Carvalho E, Ferreira-Pinto MJ, Vaz R, Aguiar P. Machine learning for adaptive deep brain stimulation in Parkinson's disease: Closing the loop. *J Neurol*. 2023;270(11):5313-5326.

doi: 10.1007/s00415-023-11873-1

28. El Maachi I, Bilodeau GA, Bouachir W. Deep 1D-Convnet for accurate Parkinson disease detection and severity prediction from gait. *Expert Syst Appl*. 2020;143:113075.

29. Goyal P, Rani R. Comparative analysis of machine learning, ensemble learning and deep learning classifiers for Parkinson's disease detection. *SN Comput Sci*. 2023;5(1):66.

30. Senturk ZK. Early diagnosis of Parkinson's disease using machine learning algorithms. *Med Hypotheses*. 2020;138:109603.

31. Ouhmida A, Raihani A, Cherradi B, Terrada O. A novel approach for Parkinson's disease detection based on voice classification and features selection techniques. *Int J Online Biomed Eng*. 2021;17(10):111-130.

32. Chintalapudi N, Battineni G, Hossain MA, Amenta F. Cascaded deep learning frameworks in contribution to the detection of Parkinson's disease. *Bioengineering* (*Basel*). 2022;9(3):116.

doi: 10.3390/bioengineering9030116

33. Ferreira MIASN, Barbieri FA, Moreno VC, Penedo T, Tavares JMR. Machine learning models for Parkinson's disease detection and stage classification based on spatial-temporal gait parameters. *Gait Posture*. 2022;98:49-55.

doi: 10.1016/j.gaitpost.2022.08.014

34. Sigcha L, Domínguez B, Borzì L, *et al*. Bradykinesia detection in Parkinson's disease using smartwatches' inertial sensors and deep learning methods. *Electronics*. 2022;11(23):3879.

35. Thakur K, Kapoor DS, Singh KJ, Sharma A, Malhotra J. Diagnosis of Parkinson's disease using machine learning algorithms. In: *Congress on Intelligent Systems*. Berlin: Springer Nature; 2022. p. 205-217.

36. Trabassi D, Serrao M, Varrecchia T, *et al*. Machine learning approach to support the detection of Parkinson's disease in IMU-based Gait analysis. *Sensors* (*Basel*). 2022;22(10):3700.

doi: 10.3390/s22103700

37. Alalayah KM, Senan EM, Atlam HF, Ahmed IA, Shatnawi HSA. Automatic and early detection of Parkinson's disease by analyzing acoustic signals using classification algorithms based on recursive feature elimination method. *Diagnostics* (*Basel*). 2023;13(11):1924.

doi: 10.3390/diagnostics13111924

38. Govindu A, Palwe S. Early detection of Parkinson's disease using machine learning. *Procedia Comput Sci*. 2023;218:249-261.

39. Martinez-Eguiluz M, Arbelaitz O, Gurrutxaga I, *et al*. Diagnostic classification of Parkinson's disease based on non-motor manifestations and machine learning strategies. *Neural Comput Appl*. 2023;35(8):5603-5617.

40. Yadav S, Singh MK, Pal S. Artificial intelligence model for parkinson disease detection using machine learning algorithms. *Biomed Mater Dev*. 2023;1(2):899-911.

41. Goyal J, Khandnor P, Aseri TC. Objective and automatic assessment of bradykinesia in Parkinson's patients using new repetitive pointing task with machine learning approach. *Multimedia Tools Appl*. 2024;83:81413-81429.

42. Faiem N, Asuroglu T, Acici K, Kallonen A, Van Gils M. Assessment of Parkinson's disease severity using gait data: a deep learning-based multimodal approach. In: *Nordic Conference on Digital Health and Wireless Solutions*. Berlin: Springer; 2024. p. 29-48.

43. Palakayala AR, Kuppusamy P. A qualitative and quantitative approach using machine learning and non-motor symptoms for Parkinson's disease classification. A hierarchical study. *Appl Comput Sci*. 2024;20(3):171-191.

44. Salsone M, Vescio B, Quattrone A, *et al.* Periodic leg movements during sleep associated with REM sleep behavior disorder: A machine learning study. *Diagnostics* (*Basel*). 2024;14(4):363.

    doi: 10.3390/diagnostics14040363

45. Wang Q, Zeng W, Dai X. Gait classification for early detection and severity rating of Parkinson's disease based on hybrid signal processing and machine learning methods. *Cogn Neurodyn*. 2024;18(1):109-132.

    doi: 10.1007/s11571-022-09925-9

46. Byeon H. Development of a depression in Parkinson's disease prediction model using machine learning. *World J Psychiatry*. 2020;10(10):234-244.

    doi: 10.5498/wjp.v10.i10.234

47. Lee KS, Ham BJ. Machine learning on early diagnosis of depression. *Psychiatry Investig*. 2022;19(8):597-605.

    doi: 10.30773/pi.2022.0075

48. Pereira PA, Trivedi DK, Silverman J, *et al*. Multiomics implicate gut microbiota in altered lipid and energy metabolism in Parkinson's disease. *NPJ Parkinsons Dis*. 2022;8(1):39.

    doi: 10.1038/s41531-022-00300-3

49. Li J, Zhao Y, Liu Y, *et al.* AI scheme for high-accuracy and contactless assessment of Parkinson's disease grades. *Biomed Signal Process Control*. 2025;100:107025.

50. Lu H, Qi G, Wu D, *et al.* A novel feature extraction method based on dynamic handwriting for Parkinson's disease detection. *PLoS One*. 2025;20(1):e0318021.

51. Castillo-Barnes D, Martinez-Murcia FJ, Ortiz A, Salas-Gonzalez D, Ramírez J, Górriz JM. Morphological characterization of functional brain imaging by isosurface analysis in Parkinson's disease. *Int J Neural Syst*. 2020;30:2050044.

    doi: 10.1142/S0129065720500446

52. Chakraborty S, Aich S, Kim HC. 3D textural, morphological and statistical analysis of voxel of interests in 3T MRI scans for the detection of Parkinson's disease using artificial neural networks. *Healthcare* (*Basel*). 2020;8(1):34.

    doi: 10.3390/healthcare8010034

53. Chakraborty S, Aich S, Kim HC. Detection of Parkinson's disease from 3T T1 weighted MRI scans using 3D convolutional neural network. *Diagnostics* (*Basel*). 2020;10(6):402.

    doi: 10.3390/diagnostics10060402

54. Huang GH, Lin CH, Cai YR, *et al*. Multiclass machine learning classification of functional brain images for Parkinson's disease stage prediction. Statistical Analysis and Data Mining. *ASA Data Sci J*. 2020;13(5):508-523.

55. Magesh PR, Myloth RD, Tom RJ. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Comput Biol Med*. 2020;126:104041.

    doi: 10.1016/j.compbiomed.2020.104041

56. Solana-Lavalle G, Rosas-Romero R. Classification of PPMI MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease. *Comput Methods Programs Biomed*. 2021;198:105793.

    doi: 10.1016/j.cmpb.2020.105793

57. Shu ZY, Cui SJ, Wu X, *et al*. Predicting the progression of Parkinson's disease using conventional MRI and machine learning: An application of radiomic biomarkers in whole-brain white matter. *Magn Reson Med*. 2021;85:1611-1624.

    doi: 10.1002/mrm.28522

58. Veetil IK, Gopalakrishnan EA, Sowmya V, Soman KP. Parkinson's disease classification from magnetic resonance images (MRI) using deep transfer learned convolutional neural networks. In: *2021 IEEE 18th India Council International Conference (INDICON)*. IEEE; 2021. p. 1-6.

59. Guo X, Tinaz S, Dvornek NC. Characterization of early stage Parkinson's disease from resting-state fMRI data using a long short-term memory network. *Front Neuroimaging*. 2022;1:952084.

    doi: 10.3389/fnimg.2022.952084

60. Tomer S, Khanna K, Gambhir S, Gambhir M. Comparison analysis of GLCM and PCA on Parkinson's disease using

structural MRI. *Int J Inform Retrieval Res*. 2022;12(1):1-15.

61. Vyas T, Yadav R, Solanki C, Darji R, Desai S, Tanwar S. Deep learning-based scheme to diagnose Parkinson's disease. *Expert Syst*. 2022;39(3):e12739.

62. Camacho M, Wilms M, Mouches P, *et al.* Explainable classification of Parkinson's disease using deep learning trained on a large multi-center database of T1-weighted MRI datasets. *Neuroimage Clin*. 2023;38:103405.

doi: 10.1016/j.nicl.2023.103405

63. Erdaş ÇB, Sümer E. A fully automated approach involving neuroimaging and deep learning for Parkinson's disease detection and severity prediction. *PeerJ Comput Sci*. 2023;9:e1485.

doi: 10.7717/peerj-cs.1485

64. Khachnaoui H, Chikhaoui B, Khlifa N, Mabrouk R. Enhanced Parkinson's disease diagnosis through convolutional neural network models applied to spect datscan images. *IEEE Access*. 2023;11:91157-91172.

65. Wang Y, He N, Zhang C, *et al.* An automatic interpretable deep learning pipeline for accurate Parkinson's disease diagnosis using quantitative susceptibility mapping and T1-weighted images. *Hum Brain Mapp.* 2023;44:4426-4438.

doi: 10.1002/hbm.26399

66. Praneeth P, Sathvika M, Kommareddy V, *et al*. Classification of Parkinson's disease in brain MRI images using deep residual convolutional neural network (DRCNN). *Int J Comput Inform Syst Ind Manag Appl*. 2023;15:13.

67. Ahalya RK, Nkondo GF, Snekhalatha U. Automated detection of Parkinson's disease based on hybrid CNN and quantum machine learning techniques in MRI images. *Biomed Eng Appl Basis Commun*. 2024;36(2):2450005.

68. Islam N, Turza MSA, Fahim SI, Rahman RM. Advanced Parkinson's disease detection: A comprehensive artificial intelligence approach utilizing clinical assessment and neuroimaging samples. *Int J Cogn Comput Eng*. 2024;5:199-220.

69. Patil P, Ford WR. Parkinson's disease recognition using decorrelated convolutional neural networks: Addressing imbalance and scanner bias in rs-fMRI Data. *Biosensors* (*Basel*). 2024;14(5):259.

doi: 10.3390/bios14050259

70. Redhya M, Jayalakshmi M. An Ensembled Grid based Machine Learning Approach For PD Classification From MRI Images. In: *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. IEEE; 2024. p. 1-6.

71. Zhang Z, Peng J, Song Q, Xu Y, Wei Y, Shu Z. Identification of depression subtypes in parkinson's disease patients via structural MRI whole-brain radiomics: An unsupervised machine learning study. *CNS Neurosci Ther*. 2025;31(2):e70182.

doi: 10.1111/cns.70182

72. Alharbi RA. Proteomics approach and techniques in identification of reliable biomarkers for diseases. *Saudi J Biol Sci*. 2020;27(3):968-974.

doi: 10.1016/j.sjbs.2020.01.020

73. Wang Z, Becker K, Donadio V, *et al.* Skin α-synuclein aggregation seeding activity as a novel biomarker for Parkinson disease. *JAMA Neurol*. 2021;78:1-11.

doi: 10.1001/jamaneurol.2020.3311

74. Lin CH, Chiu SI, Chen TF, Jang JSR, Chiu MJ. Classifications of neurodegenerative disorders using a multiplex blood biomarkers-based machine learning model. *Int J Mol Sci*. 2020;21(18):6914.

doi: 10.3390/ijms21186914

75. Maass F, Michalke B, Willkommen D, *et al*. Elemental fingerprint: Reassessment of a cerebrospinal fluid biomarker for Parkinson's disease. *Neurobiol Dis*. 2020;134:104677.

doi: 10.1016/j.nbd.2019.104677

76. Wang W, Lee J, Harrou F, Sun Y. Early detection of Parkinson's disease using deep learning and machine learning. *IEEE Access*. 2020;8:147635-147646.

77. Chung CC, Chan L, Chen JH, Bamodu OA, Chiu HW, Hong CT. Plasma extracellular vesicles tau and β-amyloid as biomarkers of cognitive dysfunction of Parkinson's disease. *FASEB J*. 2021;35(10):e21895.

doi: 10.1096/fj.202100787R

78. Vacchi E, Burrello J, Burrello A, *et al.* Profiling inflammatory extracellular vesicles in plasma and cerebrospinal fluid: An optimized diagnostic model for Parkinson's disease. *Biomedicines*. 2021;9(3):230.

doi: 10.3390/biomedicines9030230

79. Amboni M, Ricciardi C, Adamo S, *et al*. Machine learning can predict mild cognitive impairment in Parkinson's disease. *Front Neurol*. 2022;13:1010147.

doi: 10.3389/fneur.2022.1010147

80. Chen PH, Hou TY, Cheng FY, Shaw JS. Prediction of cognitive degeneration in Parkinson's disease patients using a machine learning method. *Brain Sci*. 2022;12(8):1048.

doi: 10.3390/brainsci12081048

81. Dadu A, Satone V, Kaur R, *et al.* Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *NPJ Parkinsons Dis*. 2022;8(1):172.

doi: 10.1038/s41531-022-00439-z

82. Harvey J, Reijnders RA, Cavill R, *et al*. Machine learning-based prediction of cognitive outcomes in *de novo* Parkinson's disease. *NPJ Parkinsons Dis*. 2022;8(1):150.

doi: 10.1038/s41531-022-00409-5

83. Pahuja G, Prasad B. Deep learning architectures for Parkinson's disease detection by using multi-modal features. *Comput Biol Med*. 2022;146:105610.

doi: 10.1016/j.compbiomed.2022.105610

84. Yang Y, Yuan Y, Zhang G, *et al*. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nat Med*. 2022;28(10):2207-2215.

doi: 10.1038/s41591-022-01932-x

85. Allwright M, Mundell H, Sutherland G, Austin P, Guennewig B. Machine learning analysis of the UK Biobank reveals IGF-1 and inflammatory biomarkers predict Parkinson's disease risk. *PLos One*. 2023;18(5):e0285416.

doi: 10.1371/journal.pone.0285416

86. Almgren H, Camacho M, Hanganu A, *et al*. Machine learning-based prediction of longitudinal cognitive decline in early Parkinson's disease using multimodal features. *Sci Rep.* 2023;13(1):13193.

87. Kelly J, Moyeed R, Carroll C, Luo S, Li X. Blood biomarker-based classification study for neurodegenerative diseases. *Sci Rep.* 2023;13(1):17191.

doi: 10.1038/s41598-023-43956-4

88. McFall GP, Bohn L, Gee M, *et al*. Identifying key multi-modal predictors of incipient dementia in Parkinson's disease: A machine learning analysis and Tree SHAP interpretation. *Front Aging Neurosci*. 2023;15:112432.

doi: 10.3389/fnagi.2023.1124232

89. Tsukita K, Sakamaki-Tsukita H, Kaiser S, *et al*. High-throughput CSF proteomics and machine learning to identify proteomic signatures for parkinson disease development and progression. *Neurology*. 2023;101(14):e1434-e1447.

doi: 10.1212/WNL.0000000000207725

90. Chen H, Guo S, Zhuang Z, *et al.* Intelligent identification of cerebrospinal fluid for the diagnosis of Parkinson's disease. *Anal Chem*. 2024;(6):2534-2542.

doi: 10.1021/acs.analchem.3c04849

91. Dennis AGP, Strafella AP. The identification of cognitive impairment in Parkinson's disease using biofluids, neuroimaging, and artificial intelligence. *Front Neurosci*. 2024;18:1446878.

doi: 10.3389/fnins.2024.1446878

92. Hällqvist J, Bartl M, Dakna M, *et al.* Plasma proteomics identify biomarkers predicting Parkinson's disease up to 7 years before symptom onset. *Nat Commun*. 2024;15(1):4759.

doi: 10.1038/s41467-024-48961-3

93. Callegari S, Kirk NS, Gan ZY, *et al*. Structure of human PINK1 at a mitochondrial TOM-VDAC array. *Science*. 2025;388:303-310.

doi: 10.1126/science.adu6445

94. Falchetti M, Prediger RD, Zanotto-Filho A. Classification algorithms applied to blood-based transcriptome meta-analysis to predict idiopathic Parkinson's disease. *Comput Biol Med*. 2020;124:103925.

doi: 10.1016/j.compbiomed.2020.103925

95. Su C, Tong J, Wang F. Mining genetic and transcriptomic data using machine learning approaches in Parkinson's disease. *NPJ Parkinsons Dis*. 2020;6(1):24.

96. García-Fonseca Á, Martin-Jimenez C, Barreto GE, Pachón AFA, González J. The emerging role of long non-coding RNAs and microRNAs in neurodegenerative diseases: A perspective of machine learning. *Biomolecules*. 2021;11(8):1132.

doi: 10.3390/biom11081132

97. Hu S, Li S, Ning W, *et al*. Identifying crosstalk genetic biomarkers linking a neurodegenerative disease, Parkinson's disease, and periodontitis using integrated bioinformatics analyses. *Front Aging Neurosci*. 2022;14:1032401.

doi: 10.3389/fnagi.2022.1032401

98. Lam S, Arif M, Song X, Uhlén M, Mardinoglu A. Machine learning analysis reveals biomarkers for the detection of neurological diseases. *Front Mol Neurosci*. 2022;15:889728.

doi: 10.3389/fnmol.2022.889728

99. Makarious MB, Leonard HL, Vitale D, *et al*. Multi-modality machine learning predicting Parkinson's disease. *NPJ Parkinsons Dis*. 2022;8(1):35.

doi: 10.1038/s41531-022-00288-w

100. Pantaleo E, Monaco A, Amoroso N, *et al.* A machine learning approach to Parkinson's disease blood transcriptomics. *Genes* (*Basel*). 2022;13(5):727.

doi: 10.3390/genes13050727

101. Vuidel A, Cousin L, Weykopf B, *et al.* High-content phenotyping of Parkinson's disease patient stem cell-derived midbrain dopaminergic neurons using machine learning classification. *Stem Cell Reports*. 2022;17(10):2349-2364.

doi: 10.1016/j.stemcr.2022.09.001

102. Cai L, Tang S, Liu Y, Zhang Y, Yang Q. The application of weighted gene co-expression network analysis and support vector machine learning in the screening of Parkinson's disease biomarkers and construction of diagnostic models. *Front Mol Neurosci*. 2023;16:1274268.

doi: 10.3389/fnmol.2023.1274268

103. Hajianfar G, Kalayinia S, Hosseinzadeh M, *et al.* Prediction of Parkinson's disease pathogenic variants using hybrid Machine learning systems and radiomic features. *Phys Med*.

2023;113:102647.

doi: 10.1016/j.ejmp.2023.102647

104. Wang P, Chen Q, Tang Z, *et al.* Uncovering ferroptosis in Parkinson's disease via bioinformatics and machine learning and reversed deducing potential therapeutic natural products. *Front Genet*. 2023;14:1231707.

doi: 10.3389/fgene.2023.1231707

105. Xin G, Niu J, Tian Q, *et al*. Identification of potential immune-related hub genes in Parkinson's disease based on machine learning and development and validation of a diagnostic classification model. *PLoS One*. 2023;18(12):e0294984.

doi: 10.1371/journal.pone.0294984

106. Zhang Y, Sun X, Zhang P, *et al.* Identification of Parkinson's disease associated genes through explicable deep learning and bioinformatic. In: *International Conference on Applied Intelligence*. Berlin: Springer Nature; 2023. p. 136-146.

107. Ameli A, Peña-Castillo L, Usefi H. Assessing the reproducibility of machine-learning-based biomarker discovery in Parkinson's disease. *Comput Biol Med*. 2024;174:108407.

doi: 10.1016/j.compbiomed.2024.108407

108. Banou E, Vrahatis AG, Krokidis MG, Vlamos P. Machine learning analysis of genomic factors influencing hyperbaric oxygen therapy in Parkinson's disease. *BioMedInformatics*. 2024;4(1):127-138.

109. Kumar A, Kouznetsova VL, Kesari S, Tsigelny IF. Parkinson's disease diagnosis using miRNA biomarkers and deep learning. *Front Biosci (Landmark Ed)*. 2024;29(1):4.

doi: 10.31083/j.fbl2901004

110. Peng H, Cheng Y, Chen Q, Qin L. Integrated transcriptomic and machine learning analysis identifies EAF2 as a diagnostic biomarker and key pathogenic factor in Parkinson's disease. *Int J Gen Med*. 2024;17:5547-5562.

doi: 10.2147/IJGM.S486214

111. Teng WB, Deng HW, Lv BH, Zhou SD, Li BR, Hu RT. Exploring and validating key genetic biomarkers for diagnosis of Parkinson's disease. *Brain Res Bull*. 2025;220:111165.

doi: 10.1016/j.brainresbull.2024.111165

112. Yan J, Wang Z, Li Y, Li R, Xiang K. m6A-related genes and their role in Parkinson's disease: Insights from machine learning and consensus clustering. *Medicine (Baltimore)*. 2024;103(45):e40484.

doi: 10.1097/MD.0000000000040484

113. Yang W, Xu S, Zhou M, Chan P. Aging-related biomarkers for the diagnosis of Parkinson's disease based on bioinformatics analysis and machine learning. *Aging (Albany NY)*. 2024;16(17):12191-12208.

doi: 10.18632/aging.205954

114. Yu E, Larivière R, Thomas RA, *et al*. Machine learning nominates the inositol pathway and novel genes in Parkinson's disease. *Brain*. 2024;147(3):887-899.

doi: 10.1093/brain/awad345

115. Pietrucci D, Teofani A, Unida V, *et al.* Can gut microbiota be a good predictor for Parkinson's disease? A machine learning approach. *Brain Sci*. 2020;10(4):242.

doi: 10.3390/brainsci10040242

116. Qian Y, Yang X, Xu S, *et al.* Gut metagenomics-derived genes as potential biomarkers of Parkinson's disease. *Brain*. 2020;143(8):2474-2489.

doi: 10.1093/brain/awaa201

117. Lubomski M, Xu X, Holmes AJ, *et al*. Nutritional intake and gut microbiome composition predict Parkinson's disease. *Front Aging Neurosci*. 2022;14:881872.

doi: 10.3389/fnagi.2022.881872

118. Nie S, Wang J, Deng Y, Ye Z, Ge Y. Inflammatory microbes and genes as potential biomarkers of Parkinson's disease. *NPJ Biofilms Microbiomes*. 2022;8(1):101.

doi: 10.1038/s41522-022-00367-z

119. Nishiwaki H, Ito M, Hamaguchi T, *et al*. Short chain fatty acids-producing and mucin-degrading intestinal bacteria predict the progression of early Parkinson's disease. *NPJ Parkinsons Dis*. 2022;8(1):65.

doi: 10.1038/s41531-022-00328-5

120. Sánchez XR. *Machine Learning on Gut Microbiome Reveals Potential Biomarkers for Parkinson's Diagnosis. 2022*. Bachelor's Degree in Bioinformatics (UPF-UPC-UB-UAB) Final Grade Project, Date: June 21, 2022, School of International Studies, Universitat Pompeu Fabra (UPF), Barcelona, Spain. Available from: https://repositori-api.upf.edu/api/core/bitstreams/ddb33240-c4a2-4725-999b-ddfeba0e646a/content [Last accessed on 2025 Apr 21].

121. Boodaghidizaji M, Jungles T, Chen T, *et al*. Machine learning based gut microbiota pattern and response to fiber as a diagnostic tool for chronic inflammatory diseases. *BMC Microbiol*. 2025;25(1):353.

doi: 10.1186/s12866-025-04072-7

122. Dhatrak M. *Application of Supervised Learning Classifiers on Gut Microbial Data to Predict Parkinson Disease (Doctoral Dissertation)*; 2023.

123. Li M, Liu J, Zhu J, *et al.* Performance of gut Microbiome as an independent diagnostic tool for 20 diseases: Cross-cohort validation of machine-learning classifiers. *Gut Microbes*. 2023;15(1):2205386.

doi: 10.1080/19490976.2023.2205386

124. Romano S, Wirbel J, Ansorge R, *et al.* Machine learning-based meta-analysis reveals gut microbiome alterations associated with Parkinson's disease. *Nat Commun*. 2025;16(1):4227.

doi: 10.1038/s41467-025-56829-3

125. Zhang JD, Xue C, Kolachalama VB, Donald WA. Interpretable machine learning on metabolomics data reveals biomarkers for Parkinson's disease. *ACS Cent Sci*. 2023;9(5):1035-1045.

doi: 10.1021/acscentsci.2c01468

126. Li Y, Ren HX, Chi CY, Miao YB. Artificial intelligence-guided gut-microenvironment-triggered imaging sensor reveals potential indicators of Parkinson's disease. *Adv Sci (Weinh)*. 2024;11(23):e2307819.

doi: 10.1002/advs.202307819

127. Zhao Z, Chen J, Zhao D, *et al*. Microbial biomarker discovery in Parkinson's disease through a network-based approach. *NPJ Parkinsons Dis*. 2024;10(1):203.

doi: 10.1038/s41531-024-00802-2

128. Rojas-Velazquez D, Kidwai S, Liu TC, *et al*. Understanding Parkinson's: The microbiome and machine learning approach. *Maturitas*. 2025;193:108185.

doi: 10.1016/j.maturitas.2024.108185

129. Yu B, Zhang H, Zhang M. Deep learning-based differential gut flora for prediction of Parkinson's. *PLoS One*. 2025;20(1):e0310005.

doi: 10.1371/journal.pone.0310005

130. Aijaz M, Ramesh Kumar A, Ahmad S, Raja MS, Gupta RK, Yadav R. The role of Artificial Intelligence in Parkinson's disease: A comprehensive review. *World J Adv Pharm Med Res*. 2024;10(9):270-279.

131. Ienca M, Ignatiadis K. Artificial intelligence in clinical neuroscience: Methodological and ethical challenges. *AJOB Neurosci*. 2020;11(2):77-87.

doi: 10.1080/21507740.2020.1740352

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

# Recent advances in genetic feature marker discovery through differential expression and biostatistical analysis

**Ankita Saha[1,2], Shibakali Gupta[3], Chyan Paul[4], Saurav Mallik[5,6]*, and Korhan Cengiz[7]***

[1]Department of Computer Science, Swami Vivekananda University, Barrackpore, West Bengal, India

[2]Department of Science and Management, ABS Academy of Management and Health Science, Durgapur, West Bengal, India

[3]Department of Computer Science and Engineering, University Institute of Technology, Burdwan University, West Bengal, India

[4]Department of Computer Science and Engineering, Swami Vivekananda University, Barrackpore, West Bengal, India

[5]Department of Biostatistics, University of Miami, Florida, United States of America

[6]College of Pharmacy, University of Arizona, Tucson, Arizona, United States of America

[7]Department of Electrical Engineering, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia

## Abstract

Genetic feature discovery is essential for understanding complex diseases and traits. This comprehensive review provides an in-depth comparison of differential expression analysis methods and statistical hypothesis tests—such as Student's *t*-test, Chi-square test, analysis of variance, Empirical Bayes methods, and Significant Analysis of Microarrays—used in genetic feature marker discovery. Our analysis highlights the strengths and weaknesses of these approaches in terms of methodologies, applications, performance, and accuracy. While the statistical tests provide straightforward interpretation, machine learning techniques provide superior capabilities for handling high-dimensional data and complex biological interactions. We conducted two mini-experiments: (i) Identification of differentially expressed genes, upregulated genes and downregulated genes using statistical tools (i.e., Student's t-test and Welch's t-test) under different conditions (normalization methods and *p*-value correction strategies) using the GSE31699 dataset from the NCBI Gene Expression Omnibus, and (ii) gene set enrichment analysis—covering Kyoto Encyclopedia of Genes and Genomes pathways and Gene Ontology terms like Biological process, Cellular component and Molecular function—using the GSE30760 dataset with the DAVID 2021 tool. Furthermore, we discussed the potential of hybrid approaches combining statistical tests with machine learning and optimization techniques for enhanced feature discovery. Future work will focus on multi-omics data integration, the development of explainable AI methods, and scalable algorithms. This review aims to serve as a comprehensive guide for researchers involved in genetic marker identification, highlighting both statistical and computational perspectives on differential expression and gene set enrichment studies.

***Keywords:*** Genetic feature discovery; Statistical tests; KEGG pathway analysis; Gene set enrichment analysis

***Corresponding authors:**
Saurav Mallik (sauravmtech2@gmail.com);
Korhan Cengiz (kcengiz@pmu.edu.sa)

**Citation:** Saha A, Gupta S, Paul C, Mallik S, Cengiz K. Recent advances in genetic feature marker discovery through differential expression and biostatistical analysis. *Artif Intell Health*. 2026;3(1):54-70. doi: 10.36922/AIH025180036

**Received:** April 28, 2025

**Revised:** July 17, 2025

**Accepted:** August 1, 2025

**Published online:** September 9, 2025

## 1. Introduction

Genetic feature discovery is a crucial step in understanding complex diseases and traits. It involves identifying important gene variants associated with specific phenotypes. The rapid progress in high-throughput genomics techniques has led to an exponential growth in genomic information, rendering traditional statistical methods inadequate for analyzing these vast datasets. Consequently, machine learning approaches have emerged as powerful tools in genetic feature discovery, offering improved performance and accuracy. However, the choice between statistical tests and machine learning methods remains a subject of debate among researchers. Statistical tests—such as t-tests, analysis of variance (ANOVA), and Chi-square tests—have been the cornerstone of genetic feature discovery for decades. These methods provide straightforward interpretation and hypothesis testing, making them appealing for identifying significant associations between genetic variants and traits. Nevertheless, their limitations become apparent when dealing with high-dimensional data, complex interactions, and multiple testing corrections.

Machine learning approaches, including random forests, support vector machines, and neural networks, have revolutionized genetic feature discovery by handling complex relationships and high-dimensional data. These methods excel in identifying patterns and interactions that may elude traditional statistical tests. However, the "black box" nature of many machine learning models often obscures interpretability, making it challenging to understand the underlying biological mechanisms. The integration of statistical tests and machine learning methods has emerged as a promising strategy for leveraging the strengths of both approaches. Hybrid methods can combine the hypothesis-driven framework of statistical tests with the pattern-recognition capabilities of machine learning, leading to improved feature discovery and biological interpretation.

This review aims to provide a comprehensive comparison of statistical tests and machine learning approaches in genetic feature discovery. We examine the methodologies, applications, advantages, limitations, and future directions of both paradigms, while highlighting the potential of hybrid methods and emerging trends in multi-omics integration, explainable AI, and scalable algorithms. By bridging the gap between statistical and machine learning methods, this review seeks to serve as a valuable resource for advancing genetic feature discovery and unravelling the complexities of diseases and trait etiology.

## 2. Fundamentals of the central dogma of molecular biology and drug discovery

Biomedical research is a magnificent field of science that strives to uncover pathways to confining and treating diseases that cause morbidity and death in living things. This experimental domain covers numerous scientific disciplines and relies on rigorous exploration by scientists, chemists, and biologists. The discovery of new drugs and treatments requires robust scientific testing and thorough evaluation. Researchers in this field have the responsibility to conduct their work in a beneficent, prudent, and proper manner.[1] To address difficult biomedical problems, researchers use bioinformatics—an interdisciplinary field that integrates computational tools for analyzing biomedical data.[2]

Bioinformatics has emerged from the convergence of several disciplines, including computer science, biology, mathematics, statistics, and others. Alongside related fields like computational biology and biochemistry, bioinformatics has expanded significantly in recent years, driven by the growing need to understand complex biological systems. Defining these emerging disciplines has posed a challenge to researchers and educators alike. Among them, bioinformatics has had a particularly profound impact on the medical field. It also plays an essential role in areas such as space exploration, agriculture, and more. Broadly defined, bioinformatics is the integration of computer science, statistics, biology, and mathematics to collect, organize, analyze, and interpret biological data. This integration enables the development of software applications for analyzing DNA sequences, proteins, evolutionary genetics, biomolecular interactions, and biological networks, as well as managing datasets derived from genomic, proteomic, and post-genomic studies.[3-5]

### 2.1. Central dogma of molecular biology

In molecular biology, the term "central dogma" plays a pivotal role in biomarker discovery and hub gene selection. It comprises three basic processes: transcription, translation, and replication. The process of converting DNA (deoxyribonucleic acid) to RNA (ribonucleic acid) is termed transcription, whereas the process of transforming RNA to protein is called translation (Figure 1). The process of duplicating DNA is denoted as DNA replication.

There are different kinds of RNAs, such as messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). Among these, mRNA plays a major role in the detection and prognosis of various diseases and disorders like tissue-specific cancer, Alzheimer's disease, and other neurodegenerative diseases.[6,7] Aberrant gene expression,

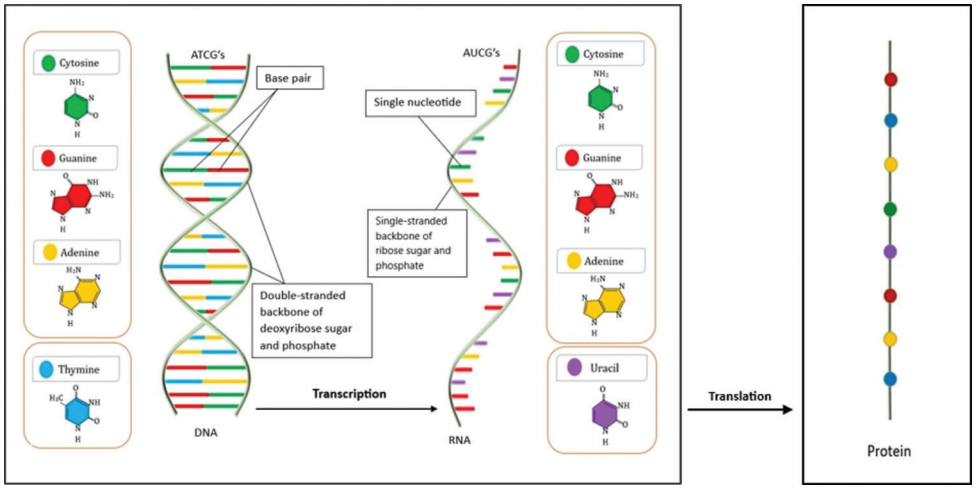



**Figure 1.** Central dogma of molecular biology. Image created by the authors.

whether significantly increased (up-regulation) or decreased (down-regulation), can contribute to disease pathogenesis in both humans and model organisms. The statistical significance of differential gene expression is typically assessed using *p*-values derived from hypothesis testing. A *p*-value represents the probability of observing an effect purely by chance.[8,9] The magnitude of gene expression changes is usually quantified using the fold change (FC) method.[10,11] There are two kinds of statistical hypothesis tests: parametric tests and non-parametric tests. Parametric tests (described in Section 2.4.1) are suitable for normally distributed data, whereas non-parametric tests (described in Section 2.4.2) are more appropriate when the assumption of normality is violated.[12,13]

## 2.2. Pre-test analysis

Before performing statistical analysis, it is important to apply pre-filtering techniques, normalize the data, and test for normality. These steps reduce analytical error and improve the reliability of results.

### *2.2.1. Pre-test filtering procedure*

High-dimensional datasets typically require several statistical tests, which can reduce statistical power and inflate error rates if not properly controlled. Various filtering methods are utilized to evaluate differential gene expression before implementing a formal statistical test. A basic approach is to examine the all-inclusive variance of every gene (row-wise) and retain only genes with minimal variance. For example, the "genevarfilter" function in MATLAB allows users to exclude genes with low variance, based on user-defined percentiles (e.g., 5$^{th}$, 30$^{th}$, or 45$^{th}$ percentile). Genes with low variance may yield seemingly significant *p*-values, despite lacking true biological relevance. Pre-filtering such genes can help reduce false positives. In the Limma package, a variance-based filtering approach is integrated into the t-test framework to minimize spurious detections, especially in studies with small sample sizes.[14] Proper filtering is beneficial only when the false positive rate is adequately controlled. However, improper filtering—especially if not aligned with class labels—can adversely affect the control of Type I errors.

Consider a data matrix of dimension $m \times n$, where $m$ indicates total number of genes and $n$ indicates the number of samples. Let the expression data for gene $i$ be denoted by $Y_i = (Y_{i1},\ldots,Y_m)^t$. If $Y_{i1},\ldots,Y_m$ are independently and normally distributed for every $i \in H_0$ (i.e., under the null hypothesis), then the test statistics before and after filtering are marginally independent. This implies that unadjusted *p*-values remain valid after filtering and applying the two-sample t-test. Thus, the un-adjusted *p*-value will be correct after applying two test statistics and filtering. When the sample size is bigger, the implementation of an experimental null distribution helps ensure accurate estimation of conditional effects introduced by filtering.

Indeed, if the null hypothesis is false, the test statistics and the filtering criterion are not necessarily independent. Filtering enhances detection power only if the test statistic and filtering criterion are positively correlated under the alternative hypothesis.

FC represents another filtering approach. Tools like "volcanoplot[1]" integrate FC thresholds with test statistics to identify significantly differentially expressed genes. Genes with FC values exceeding a lower threshold (up-regulated) or below an upper threshold (down-

[1] More information is available online at: http://www.mathworks.in/help/bioinfo/ref/mavolcanoplot.html

regulated) are considered as having passed the filter. The Bioconductor package "genefilter[2]" provides additional tools for implementing pre-filtering strategies.

Another commonly used method involves filtering based on intensity variation or the highest within-class mean. Consider two classes (i.e., disease vs. control). If the dataset follows a normal distribution (Gaussian distribution) with known common variance $\sigma_2$, the within-class variance for the mean is $\overline{\sigma}_2 = 2\sigma_2 / n$. Genes for which the maximum of the average expression across the two groups exceeds a threshold $u^*$ (i.e., $\max\{\overline{Y}_{i,1_g}, \overline{Y}_{i,2_g}\} > u^*$ are retained for further analysis.

#### 2.2.2. Data normalization techniques

Once the pre-filtering approach is complete, gene expression data must be normalized to bring measurements from various scales to a common scale. Gene-wise normalization techniques such as zero mean normalization, median normalization, and min-max normalization are commonly applied.[15,16] Other standard techniques include statistical column normalization,[15] variance stabilizing normalization,[17,18] and quantile normalization.[17,19]

#### 2.2.3. Normality tests (NT) in data analysis

Following normalization, it is critical to apply NT[20] to each gene's expression data to assess whether the dataset conforms to a normal distribution, which may affect the accuracy of findings. Confirming normality is important for ensuring the assumptions underlying parametric statistical tests are met. Several methods are available for normality testing,[21] including the Jarque-Bera (JB) test,[22,23] Shapiro-Wilk test,[24] Anderson-Darling (AD) test,[24] Kolmogorov–Smirnov (KS) test,[24] and Lilliefors test.[24] Based on the results of these tests, parametric tests may be used for data that follow a normal distribution, whereas non-parametric tests are more appropriate for non-normally distributed data. For datasets with very small sample sizes (e.g., 1–5 samples), statistical testing may not be meaningful. In such cases, only FC methods can be applied to assess gene expression differences.

The JB test is used to evaluate whether a dataset exhibits characteristics of a normal distribution based on skewness and kurtosis. Skewness measures asymmetry, while kurtosis quantifies the "tailedness" or sharpness of the distribution peak.[25] This test does not require prior calculation of mean or standard deviation to be methodically implemented.[26] First proposed by Carlos Jarque and Anil Bera in 1980, it has become a standard method for testing normality in statistical research. The JB test statistic is defined in Equation I.

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \quad \text{(I)}$$

where n, *S*, and *K* denote sample size, sample skewness, and sample kurtosis, respectively.

For 2,000 or more sample sizes, the test statistic is compared with the Chi-squared distribution[27] with 2 degrees of freedom. If the computed statistics are larger than the critical Chi-squared value, normality is rejected. Chi-squared estimation demands a large sample size for accurate results.[28] Thus, simulation-based methods are used when sample sizes are below 2000. Typically, 100,000 normally distributed samples—generated with similar mean value and standard deviation (SD) as the original data—are used to estimate a reference distribution for the JB test statistic.

### 2.3. FC

FC is a basic yet widely used method for determining gene expression patterns.[29] According to the literature, two definitions of FC are available.[11] For real-time expression values, the FC for gene g is calculated as the ratio of average expression values between two groups, as shown in Equation II.

$$FC = \frac{\overline{x}_{1_g}}{\overline{x}_{2_g}} \quad \text{(II)}$$

Where $\overline{x}_{1_g}$ represents the average expression value of gene g in the experimental (case) group, and $\overline{x}_{2_g}$ represents the average expression value in the control (normal) group.

### 2.4. Statistical test

A statistical test is a method used to draw conclusions regarding the larger population by examining a smaller set of samples.[12] It involves using statistical techniques to analyze the data and determine whether the results are due to chance or if they are statistically significant.

#### 2.4.1. Parametric distributions in statistics

Parametric testing methods, known as traditional or classical statistical tests, assume that the samples are drawn from normally distributed populations with similar variances across groups.[13] If the data do not fit these assumptions, nonparametric statistical tests are applied. Parametric tests are typically based on the mean expression value of genes.[30,31] Various commonly used parametric tests are briefly discussed below.

One of the most common statistical tests is the Student's *t*-test,[32] particularly the two-sample t-test, which is used to assess differences between the means of two independent

[2] The tool can be accessed at: http://www.bioconductor.org/packages/2.12/bioc/html/genefilter.html

groups. This test calculates a *p*-value using the cumulative distribution function of the t-distribution. The *p*-value measures the probability of observing a t-value as extreme as, or more extreme than, the observed one under the null hypothesis. According to conventional statistical thresholds, a $p<0.05$ indicates a statistically significant difference.

Let us assume, for a given gene g: group 1: $n_1$ experimental samples, $mean = \bar{x}_{1_g}$, and $SD = s_{1_g}$; and group 2: $n_2$ control samples, $mean = \bar{x}_{2_g}$, and $SD = s_{2_g}$. The t-statistic ($t$) can be calculated using Equation III, and the standard error of the difference in means ($se_g$) is calculated using Equation IV. The pooled standard deviation is computed using Equation V.

$$t = \frac{(\bar{x}_{1_g} - \bar{x}_{2_g})}{se_g} \tag{III}$$

$$\left( se_g = spooled * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \tag{IV}$$

$$\left( spooled = \sqrt{\frac{(n_1 - 1) * s_{1_g}^2 + (n_2 - 1) * s_{2_g}^2}{(n_1 + n_2 - 2)}} \right) \tag{V}$$

where $\bar{x}_{1_g}$ and $\bar{x}_{2_g}$ are the means of group 1 and group 2; $s_{1_g}$ and $s_{2_g}$ are the SDs of group 1 and group 2; and $n_1$ & $n_2$ are the sample sizes of group 1 and group 2.

The degrees of freedom for this test are given by $n_1 + n_2 - 2$. In this method, the populations are assumed to have equal variances. Testing or estimating variability of two different groups is called as "Behrens-Fisher" problem.[33,34] This problem arises when comparing means from two normally distributed but heteroscedastic populations. To address this issue, it is important to test whether the variance of each population is equivalent to the others. If they are unequal, the Welch's t-test should be used, which does not assume equal population variances. In this case, the $t$ can be calculated using Equation VI, which provides an unpooled estimate of the population standard deviation.[35]

$$t = \left( (\bar{x}_{1_g} - \bar{x}_{2_g}) \Big/ \sqrt{(s_{1_g}^2 / n_1) + (s_{2_g}^2 / n_2)} \right) \tag{V}$$

where $\bar{x}_{1_g}$ and $\bar{x}_{2_g}$ are the means of group 1 and group 2; $s_{1_g}$ and $s_{2_g}$ are the SDs of group 1 and group 2; and $n_1$ & $n_2$ are the sample sizes of group 1 and group 2.

The assumption of normality assumption (NA) may not always hold in gene expression data. Let each gene g have a common variance $\sigma^2$, such that the class-specific mean variance becomes $\bar{\sigma}^2 = 2\sigma^2 / n$. To pre-filter genes, one might compute a test statistic such as: $U_g^I = max\{\bar{x}_{1_g}, \bar{x}_{2_g}\}$ and retain genes for which $U_g^I$ exceeds the user-defined threshold $u^*$. Alternatively, the test statistic, which resembles a standardized t-statistic with known variance, can be used: $U_g^{II} = (\bar{x}_{1_g} - \bar{x}_{2_g}) / \sqrt{2}\bar{\sigma}$. This allows for more robust detection of differentially expressed genes while explicitly incorporating variance assumptions.

### 2.4.2. Primary non-parametric statistical techniques

Non-parametric methods, also known as distribution-free methods, do not assume any specific underlying data distribution. These techniques are particularly useful when data violate the assumptions of parametric tests, such as normality or equal variance. Below are some commonly used non-parametric statistical methods for identifying differentially expressed genes.

The Wilcoxon rank-sum test (RST) is widely used for small sample sizes or when the data are not normally distributed. In such cases, the t-test may not be reliable, and RST provides a robust alternative. This test ranks all values from both groups together and then calculates the sum of ranks for each group.

Let $W_1 = \sum ranks_{group1}$ and $W_2 = \sum ranks_{group2}$. If the sample sizes $n_1$ and $n_2$ of group 1 and group 2 are equal, the test statistics $T$ is defined as $T = min(T_1, T_2)$, where $T_1$ and $T_2$ are the rank sums of each group. If the sizes are not equal, the statistic is computed as follows: $T_2 = n_1 (n_1 + n_2 + 1) - T_1$. A significantly lower value of $T$ suggests rejecting the null hypothesis of equal sample means. For small samples, critical values of $T$ are tabulated. The z-score for each gene is computed using Equations VI, VII, and VIII.

$$z = \frac{(\left| T - mean_{w_1} \right| - 0.5)}{\sqrt{var_{w_1}}} \tag{VI}$$

$$var_{w_1} = n_2 * mean_{w_1} / 6 = n_1 * n_2 * (n_1 + n_2 + 1) / 12 \tag{VII}$$

$$mean_{w_1} = n_1 * (n_1 + n_2 + 1) / 2 \tag{VIII}$$

The RST and Mann-Whitney U test are mathematically equivalent, but RST is computationally slower.[36,37] The test statistic is given in Equation IX.

$$z = \frac{\left( u_1 - mean_{u_1} \right)}{\sqrt{var_{u_1}}} \tag{IX}$$

where $var_{u_1} = n_1 * n_2 * (n_1 + n_2 + 1) / 12$ and $mean_{u_1} = n_1 * n_2 / 2$. Here, $u_1 = T_1 - n_1 * (n_1 + 1)/2$ and $T1 = \sum ranks_{group1}$.

One limitation of the *t*-test is its sensitivity to small standard errors, especially in low-expression genes, which can yield artificially large test statistics. To overcome this, SAM introduces a small positive constant $s_0$ (also called a "fudge factor") to stabilize variance.[38] The SAM statistics, as proposed by Tusher *et al.*,[39] are described in Equation X.

$$t_{sam} = \frac{\bar{x}_{1_g} - \bar{x}_{2_g}}{se_g + s_0} \quad (X)$$

where $se_g = sPooled * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is the standard error of the group means value

### 2.4.3. Other non-parametric tests

Introduced by Kruskal and Wallis, the KW test is utilized to determine whether multiple independent samples follow the same distribution.[40-42] It is the non-parametric equivalent of one-way ANOVA.

Another common test is the ideal discriminator (ID) method,[42] which is a resampling-based technique used to identify genes that are maximally expressed in one group and minimally expressed in another.[43] The technique justifies significance by selecting the genes (or miRNAs) that have the maximum Pearson's correlation coefficient (PCC) relative to the ID. Significance is determined by comparing the observed PCC obtained from permutations of 50,000 random columns of data.

The KS test[43,44] is one of the useful non-parametric tests utilized to assess equality of continuous, one-dimensional probability distributions.[44,45] It compares either an empirical distribution with a reference cumulative distribution, or two empirical distributions. The test statistic represents the maximum deviation between the two distributions and is sensitive to differences in both location and shape. The KS test is particularly useful for evaluating distributional differences between two samples. Its asymptotic *p*-values are reliable when: $(n_1 * n_2)/(n_1 + n_2)4$, where $n_1$ and $n_2$ are the sample sizes of the two groups.

## 2.5. Testing errors and performance metrics

This section describes different types of hypothesis testing errors in statistics, followed by performance metrics used to evaluate statistical methods. Two popular types of errors in statistics are type I and type II errors (Figure 2),[45,46] which describe specific flaws in experimental inference.[46,47]

Let us assume, $m$ is the number of null hypotheses (e.g., genes), $n$ is the number of conditions (samples), and $R$ is the number of rejected null hypotheses. A Type I error occurs when the null hypothesis $H_0$ is incorrectly rejected, even though it is actually true. The probability of

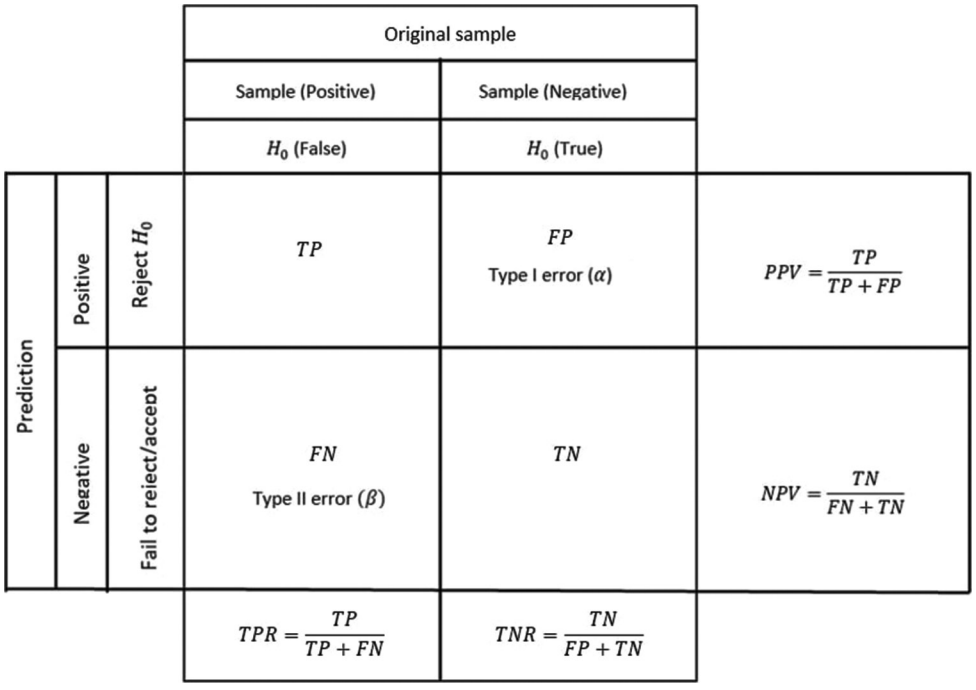



**Figure 2.** Relationship between hypothesis test outcome and error types. Image created by the authors.
Abbreviations: α: Type I error; β: Type II error; $H_o$: Null hypothesis; FN: False negative; FP: False positive; NPV: Negative predictive value; PPV: Positive predictive value; TN: True negative; TP: True positive; TNR: True negative rate; TPR: True positive rate.

committing a Type I error is referred to as the statistical significance level, denoted by $\alpha$. Type I error rates can be characterize via four ways:[47] (1) Per comparison error rate (*PCER*), which is the expected proportion of false positive value (*E*) of Type I error (*EP*) is divided by the total number across all hypothesis (*m*) ($PCER = E(FP)/m$); (2) Per-family error rate (*PFER*), which is the expected number (*E*) of *FP* in the set $PFER = E(FP)$; (3) Family-wise error rate (*FWER*), which measures the probability (P) of one or more *FP* ($FWER = P(FP \geq 1)$); and (4) False discovery rate (*FDR*), introduced by Benjamini and Hochberg, is the expected proportion (E) of *FP* among the rejected hypothesis ($FDR = E(EP/(TP+FP)), R > 0$).[48]

Fundamentally, in multiple testing process, $PCFR \leq FWER \leq PFER$. Therefore, the *PFER* method is more conventional than the *FWER* technique, while *FWER* practice is more conventional the *PCER*. *PFER* produces more false positive errors than the *PWER*, while *FWER* produces more false positive errors than *PCER*.

A type II error occurs when the alternative hypothesis $H_1$ is rejected even though it is actually true. The probability of this error is denoted by $\beta$, and the power of the test is defined as $1-\beta$, representing the probability of selecting $H_1$ when it is true. In statistical classification—particularly in medicine and bioinformatics—there are two key performance indicators: (1) Sensitivity (true positive, TP, rate): proportion of correctly identified TPs. (2) Specificity (true negative, TN, rate): proportion of correctly identified TNs. These metrics are intrinsically linked to type I and type II errors and are summarized along with related metrics in Table 1.

In 1975, a biochemist named Brian W. Matthews introduced the Matthews correlation coefficient (MCC), which is a quality-centric measurement of binary classifications. It is especially suitable for imbalanced datasets, providing a single-valued metric that summarizes the confusion matrix.[49] MCC stands out as a robust and balanced metric, particularly well-suited for handling datasets of varying sizes and uneven class distributions, making it an effective tool for evaluating performance in such scenarios. MCC ranges from -1 (complete disagreement between prediction and actual outcome) and +1 (perfect prediction). An MCC value of 0 indicates an arbitrary prediction of the actual value. This statistic is also called the phi coefficient. The MCC[50] can be computed using Equation XI as follows.[51]

$$MCC = \frac{\left(TP * TN\right) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad \text{(XI)}$$

where TP is the true positives, TN is the true negatives, FP is the false positives, and FN is the false negatives.

### 2.6. Post-test method: Various *p*-value corrections

Multiple testing correction refers to the adjustment of *p*-values when statistical analyses are repeated multiple times on the same dataset. When each test is conducted at a 5% significance level, every individual test carries a 5% probability of committing a type I error, i.e., incorrectly rejecting the null hypothesis.[51] However, when many tests are conducted simultaneously, the FWER—the probability of committing at least one type I error—can exceed 5% and may reach approximately 30% in some scenarios.[52] Therefore, it is essential to control this cumulative error rate to avoid false discoveries.

To maintain the FWER at a desired level (commonly set at $\alpha = 0.05$), the significance threshold for individual tests must be adjusted to be more stringent.[53,54] *p*-value

**Table 1. Advantages and disadvantages of parametric and non-parametric tests**

| Test | Advantages | Disadvantages |
|---|---|---|
| FC | • Useful for a small number of samples (e.g., 1–2).<br>• Simpler biological interpretation. | • Prone to high FPs.<br>• Ignore variability and is highly affected by outliers. |
| Student's t-test | • Performs well with a large sample size from two groups with similar variance. | • Performs poorly with a small number of samples.<br>• FP rate increases with increasing variance.<br>• Not suitable for comparing more than two groups. |
| Mann–Whitney U test (Non-parametric) | • Robust to outliers and works well for non-normal data.<br>• Useful for ranked or ordinal data.<br>• Detects median difference.<br>• More efficient and fewer chances of mistakes. | • Less powerful than the t-test for normally distributed data.<br>• Calculations are more complicated, especially for small sample sizes.<br>• Not appropriate for comparing more than two groups. |
| SAM | • Avoids the problem of having small variances.<br>• Suitable for small sample sizes.<br>• Uses permutation to account for gene correlation<br>• Does not rely on parametric assumptions.<br>• Reports local FDR and links expression changes over time. | • Performance is inconsistent for small sample sizes.<br>• The correlation method for sample variance is not model-motivated. |

Abbreviations: FC: Fold change; FDR: False discovery rate; FP: False positive; SAM: Significance Analysis of Microarrays.

corrections are applied to account for the increased likelihood of FP resulting from multiple comparisons. In genomic data analysis, each gene or miRNA is typically tested independently, and the FP rate is directly proportional to both the number of comparisons and the chosen *p*-value threshold. GeneSpring and other statistical platforms categorize multiple testing correction methods into four main types.

#### 2.6.1. Bonferroni correction

The Bonferroni correction is a conservative method used to adjust *p*-values when various dependent or independent statistical tests are performed on an individual dataset.[50] This technique controls the FWER by dividing the desired significance level ($\alpha$) by the number of comparisons ($m$). Alternatively, it can be implemented by multiplying each unadjusted *p*-value by the number of hypotheses tested: *Adjusted* $p = p \times m$. If the adjusted *p*-value is still less than the chosen significance threshold (e.g., 0.05), the test result is considered statistically significant.[55] In this context, $m$ represents the total number of genes (or miRNAs) tested. The Bonferroni method is particularly effective for strong control of the FWER when many pairwise tests are involved. This method discards null hypothesis, $H_g$ if the unadjusted *p*-value is equal to or lesser than $\alpha/m$. The single-step Bonferroni-corrected *p*-value is calculated as $\tilde{p}_g = min(m\tilde{p}_g, 1)$. Then, based on Boole's inequality, the FWER is bounded following Equation XII.

$$\begin{aligned} FWER = Pr(FP \geq 1) &= \Pr\left(\bigcup_{g=1}^{m_0}\left\{\tilde{p}_g \leq \alpha\right\}\right) \\ &\leq \sum_{g=1}^{m_0} Pr\left(\tilde{p}_g \leq \alpha\right) \\ &\leq \sum_{g=1}^{m_0} Pr\left(p_g \leq \frac{\alpha}{m}\right) \leq \frac{m_0\alpha}{m} \end{aligned} \quad \text{(XII)}$$

Where $m_0$ implies the total number of the true null hypothesis and $p_g$ is the unadjusted *p*-value for gene *g*. This final inequality follows from the assumption that under the null hypothesis $H_g$, the probability $Pr\ (p_g \leq (x|H_g)) \leq x$ for $x \in [0,1]$.[56]

#### 2.6.2. Bonferroni-Holm (step-down) correction

Although similar to Bonferroni correction, the Bonferroni-Holm correction is slightly less stringent, offering improved statistical power.[57] In this method, the unadjusted *p*-value of each gene (or miRNA) is first sorted in increasing sequence.[58] Then, a sequence of comparisons is performed: (1) The smallest *p*-value is multiplied by the total number of hypotheses $m$; (2) the second smallest is multiplied by $m-1$; (3) the third by $m-2$, and so on.[55] This process continues until a *p*-value fails to meet the significance threshold (e.g., 0.05), at which point the procedure stops, and all remaining hypotheses are not rejected. Let us assume $Pr_1 \leq Pr_2 \leq \ldots Pr_m$ denote the observed unadjusted *p*-values and $Hr_1, Hr_2, \ldots, Hr_m$ indicates the null hypotheses. As stated by Holm (1979), the index can be defined as $g^* = min\left\{g : Pr_g > \frac{\alpha}{m-g+1}\right\}$. and the hypotheses $Hr_g$, where g = 1,…, $g^*-1$, are all rejected. If no such $g^*$ exist, then all hypotheses are rejected. Since the correction becomes progressively less stringent as the *p*-value increases, this method is uniformly more powerful than the Bonferroni correction.[59]

#### 2.6.3. Westfall and Young permutation method

Unlike Bonferroni and Holm procedures—which are single-step methods that adjust *p*-values independently—the Westfall and Young permutation method incorporates the dependency structure between tests.[60] This approach is particularly suitable for genomic data such as DNA microarrays, where expression levels of many genes are often highly correlated.[61]

The method follows a step-down process, similar to Holm's, but uses permutations to create resampled datasets. Specifically, the data are randomly partitioned into two artificial groups (e.g., control and treatment), and *p*-values for all genes are calculated within each permuted dataset. This process is repeated many times to generate a null distribution of *p*-values.

The single-step min *p* adjusted *p*-values, for a gene g, are corrected using Equation XII.

$$\tilde{p}_g = \Pr\left(\min_{1\leq l\leq m} P_l \leq p_g \middle| H_0^C\right) \quad \text{(XII)}$$

where $P_l$ refers to the unadjusted *p*-value of the $l^{th}$ hypothesis and $H_0^C$ denotes the complete null hypothesis.

Alternatively, the single-step max T adjusted *p*-value is defined as in Equation XIII.

$$\tilde{p}_g = \Pr\left(\max_{1\leq l\leq m} |T_l| \geq |t_g| \middle| H_0^C\right) \quad \text{(XIII)}$$

where $T_l$ denotes the non-normally distributed test statistic of the $l^{th}$ hypothesis (e.g., t-statistic of $l^{th}$ hypothesis with different degrees of freedom across tests). This permutation-based correction provides one of the most powerful FWER control methods, as it directly accounts for correlation between test statistics. However, due to its computational intensity, it may be impractical for large datasets or high-throughput applications.

#### *2.6.4. Benjamini-Hochberg (BH) false discovery rate (FDR) correction*

The BH FDR correction is a less stringent multiple testing correction compared to Bonferroni or Holm. Unlike those approaches which control the FWER, the BH method controls the FDR, which is the expected proportion of FP among all rejected hypotheses. This relaxation allows more discoveries (TP) to emerge.[62,63]

The BH method was introduced by Yoav Benjamini and Yosef Hochberg (1995)[64,65] and it is widely used due to its balance between sensitivity (true discovery) and error control. In comparison to other corrections, the BH procedure is less conservative, allowing more hypotheses to be rejected. It tolerates a small proportion of FP, leading to fewer FN. It also assumes independence or positive dependence among test statistics, though empirical extensions have addressed violations of these assumptions.

To apply the correction, the unadjusted *p*-values are first sorted in ascending order: $Pr_1 \leq Pr_1 \leq Pr_m$. To control the FDR at a chosen level $\alpha$, the largest index g* is determined using Equation XIV.

$$\max\{g : p_{r_g} \leq (g/m)\alpha\} \quad \text{(XIV)}$$

where $Pr_g$ is the unadjusted *p*-value, $m$ is the total number of hypotheses tested, and $\alpha$ is the chosen significance level. Then, all hypotheses $H_{r_g}$, where g = 1,…, g*, are rejected. If no such g* exists, then no hypothesis is rejected. The adjusted p values are then calculated using Equation XV.

$$\tilde{p}_{r_g} = \min_{k=g,\ldots,m}\left\{\min\left(\frac{m}{k}p_{r_k},1\right)\right\} \quad \text{(XV)}$$

For large-scale data, the FDR can also be estimated using Empirical Bayes methods.[66,67] These approaches combine frequentist inference with Bayesian shrinkage techniques and provide a multivariate estimation strategy for both effect sizes and error rates.

### 2.7. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and gene ontology (GO) analysis

In addition to *p*-value-based gene filtering used for hypothesis testing, KEGG pathway and GO filtering can be applied to identify biologically meaningful gene subsets. For instance, KEGG identifiers such as hsa0521 (bladder cancer), hsa05216 (thyroid cancer), hsa05310 (asthma), and hsa05144 (malaria) allow for pathway-level filtering, while GO identifiers like GO:0000122, GO:0005829, and GO:0005515 are used for ontology-based categorization. To perform such filtering, web-based tools like DAVID (Database for Annotation, Visualization, and Integrated Discovery) are commonly used. In DAVID, users can upload a list of gene identifiers along with the species name (e.g., Homo sapiens) to retrieve KEGG Pathway and GO terms associated with the input gene set. Internally, DAVID uses Fisher's Exact Test to determine the statistical significance of gene-term enrichment. It evaluates whether the overlap between the gene set and a specific pathway or GO term occurs more frequently than expected by chance. A *p*-value threshold (commonly $p<0.005$) is used to determine significantly enriched pathways or ontology categories.

GO is a structured, controlled vocabulary representing gene product attributes across all species. It covers three domains: (1) Biological process (BP): describes the broader biological objectives or pathways to which a gene contributes; (2) Cellular component (CC): indicates the subcellular location or structure where the gene product is active; and (3) Molecular function (MF): refers to the biochemical activity of a gene product.

This enrichment-based approach facilitates Gene Set Enrichment Analysis (GSEA) by revealing potential functional implications of gene expression changes and connecting statistically significant genes to known biological contexts.

### 2.8. Detection of miRNA target

A single miRNA can regulate multiple genes. In general, a miRNA can reduce the expression of its target genes. A miRNA is a non-coding biomolecule are crucial biomarker for various diseases, particularly in cancer diagnosis and prognosis. Various methods have been developed for miRNA target detection, such as bioinformatics tools using high-throughput sequencing data, conserved seed region matching, and hybrid deep learning-based approaches that integrate convolutional and recurrent neural networks.[68] Popular computational tools for miRNA target prediction include miRanda,[1] TargetScan, and DIANA-microT-CDS.

## 3. Comparative study of statistical tests and other computational tools in genetic feature discovery

### 3.1. Differential expression analysis

#### *3.1.1. Dataset*

In this study, we utilized microarray gene expression data related to uterine leiomyoma retrieved from the NCBI Gene Expression Omnibus (GEO) under the accession number GSE31699. The dataset comprises two kinds of samples: (i) 18 uterine leiomyoma (diseased/experimental)

samples (denoted as UL) and (ii) 18 myometrial (matched normal/control) samples (denoted as MM), all derived from African-American women.

#### 3.1.2. Comparison of different statistical tests

Firstly, we removed the genes with missing values (NA) and then low variance. Thereafter, NT using the JB test was performed to separate the dataset into normally distributed and non-normally distributed subsets. We considered only the matched pairs—16 UL and 16 MM samples—for further analysis.

Given the limitations of parametric tests on non-normally distributed data, our analysis emphasized this more challenging subset. Our analysis emphasized this more challenging subset (i.e., zero-mean normalization, min-max normalization) and thereafter applied statistical hypothesis tests (i.e., Student's two-sample t-test and Welch's two-sample t-test) without *p*-value correction as well as with *p*-value correction (using Bonferroni, BH, FDR, Holm, and Hochberg methods). Finally, we computed the differential expression analysis to determine (i) the number of differentially expressed (DE) genes, (ii) the number of up-regulated genes (UpG), and (iii) the number of down-regulated genes (DownG). We set $p<0.05$ for significance as well as FC ≥1.10 (for UpG) and FC ≤−1.10 (for DownG) as cut-offs.

The results are presented in Table 2. Overall, Student's two-sample t-test and Welch's t-test exhibited only slight differences in the number of DE, UpG, and DownG under different conditions, particularly when applied to non-normally distributed data using different normalization methods and *p*-value correction techniques.

In recent years, several advanced statistical methods and tools have been developed for differential expression analysis in microarray, RNA-Seq, and other omics datasets. Among parametric tools, Limma[14] and DESeq2[69] are widely

**Table 2. Comparative analysis of statistical hypothesis tests under varying normalization methods and *p*-value corrections using the non-normally distributed microarray dataset (GSE31699).**

| ID | Statistical hypothesis test | Normalization method | *p*-value correction | DE genes (*n*) | Up-regulated genes (*n*) | Down-regulated genes (*n*) |
|---|---|---|---|---|---|---|
| CS1 | Student's t-test | ZM | Without correction | 570 | 16 | 12 |
| CS2 | Student's t-test | ZM | Bonferroni | 3 | 1 | 0 |
| CS3 | Student's t-test | ZM | BH | 111 | 10 | 7 |
| CS4 | Student's t-test | ZM | FDR | 111 | 10 | 7 |
| CS5 | Student's t-test | ZM | Holm | 3 | 1 | 0 |
| CS6 | Student's t-test | ZM | Hochberg | 3 | 1 | 0 |
| CS7 | Student's t-test | MM | Without correction | 570 | 16 | 12 |
| CS8 | Student's t-test | MM | Bonferroni | 3 | 1 | 0 |
| CS9 | Student's t-test | MM | BH | 111 | 10 | 7 |
| CS10 | Student's t-test | MM | FDR | 111 | 10 | 7 |
| CS11 | Student's t-test | MM | Holm | 3 | 1 | 0 |
| CS12 | Student's t-test | MM | Hochberg | 3 | 1 | 0 |
| CS13 | Welch t-test | ZM | Without correction | 559 | 16 | 12 |
| CS14 | Welch t-test | ZM | Bonferroni | 1 | 0 | 0 |
| CS15 | Welch t-test | ZM | BH | 111 | 10 | 7 |
| CS16 | Welch t-test | ZM | FDR | 54 | 7 | 5 |
| CS17 | Welch t-test | ZM | Holm | 1 | 0 | 0 |
| CS18 | Welch t-test | ZM | Hochberg | 1 | 0 | 0 |
| CS19 | Welch t-test | MM | Without correction | 559 | 16 | 12 |
| CS20 | Welch t-test | MM | Bonferroni | 1 | 0 | 0 |
| CS21 | Welch t-test | MM | BH | 54 | 7 | 5 |
| CS22 | Welch t-test | MM | FDR | 54 | 7 | 5 |
| CS23 | Welc1h t-test | MM | Holm | 1 | 0 | 0 |
| CS24 | Welch t-test | MM | Hochberg | 1 | 0 | 0 |

Abbreviations: DE: Differentially expressed; BH: Benjamini-Hochberg; FDR: False discovery rate; MM: Min-max normalization; ZM: Zero-mean normalization.

adopted. Among non-parametric approaches, commonly used methods include the Mann–Whitney U test, Shrink *t*-test, and SAM.[39] For high-throughput sequencing data analysis, the Genome Analysis Toolkit developed by the Broad Institute[70] is widely used. It supports functionalities such as identifying Single-nucleotide polymorphisms, assessing copy number variations, and detecting structural variations, in addition to differential expression analysis.

Deep learning-based frameworks have also emerged. One notable example is DeepDiff,[71] which predicts differential gene expression scores from histone modification data. In parallel, recent studies have proposed improved methodologies to optimize DE analysis. Gomez *et al.*[72] demonstrated a computational drug discovery pipeline using DE signatures. Peng *et al.*[73] developed a high-performance ensemble-based inference framework for proteomics data. Aurelio *et al.*[74] proposed a DE analysis pipeline tailored to non-model species (e.g., *Cedrela odorata*).

For single-cell sequencing, dedicated tools such as DEsingle,[75] Pagoda2,[76] Seurat,[77] and Ascend[78] offer specialized functions for differential expression and methylation analyses. Several other recent well-known cancer diagnosis methods that use machine learning, deep learning, or optimization include mammography-based diagnosis,[79] integrated ultrasound and mammography approaches,[80] and skin lesion classification.[81-83]

## 3.2. Gene set enrichment study

In addition to the differential expression analysis, we also conducted a traditional GSEA study using the NCBI GEO dataset GSE30760. The analysis was performed using DAVID 2021 (December 2021) with the latest DAVID Knowledgebase v2023q4 enrichment tool.[84]

### 3.2.1. KEGG pathway analysis

Using a corrected *p*-value threshold of <0.05, we obtained 138 enriched KEGG pathways. Using FDR-corrected $p<0.05$, we identified 120 enriched KEGG pathways. A detailed summary of the top 10 KEGG pathways and the associated statistics is provided in Table 3. Additionally, the complete list of enriched KEGG pathways is available in Supplementary file[3].

### 3.2.2. GO:BP

In this analysis, we obtained 745 enriched GO:BP terms based on *p*-value correction <0.05. For FDR-corrected $p<0.05$, we identified 182 enriched GO:BP terms. The top five most significantly enriched GO:BP terms include signal transduction, positive regulation of transcription by RNA polymerase II, cell adhesion, positive regulation of DNA-templated transcription, and negative regulation of transcription by RNA polymerase II. A summary of the top 10 GO:BP terms with the associated statistics are provided in Table 4. Complete details of all enriched GO:BPs terms are provided in Supplementary File[4].

### 3.2.3. GO:CC

This analysis identified 183 enriched GO:CC terms with $p<0.05$ and 91 enriched GO:CC terms when FDR-corrected $p<0.05$ was applied. The top five enriched GO:CC terms are as follows: cytosol, plasma membrane, membrane, cytoplasm, and extracellular exosome. The top 10 GO:CC terms and the associated statistics are provided in Table 5,

[3] Data available at GSE30760_DAVID_allgenes_geneset_encriched_KEGG_path.csv

[4] Data available at GSE30760_DAVID_allgenes_geneset_enriched_GO_BP.csv

**Table 3. Top ten enriched KEGG pathways GSE30760 gene expression data using DAVID 2021 and DAVID Knowledgebase v2023q4**

| KEGG pathway ID & name | Genes (*n*) | *p* | Gene names | FDR |
|---|---|---|---|---|
| hsa05200: Pathways in cancer | 245 | 6.09E-09 | *RB1, SPI1, HHIP, KEAP1, CALML3* | 9.09E-07 |
| hsa05205: Proteoglycans in cancer | 110 | 8.15E-09 | *IHH, FZD10, FGF2, ELK1, TNF* | 9.09E-07 |
| hsa04550: Signaling pathways regulating pluripotency of stem cells | 82 | 3.52E-08 | *GSK3B, WNT2B, RIF1, ONECUT1, PIK3CD* | 2.62E-06 |
| hsa04015: Rap1 signaling pathway | 111 | 5.68E-08 | *ITGA2B, CTNND1, CALML3, CALML4, FGF2* | 3.17E-06 |
| hsa04510: Focal adhesion | 106 | 1.40E-07 | *MYLK2, ITGA2B, ELK1, ACTB, MYLK* | 6.24E-06 |
| hsa04514: Cell adhesion molecules | 84 | 1.25E-06 | *CD86, CD40, PTPRS, ITGAM, ITGB2* | 3.87E-05 |
| hsa04820: Cytoskeleton in muscle cells | 115 | 1.28E-06 | *ITGA2B, ENO3, ACTB, ACTG2, COMP* | 3.87E-05 |
| hsa04072: Phospholipase D signaling pathway | 80 | 1.39E-06 | *DGKG, DGKE, DGKD, DGKA, PIK3CD* | 3.87E-05 |
| hsa04611: Platelet activation | 69 | 3.29E-06 | *MYLK2, ITGB3, ITGA2B, PIK3CD, PIK3CB* | 8.16E-05 |
| hsa05414: Dilated cardiomyopathy | 58 | 1.60E-05 | *ITGB3, ITGA2B, TNF, ACTB, SLC8A2* | 3.42E-04 |

Abbreviation: FDR: False discovery rate.

**Table 4. Top ten GO: BP terms from GSE30760 gene expression data using DAVID 2021 and DAVID Knowledgebase v2023q4**

| GO: BP ID and name | Genes (*n*) | *p* | Gene names | FDR |
|---|---|---|---|---|
| GO: 0007165: Signal transduction | 547 | 2.27E-30 | *CNTFR, GMFB, GMFG, GLDN, CRHBP* | 2.07E-26 |
| GO: 0045944: Positive regulation of transcription by RNA polymerase II | 508 | 3.56E-29 | *ATF1, RB1, EHF, SPI1, GABPB2* | 1.62E-25 |
| GO: 0007155: Cell adhesion | 246 | 3.74E-17 | *SLC23A2, APP, SPON1, COL12A1, ICAM2* | 1.14E-13 |
| GO: 0045893: Positive regulation of DNA-templated transcription | 299 | 1.66E-16 | *TRRAP, GPATCH3, ELK1, ACTB, PSMD9* | 3.79E-13 |
| GO: 0000122: Negative regulation of transcription by RNA polymerase II | 372 | 1.72E-14 | *ZNF177, RB1, TCERG1, APP, ZNF296, SPI1* | 3.13E-11 |
| GO: 0001525:Angiogenesis | 131 | 1.70E-13 | *PLXND1, ITGA2B, SERPINE1, UBP1, RORA* | 2.58E-10 |
| GO: 0098609: Cell-cell adhesion | 102 | 1.10E-12 | *CLSTN3, CTNND2, ITGA2B, CTNND1, ICAM2* | 1.44E-09 |
| GO: 0008284: Positive regulation of cell population proliferation | 209 | 9.71E-11 | *CNTFR, VIPR1, ACTB, MYC, KDR* | 1.11E-07 |
| GO: 0007268: Chemical synaptic transmission | 110 | 1.20E-10 | *CHRM1, CHRM4, RPS6KA3, HTR6, HTR7* | 1.21E-07 |
| GO: 0048009: Insulin-like growth factor receptor signaling pathway | 51 | 1.46E-10 | *DDR1, RET, ALK, FLT1, IRS1, FLT4* | 1.33E-07 |

Abbreviations: BP: Biological process; FDR: False discovery rate; GO: Gene Ontology.

**Table 5. Top ten GO: CC terms from GSE30760 gene expression data using DAVID 2021 and DAVID Knowledgebase v2023q4**

| GO: CC ID & name | Genes (*n*) | *p* | Gene names | FDR |
|---|---|---|---|---|
| GO: 0005829: Cytosol | 1935 | 2.01E-44 | *SCOC, NUP107, TESK1, SLA2, SCP2* | 2.70E-41 |
| GO: 0005886: Plasma membrane | 1828 | 1.45E-30 | *TFRC, SLA2, HTR6, HTR7, AKT2* | 9.75E-28 |
| GO: 0016020: Membrane | 1773 | 4.68E-30 | *PGLYRP3, SPI1, NUP107, TFRC, NDST1* | 2.10E-27 |
| GO: 0005737: Cytoplasm | 1933 | 8.47E-29 | *TSKS, POP7, TESK1, SLA2, ALKBH6* | 2.85E-26 |
| GO: 0070062: Extracellular exosome | 826 | 1.46E-28 | *TFRC, ISLR, PSMD7, PSMD2, DPYSL2* | 3.92E-26 |
| GO: 0005654: Nucleoplasm | 1371 | 6.47E-26 | *ATF1, SCOC, POP7, SPI1, PWWP2B* | 1.45E-23 |
| GO: 0009986: Cell surface | 298 | 1.75E-24 | *APP, SLC46A2, SPARC, TFRC, HHIP* | 3.37E-22 |
| GO: 0009897: External side of plasma membrane | 190 | 3.62E-17 | *FCN1, CD86, CNTFR, CD84, CSF3R* | 6.07E-15 |
| GO: 0048471: Perinuclear region of cytoplasm | 303 | 2.47E-16 | *IFITM3, CYFIP1, APP, EIF4A1, TFRC* | 3.68E-14 |
| GO: 0000785: Chromatin | 422 | 3.43E-15 | *ATF1, RB1, EHF, SPI1, RAX* | 4.62E-13 |

Abbreviations: CP: Cellular process; FDR: False discovery rate; GO: Gene Ontology.

**Table 6. Top ten GO: MF terms from GSE30760 gene expression data using DAVID 2021 and DAVID Knowledgebase v2023q4**

| GO: MF ID & name | Genes (*n*) | *p* | Gene names | FDR |
|---|---|---|---|---|
| GO: 0005515: Protein binding | 4522 | 1.62E-99 | *PGLYRP3, SCOC, NUP107, TFRC, PWWP2B* | 5.07E-96 |
| GO: 0042802: Identical protein binding | 676 | 2.92E-19 | *ATF1, RB1, GABPB2, TFRC, ACCS* | 4.57E-16 |
| GO: 1990837: Sequence-specific double-stranded DNA binding | 249 | 1.17E-14 | *ZNF177, ZNF296, GFI1, FOXI1, RAX* | 1.22E-11 |
| GO: 0019904: Protein domain specific binding | 107 | 2.73E-10 | *FOXA1, APP, PLXND1, ZFYVE9, ZMYND8* | 2.13E-07 |
| GO: 0003700: DNA-binding transcription factor activity | 245 | 9.49E-10 | *ATF1, ZNF296, SPI1, GFI1, FOXI1* | 5.94E-07 |
| GO: 0005178: Integrin binding | 83 | 1.59E-09 | *APP, ITGAM, ITGB3, ITGA2B, ITGB2* | 8.31E-07 |
| GO: 0140801: Histone H2AXY142 kinase activity | 65 | 3.44E-09 | *DDR1, RET, ALK, DYRK4, ITK* | 1.35E-06 |
| GO: 0035401: Histone H3Y41 kinase activity | 65 | 3.44E-09 | *DDR1, RET, ALK, DYRK4, ITK* | 1.35E-06 |
| GO: 0005524: ATP binding | 544 | 8.52E-09 | *PI4K2B, TESK1, SMC3, SMC2, MYLK* | 2.96E-06 |
| GO: 0001228: DNA-binding transcription activator activity, RNA polymerase II-specific | 202 | 1.02E-08 | *ATF1, EHF, FOXI1, RAX, SOX21* | 2.98E-06 |

Abbreviations: FDR: False discovery rate; GO: Gene Ontology; MF: Molecular function.

and the complete list of all the enriched GO:CCs terms is listed in Supplementary File[5].

#### 3.2.4. GO:MF

In this analysis, we identified 208 enriched GO:MF terms with a $p<0.05$. For the FDR-corrected $p<0.05$, the enriched terms were reduced to 91. The top five most enriched GO:MF terms include protein binding, identical protein, sequence-specific double-stranded DNA binding, protein domain-specific binding, and DNA-binding transcription factor activity. The top 10 GO:MF terms and the associated statistics are provided in Table 6. Additionally, all the enriched GO:MFs terms are presented in Supplementary File[6].

## 4. Conclusion

In recent times, the discovery of genetic and epigenetic features—such as gene and methylation markers—has played an important role in understanding complex diseases and traits. This study provides a comprehensive review and comparative study of various well-known statistical hypothesis testing methods (i.e., Student's *t*-tests, ANOVA, Chi-square tests) in the context of genetic feature discovery and gene set enrichment analysis for microarray or RNA-seq datasets. Our analysis highlights the strengths and weaknesses of each approach, examining their methodologies, applications, performance, accuracy, and future directions. While classical statistical tests offer transparent and interpretable results, machine learning and deep learning techniques demonstrate superior capacity for managing high-dimensional data and modeling intricate biological interactions. We also explore the emerging potential of hybrid strategies that integrate statistical inference with machine or deep learning models to improve the reliability and efficiency of feature discovery. Looking ahead, promising directions include the integration of multi-omics data, the development of explainable AI models, and the advancement of scalable computational frameworks. This review serves as a resourceful guide for researchers aiming to harness the complementary strengths of statistical and machine learning methodologies in genetic and epigenetic biomarker discovery. In future work, we plan to conduct experimental evaluations using publicly available RNA-seq and Illumina DNA methylation datasets to identify robust biomarkers for various biological conditions and disease states.

## Acknowledgments

We thank all lab members and researchers from our department at Swami Vivekananda University, Kolkata, India.

## Funding

None.

## Conflict of interest

The authors declare that they have no competing interests.

## Author contributions

*Conceptualization:* Ankita Saha, Shibakali Gupta, Chyan Paul

*Visualization:* Ankita Saha, Shibakali Gupta, Chyan Paul, Saurav Mallik

*Writing–original draft:* Ankita Saha, Shibakali Gupta, Chyan Paul

*Writing–review & editing:* Shibakali Gupta, Chyan Paul, Saurav Mallik, Korhan Cengiz

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

R code and Supplementary Files are available on https://drive.google.com/drive/folders/1oQoQZFwiPB6Zpeb0s7TmYbq WKrY0z_2_?usp=drive_link.

## References

1. What is Biomedical Research? *California Biomedical Research Association*. Available from: https://statesforbiomed.org/education/background-on-biomedical-research/what-is-biomedical-research [Last accessed on 2024 Oct 09].
2. Bayat A. Clinical review science, medicine, and the future bioinformatics. *BMJ*. 2002;324:1018-1022.
   doi: 10.1136/bmj.324.7344.1018
3. Chowdhary M, Rani A, Parkash J, Shahnaz M, Dev D. Bioinformatics: An overview for cancer research. *J Drug Deliv Ther*. 2016;6(4):69-72.
   doi: 10.22270/jddt.v6i4.1290
4. Zhang S, Liu K, Liu Y, Hu X, Gu X. The role and application of bioinformatics techniques and tools in drug discovery. *Front Pharmacol*. 2025;16:1547131.
   doi: 10.3389/fphar.2025.1547131
5. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthc J*. 2021;8(2):e188-e194.
   doi: 10.7861/fhj.2021-0095
6. Khan FA, Nsengimana B, Khan NH, *et al*. Differential expression profiles of circRNAs in cancers: Future clinical

[5] Data available at GSE30760_DAVID_allgenes_geneset_enriched_GO_CC.csv

[6] Data available at GSE30760_DAVID_allgenes_geneset_enriched_GO_MF.csv

and diagnostic perspectives. *Gene Protein Dis*. 2022;1(2):138.

doi: 10.36922/gpd.v1i2.138

7. Yeh C, Madison T, Plas K. Exploring the cell-to-cell communication network to better defeat cancer. *Tumor Discov.* 2025;4(2):92.

doi: 10.36922/td.8323

8. Bandyopadhyay S, Mallik S, Mukhopadhyay A. A survey and comparative study of statistical tests for identifying differential expression from microarray data. *IEEE/ACM Trans Comput Biol Bioinform*. 2014;11(1):95-115.

doi: 10.1109/TCBB.2013.147

9. Biomolecule. *Encyclopaedia Britannica*; 2022. Available from: https://www.britannica.com/science/biomolecule [Last accessed on 2023 Mar 15].

10. Morey JS, Ryan JC, Van Dolah FM. Microarray validation: Factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biol Proced Online*. 2006;8(1):175-193.

doi: 10.1251/bpo126

11. Adler M, Alon U. Fold-change detection in biological systems. *Curr Opin Syst Biol*. 2018;8:81-89.

doi: 10.1016/j.coisb.2017.12.005

12. Vickers AJ. Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Med Res Methodol.* 2005;5:35.

doi: 10.1186/1471-2288-5-35

13. Hopkins S, Dettori JR, Chapman JR. Parametric and nonparametric tests in spine research: Why do they matter? *Global Spine J*. 2018;8(6):652-654.

doi: 10.1177/2192568218782679

14. Ritchie ME, Phipson B, Wu D, *et al*. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.

doi: 10.1093/nar/gkv007

15. Sinsomboonthong S. Performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification. *Int J Math Math Sci*. 2022;2022:3584406.

doi: 10.1155/2022/3584406

16. Henderi H, Wahyuningsih T, Rahwanto E. Comparison of min-max normalization and Z-score normalization in the K-nearest neighbor (KNN) algorithm to test the accuracy of types of breast cancer. *Int J Inform Informat Syst.* 2021;4(1):13-20.

17. Välikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform*. 2018;19(1):1-11.

doi: 10.1093/bib/bbw095

18. Li B, Tang J, Yang Q, *et al*. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. *Sci Rep.* 2016;6:38881.

doi: 10.1038/srep38881

19. Uh HW, Klaric L, Ugrina I, Lauc G, Smilde AK, Houwing-Duistermaat JJ. Choosing proper normalization is essential for discovery of sparse glycan biomarkers. *Mol Omics*. 2020;16(3):231-242.

doi: 10.1039/c9mo00174c

20. Kwak SG, Park SH. Normality test in clinical research. *J Rheum Dis*. 2019;26(1):5-11.

doi: 10.4078/jrd.2019.26.1.5

21. Khatun N. Applications of normality test in statistical analysis. *Open J Stat*. 2021;11(1):113-122.

doi: 10.4236/ojs.2021.111006

22. Das KR. A brief review of tests for normality. *Am J Theor Appl Stat*. 2016;5(1):5.

doi: 10.11648/j.ajtas.20160501.12

23. Thadewald T, Büning H. *Jarque-Bera Test and its Competitors for Testing Normality: A Power Comparison. Diskussionsbeiträge.* Freie Universität Berlin, Fachbereich Wirtschaftswissenschaft, Berlin; 2004. Available from: https://hdl.handle.net/10419/49919 [Last accessed on 2025 Apr 19].

24. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-darling tests. *J Stat Model Anal*. 2011;2:21-33.

25. Thadewald T, Büning H. *Jarque-Bera Test and its Competitors for Testing Normality: A Power Comparison. Diskussionsbeiträge*. Freie Universität Berlin, Fachbereich Wirtschaftswissenschaft, Berlin; 2004.

26. Livingston EH. The mean and standard deviation: What does it all mean? *J Surg Res*. 2004;119(2):117-123.

doi: 10.1016/j.jss.2004.02.008

27. Ugoni A, Walker BF. The chi square test: An introduction. Aust Chiropr Osteopathy. 1995;4(3):85-91.

28. McHugh ML. The Chi-square test of independence. *Biochem Med (Zagreb)*. 2013;23(2):143-149.

doi: 10.11613/BM.2013.018

29. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*. 2009;25(6):765-771.

doi: 10.1093/bioinformatics/btp053

30. Thanavathi C. *Advanced Educational Research and Statistics*; 2017. Available from: https://www.researchgate.net/publication/337991541 [Last accessed on 2025 Apr 19].

31. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis

of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*. 2019;20:40.

doi: 10.1186/s12859-019-2599-6

32. Boareto M, Caticha N. t-Test at the probe level: An alternative method to identify statistically significant genes for microarray data. *Microarrays*. 2014;3(4):340-351.

doi: 10.3390/microarrays3040340

33. Zhang L, Zhu T, Zhang JT. Two-sample Behrens-Fisher problems for high-dimensional data: A normal reference scale-invariant test. *J Appl Stat*. 2023;50(3):456-476.

doi: 10.1080/02664763.2020.1834516

34. Hong S, Coelho CA, Park J. An exact and near-exact distribution approach to the Behrens-fisher problem. *Mathematics*. 2022;10(16):2953.

doi: 10.3390/math10162953

35. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol.* 2014;14:135.

doi: 10.1186/1471-2288-14-135

36. Dao PB. On Wilcoxon rank sum test for condition monitoring and fault detection of wind turbines. *Appl Energy*. 2022;318:119209.

doi: 10.1016/j.apenergy.2022.119209

37. Botlagunta M, Khatri K, Devi BM, Doneti R, Pasha A, Pawar SC. Differential expression of DDX3 and microRNAs in response to hormone and cisplatin against cervical cancer. EJMO. 2022;6(4):307-316.

doi: 10.14744/ejmo.2023.96531

38. Larsson O, Wahlestedt C, Timmons JA. Considerations when using the significance analysis of microarrays (SAM) algorithm. *BMC Bioinformatics*. 2005;6:129.

doi: 10.1186/1471-2105-6-129

39. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98(9):5116-5121.

doi: 10.1073/pnas.091062498

40. Bewick V, Cheek L, Ball J. Statistics review 10: Further nonparametric methods. *Crit Care*. 2004;8(3):196-199.

doi: 10.1186/cc2857

41. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc.* 1952;47(260):583-621.

doi: 10.2307/2280779

42. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*. 2002;18(11):1454-1461.

doi: 10.1093/bioinformatics/18.11.1454

43. Massey FJ Jr. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc*. 1951;46(253):68-78.

doi: 10.2307/2280095

44. Steinskog DJ, Tjøtheim DB, Kvamstø NG. A cautionary note on the use of the Kolmogorov-Smirnov test for normality. *Mon Weather Rev*. 2007;135(3):1151-1157.

doi: 10.1175/MWR3326.1

45. Pushap AC, Sudershan S, Sudershan A. Type of error in statistics: A review. *Haya Saudi J Life Sci*. 2023;8(03):39-43.

doi: 10.36348/sjls.2023.v08i03.001

46. Kaur P, Stoltzfus J. Type I, II, and III statistical errors: A brief overview. *Int J Acad Med.* 2017;3(2):268-270.

doi: 10.4103/IJAM.IJAM_92_17

47. Shaffer JP. *Multiple Hypothesis Testing: A Review. Technical Report No. 23.* Research Triangle Park, NC: National Institute of Statistical Sciences; 1994. Available from: https://www.niss.org [Last accessed on 2025 Apr 19].

48. El-Gohary TM. Hypothesis testing, type I and type II errors: Expert discussion with didactic clinical scenarios. *Int J Health Rehabil Sci*. 2019;8(3):132.

doi: 10.5455/ijhrs.0000000180

49. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1):6.

doi: 10.1186/s12864-019-6413-7

50. Lise S, Archambeau C, Pontil M, Jones DT. Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinformatics*. 2009;10:365.

doi: 10.1186/1471-2105-10-365

51. Gohary T. Hypothesis testing, type I and type II errors: Expert discussion with didactic clinical scenarios. *Int J Health Rehabil Sci*. 2019;8(3):132.

doi: 10.5455/ijhrs.0000000180

52. Jafari M, Ansari-Pour N. Why, when and how to adjust your P values? *Cell J*. 2019;20(4):604-607.

doi: 10.22074/cellj.2019.5992

53. Greenland S, Senn SJ, Rothman KJ, *et al.* Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337-350.

doi: 10.1007/s10654-016-0149-3

54. Silicon Genetics. *Multiple Testing Corrections.* Redwood City, CA: Silicon Genetics; 2003.

55. Vasilopoulos T, Morey TE, Dhatariya K, Rice MJ. Limitations of significance testing in clinical research:

A review of multiple comparison corrections and effect size calculations with correlated measures. *Anesth Analg*. 2016;122(3):825-830.

doi: 10.1213/ANE.0000000000001107

56. Sedgwick P. Multiple significance tests: The Bonferroni correction. *BMJ*. 2012;344:e509.

doi: 10.1136/bmj.e509

57. Vickerstaff V, Omar RZ, Ambler G. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. *BMC Med Res Methodol*. 2019;19(1):129.

doi: 10.1186/s12874-019-0754-4

58. Blakesley RE, Mazumdar S, Dew MA, *et al*. Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology*. 2009;23(2):255-264.

doi: 10.1037/a0012850

59. Kang G, Ye K, Liu N, Allison DB, Gao G. Weighted multiple hypothesis testing procedures. *Stat Appl Genet Mol Biol*. 2009;8(1):23.

doi: 10.2202/1544-6115.1437

60. Cox DD, Lee JS. Pointwise testing with functional data using the Westfall-Young randomization method. *Biometrika*. 2008;95(3):621-634.

doi: 10.1093/biomet/asn021

61. Westfall PH, Young SS. p Value adjustments for multiple tests in multivariate binomial models. *J Am Stat Assoc*. 1989;84(407):780-786.

doi: 10.1080/01621459.1989.10478837

62. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*. 2005;21(13):3017-3024.

doi: 10.1093/bioinformatics/bti448

63. Acharya A. *A Complete Review of Controlling the False Discovery Rate in a multiple Comparison Problem Framework: The Benjamini-Hochberg Algorithm*. *arXiv:1406.7117v1 [stat.ME]*; 2014.

doi: 10.48550/arXiv.1406.7117

64. Benjamini Y. Discovering the false discovery rate. *J R Stat Soc Series B Stat Methodol*. 2010;72(4):405-416.

doi: 10.1111/j.1467-9868.2010.00746.x

65. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4):1165-1188.

doi: 10.1214/aos/1013699998

66. Chakraborty A, Jiang G, Boustani M, Liu Y, Skaar T, Li L. Simultaneous inferences based on empirical Bayes methods and false discovery rates in eQTL data analysis. *BMC Genomics*. 2013;14(Suppl 8):S8.

doi: 10.1186/1471-2164-14-S8-S8

67. Efron B. Microarrays, empirical Bayes and the two-groups model. *Stat Sci*. 2008;23(1):1-22.

doi: 10.1214/07-STS236

68. Gu T, Zhao X, Barbazuk WB, Lee JH. miTAR: A hybrid deep learning-based approach for predicting miRNA targets. *BMC Bioinformatics*. 2021;22(1):96.

doi: 10.1186/s12859-021-04026-6

69. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.

doi: 10.1186/s13059-014-0550-8

70. Available from: https://gatk.broadinstitute.org/hc [Last accessed 2025 Jul 03].

71. Sekhon A, Singh R, Qi Y. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics*. 2018;34(17):i891-i900.

doi: 10.1093/bioinformatics/bty612

72. Gomez CG, Rosa-Calatrava M, Fouret J. Optimizing *in silico* drug discovery: Simulation of connected differential expression signatures and applications to benchmarking. *Brief Bioinform*. 2024;25(4):bbae299.

doi: 10.1093/bib/bbae299

73. Peng H, Wang H, Kong W, *et al*. Optimizing differential expression analysis for proteomics data via high-performing rules and ensemble inference. *Nat Commun*. 2024;15:3922.

doi: 10.1038/s41467-024-47899-w

74. Aurelio AMM, Fabián CAF, Iván CCC, Felipe GL. Optimized method for differential gene expression analysis in non-model species: Case of *Cedrela odorata* L. *MethodsX*. 2023;11:102449.

doi: 10.1016/j.mex.2023.102449

75. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*. 2018;34:3223-3224.

doi: 10.1093/bioinformatics/bty332

76. Available from: https://github.com/kharchenkolab/pagoda2 [Last accessed on 2025 Jul 15].

77. Hao Y, Stuart T, Kowalski MH, *et al*. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol*. 2024;42(2):293-304.

doi: 10.1038/s41587-023-01767-y

78. Senabouth A, Lukowski SW, Hernandez JA, *et al*. ascend: R package for analysis of single-cell RNA-seq data. *Gigascience*. 2019;8(8):giz087.

doi: 10.1093/gigascience/giz087

79. Hussain SI, Toscano E. Optimized deep learning for mammography: Augmentation and tailored architectures. *Information*. 2025;16(5):359.

doi: 10.3390/info16050359

80. Xu Z, Zhong S, Gao Y, *et al.* Optimizing breast lesions diagnosis and decision-making with a deep learning fusion model integrating ultrasound and mammography: A dual-center retrospective study. *Breast Cancer Res*. 2025;27:80.

doi: 10.1186/s13058-025-02033-6

81. Shetty B, Fernandes R, Rodrigues AP, *et al.* Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Sci Rep.* 2022;12:18134.

doi: 10.1038/s41598-022-22644-9

82. Hussain SI, Toscano E. An extensive investigation into the use of machine learning tools and deep neural networks for the recognition of skin cancer: Challenges, future directions, and a comprehensive review. *Symmetry*. 2024;16(3):366.

doi: 10.3390/sym16030366

83. Hussain SI, Toscano E. Enhancing recognition and categorization of skin lesions with tailored deep convolutional networks and robust data augmentation techniques. *Mathematics*. 2025;13(9):1480.

doi: 10.3390/math13091480

84. Available from: https://davidbioinformatics.nih.gov/home.jsp [Last accessed on 2025 Jul 02].

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

PERSPECTIVE ARTICLE

# Healthcare leadership in the modern age of artificial intelligence: Are we organizationally ready?

**Justin Iannello***

Veterans Health Administration, VISN 21 (Sierra Pacific Network), Pleasant Hill, California, United States of America

## Abstract

Organizational artificial intelligence (AI) readiness in healthcare has gained significant attention, given the excitement of transforming healthcare in new and innovative ways. While many healthcare organizational leaders have expressed a strong desire to leverage AI to automate processes, improve productivity, and increase staff/patient satisfaction, most are still in the beginning stages of identifying, strategizing, and implementing foundational elements that transform an organization from one that utilizes AI to one that embraces AI as a key partner in delivering healthcare. Given the complex nature of integrating technology tools/solutions like AI within clinical and operational workflows, the paper will highlight widespread interest for AI integration among healthcare leaders; introduce individual and organizational factors that affect adoption of AI tools/technologies; and provide an overview of organizational AI readiness examples to assist healthcare organizations on their AI journey.

***Keywords:*** Organization; Artificial intelligence; Readiness; Adoption factors; Technology; Healthcare; Transformation

***Corresponding author:**
Justin Iannello
(JLIannello22@gmail.com)

**Citation:** Iannello J. Healthcare leadership in the modern age of artificial intelligence: Are we organizationally ready? *Artif Intell Health*. 2026;3(1):71-76. doi: 10.36922/AIH025230051

**Received:** June 5, 2025

**1st revised:** June 27, 2025

**2nd revised:** July 2, 2025

**Accepted:** July 2, 2025

**Published online:** July 18, 2025



## 1. Introduction

Artificial intelligence (AI) in healthcare is often referenced given its promise to improve productivity, increase operational efficiency, and solve some of healthcare's most complex challenges. There is little doubt that the present and future of healthcare will advance in partnership with machine learning, neural networks, deep learning, and various automation tools.

According to the American Hospital Association Center for Health Innovation, approximately 43% of healthcare executives are prioritizing business process automation, and over 90% of healthcare executives reported that hiring and training AI talent is a vital part of their organizational plan.[1]

Yet, despite enthusiasm amongst healthcare system executives regarding AI's potential impact in healthcare, just 6% have organizational strategies for AI integration.[2] Furthermore, the Scottsdale Institute (2024) reported that >70% of healthcare executives believe that demonstrating an informatics return on investment is a "moderate" or "severe" problem organizationally, whereas just 22% report that informatics in their organization is "leading the way."[3]

As technological innovation continues to rapidly permeate the healthcare space, identifying and strategizing how critical organizational and individual AI readiness factors will be approached systematically is key to value-added AI transformation. Unsurprisingly, the leap for healthcare organizations to a successful AI-integrated system is rooted in the very fundamentals that shape an institution's structure and function.

## 2. Organizational AI readiness and the fundamentals

### 2.1. Organizational AI readiness models

According to Sriharan *et al.*,[4] AI transformation in healthcare systems requires that organizations successfully execute AI integration across four functional areas: Technological (e.g., AI technical skills, innovation, and interdisciplinary subject matter expertise); strategic (e.g., organizational alignment, workflow integration, effective communication, and change management); operational (e.g., governance, ethics and risk management, data privacy and security, and an understanding of regulatory factors); and organizational (e.g., culture, building trust, stakeholder engagement, incentivizing staff, and a collaborative work environment).[4] Functional domains are further supported by organizational healthcare leaders that display technical capacity such as AI literacy, subject matter expertise, change management, and innovative thinking; adaptive capacity including systems thinking, the ability to identify opportunistic emerging technologies, agile thinking, and transformational mindset; and interpersonal capacity such as teambuilding skills, the ability to develop partnerships across numerous organizational areas, understanding various stakeholder perspectives, integrity, a sense of humility, and the capacity to influence adoption.[4]

Finally, additional vital areas include contextual factors that affect organizational AI readiness, such as organizational (e.g., transparency and innovation spirit), technological (e.g., subject matter expertise, technical talent, and resource management), and regulatory impact.[4]

In addition, other organizational AI readiness models from the literature share similar ideas such as the importance of organizational culture, strategy, healthcare leadership support and communication, stakeholder engagement, digital literacy, governance, technological alignment, workflow integration, ethical decision-making, and regulatory factors.[5-12]

### 2.2. Practical AI healthcare applications: A review of organizations demonstrating readiness

While reaching the pinnacle of organizational AI readiness in the healthcare industry is still a work in progress, several healthcare institutions are spearheading AI readiness in unique and exciting ways. Included below are three examples of healthcare systems that demonstrate their journey to organizational AI readiness in functional areas (technological, strategic, operational, organizational); healthcare leadership capacity (technical capacity, adaptive capacity, interpersonal capacity); and/or contextual domains (organizational, technological, regulatory).[4]

#### *2.2.1. Case A: Stanford university[13]*

Stanford University implemented an AI playground that provides an opportunity for faculty, students, and additional staff to test-drive large language models and other AI solutions on their custom-developed platform. The "Stanford AI Playground" is a novel concept that demonstrates organizational AI readiness from a functional standpoint (e.g., technological focus, such as development and access to tools to improve AI literacy, and organizational alignment through stakeholder engagement and establishing a culture that supports AI innovation).[4]

#### *2.2.2. Case B: Duke Health[14]*

Duke Health developed a governance framework for the management of AI tool deployment to ensure safe and effective integration of AI organizationally. Their framework includes quality assurance (known as "checkpoint gates") for the entire AI lifecycle from creation of models to evaluation phases to deployment and continuous monitoring.

Duke Health's structure highlights functional domains such as operational (robust governance, inclusive of factors such as data-driven decision-making, risk management, and regulatory compliance to ensure the highest quality) as well as strategic (given their emphasis on breaking down siloes, aligning technology with workflows, and rigorous validation). Furthermore, contextual factors are apparent such as organizational (demonstrated by organizational-wide transparency, consistency, standardization, and knowledge sharing with staff) and technological (e.g., a commitment to resourcing governance and the involvement of interdisciplinary subject matter experts such as data scientists, biostatisticians, informatics, policy and legal personnel, implementation specialists, and clinical staff that provide guidance on interpretation and appropriate application of AI model output during clinical decision-making).[4]

#### *2.2.3. Case C: The Permanente Medical Group (TPMG)[15]*

TPMG, one of the largest integrated health systems in the United States, was an early pioneer implementing ambient dictation in the clinical setting. While ambient dictation (a natural language processing tool that captures

audio dialogue between providers and patients via the use of a smart phone) has been commonly utilized across numerous healthcare systems, TPMG's approach to organizational readiness stands out. During deployment, TPMG included a variety of strategies for approximately 10,000 providers such as training sessions, peer-to-peer support, patient resources, and sustainment plans demonstrating measurable outcomes such as increased quality of documentation, decreased provider workload burden, and patient satisfaction.

TPMG's focus on functional areas include strategic (e.g., effective workflow integration, change management planning, and provider, staff, and patient support before and during AI tool deployment); operational (e.g., data privacy during provider-patient encounters, regulatory requirements involving clinical documentation, and risk mitigation preparedness involving use of the technology); and organizational (such as building trust across the organization and encouraging collaboration and a strong teamwork culture). In addition, healthcare leadership capacity areas are emphasized including that of adaptive capacity (e.g., transformational thinking and taking advantage of emerging technologies to benefit patients, providers, and staff); interpersonal capacity (such as identifying clinical champions, engaging staff, and facilitating AI adoption); and technical capacity (e.g., developing provider training plans to grow AI literacy and utilize this technology in meaningful and effective ways).

## 3. AI readiness adoption factors: Planning your work and working your plan

Healthcare management is complex and multifaceted. Any AI or technological alignment with the organization's mission and vision is challenging; however, achieving this goal and realizing return on investment is possible if you plan for it. In fact, the Scottsdale Institute (2024) reported that over 20% of healthcare organizations believe their informatics program is "leading the way" from analytics and quality to improved operational processes, outcomes, and patient experience.[3] Healthcare institutions that are committed to an organizational AI transformation involving people, process, and technology have demonstrated correlation with organizational performance.[6]

### 3.1. "Planning your work" (individual and organizational AI adoption factors)[1-12]

Healthcare system AI readiness can be affected by numerous individual and organizational adoption factors. Figure 1 outlines essential questions healthcare organizations may consider exploring during their AI journey.

Individual AI adoption factors
- Clinician attitudes: What is the digital literacy amongst interdisciplinary team members? Are front-line staff aware of the future direction of the organization and how their skillset in alignment with AI supports (not substitutes for) their workflows?
- Clinical context: Has clinical context, risk, and level of human involvement been discussed?
- AI model design: Has there been consultation with various clinical and technical subject matter experts on data sources and types/subtypes of AI model designs to utilize for your area of study?
- Workflow integration: Have baseline workflow assessments been completed? Have specific opportunities been identified to identify scope and focus of patient or organizational impact?
- Cognitive biases: Have cognitive biases been articulated? Is there a plan to address them?
- Guidelines: Are AI point of contact (s) and AI use case process owners aware of current organization or industry-wide AI guidelines?
- Liability: Are clinicians aware that clarity regarding liability is unknown?

Organizational AI adoption factors
- Organizational culture: Are AI and other technology applications aligned with the organizational mission and strategy? Does the organization have a "build" or "buy" culture?
- Organizational strategy: Is the organizational mission and strategy aligned with AI innovation?
- Organizational priorities: Is the project of focus considered a high organizational priority? Is the timing right? Are high-priority organizational initiatives seeking AI solutions as opportunities to improve (not fix) operational or clinical areas of care?
- Organizational resources and governance: Has the organization committed to resources and established governance (including ethics) for AI oversight, AI lifecycle strategic planning, and AI innovation?
- Organizational teams and structure: Does the organization have the necessary structure in place to manage AI projects? Are the appropriate clinical, operational, AI, and other technical subject matter experts in place?
- Organizational change management: Does the organization invest in change management resources? Are there organizational training and change management plans? Opportunities for local involvement with validation?
- Organizational training: Has the organization committed to hiring/developing AI talent and investing in AI growth? Does the organization have a plan to increase digital literacy across the healthcare system?
- Local validation: Is there a trusted local clinical and AI subject matter expert who can assess risk and provide input on local AI model training, testing, validation, and deployment?
- Systems impact: Has the totality of AI model design, training, and validation been considered prior to deployment such as transparency, workflow impact, limiting bias, fair and equitable systems, etc.?
- Regulation and standards: Is the project lead a subject matter expert, or does the project team have an AI regulation and standards expert?
- Evaluation and validation: Is there a clinical and AI subject matter expert who can provide input on local AI model design or engage with vendors regarding intricate details pertaining to explainability, outcomes, and risk?

**Figure 1.** Outline of key individual and organizational artificial intelligence adoption factors

**Table 1. Transforming organizational AI thinking–healthcare leaders, clinical staff, and additional healthcare professionals**

| For healthcare and administrative leaders | For healthcare clinical staff | For any healthcare professional |
|---|---|---|
| • AI healthcare system transformation is hard and alignment with the organizational mission and vision are critical to realize return on investment utilizing AI tools. | • AI is not a substitute for clinical decision-making. Responsible, safe, and trustworthy AI in healthcare should be seen as augmented intelligence (humans and AI working together to improve decision-making and new discovery). | • Start with the basics and focus on the fundamentals which are vital to the success of any AI, technology, or process improvement project. |
| • Neither AI nor any technology will fix organizational processes, operations, or culture. | • AI model bias may be reflective of human bias. | • Technology is not the answer if there are people and process gaps. |
| • An executive leadership team member champion or executive sponsor for AI-based projects/initiatives and governance demonstrates support and engagement to organizational staff. | • A clinical expert does not automatically become an AI expert, just like an AI subject matter expert does not automatically become a clinical expert. | Educate yourself first before educating others.[a] |
| • Leadership engagement, commitment, and resource support for multi-disciplinary teams are some of the most important factors involved with AI and technology adoption. | • Respective clinical subject matter experts with domain expertise and workflow know-how should always be involved in providing AI model training and conducting local validation. | • Ensure your project aligns with the organizational mission and/or priorities. |
| • Organizational focus on recruitment and/or developing people in the AI industry is one of the most essential and value-added investments one can make (upskilling/reskilling is a strategic resource). | • Interpreting AI studies and AI model performance outcomes differs from historical clinician training involving statistical analysis. | • Building multi-disciplinary teams with respective clinical and technical is a good start; identifying like-minded people and champions is even better. |
| • Organizational structure and governance involving AI technology standards, guidelines, innovation, design, implementation, adoption, and fine-tuning/maintenance are vital. | • AI regulations and standards are rapidly evolving, and clarity regarding liability remains an unanswered question. | • AI can identify correlation but cannot validate causation. |

Notes: [a]Per the National Health Service AI Lab and Health Education England, a suggested education and training approach may consider three steps in the following.

### 3.2. "Working your plan" (transforming organizational AI thinking – healthcare leaders, clinical staff, and additional healthcare professionals)[1-12]

In addition to examining vital individual and organizational AI adoption factors, it is also important for healthcare systems to identify and bridge organizational gaps throughout all levels of the organization (Table 1).

The National Health Service AI Lab and Health Education England suggests the following AI education and training approach:

(1) Educational groundwork: Increase workforce AI awareness and facilitate the "adoption of change and innovation in healthcare settings." For example, what AI is and what it isn't (AI is not a substitute for clinical decision-making).

(2) Foundational and advanced AI education:
- Foundational: Improve general AI literacy including understanding "limitations and risks of using AI technologies." For instance, staff should be aware of a general sense that AI models usually suggest correlation, not causation.
- Advanced: Developing staff with more in-depth AI skills/knowledge in alignment with their domain expertise. For example, a board certified clinical informaticist that combines AI knowledge with clinical subject matter input to deploy AI tools.

(3) Product-specific training: A deeper understanding of specific AI technologies integrated into workflow settings. For instance, a trained, tested, and validated machine learning model(s) application being utilized in patient flow operations.[10]

## 4. Conclusion

Organizational areas such as strategic insight, technological alignment, operational planning, workflow integration, organizational culture, leadership support and communication, stakeholder engagement, governance oversight, increasing organizational-wide digital literacy, and resource commitment to invest in reskilling/upskilling are germane to realizing AI return on investment.

While most healthcare organizations are still in the beginning stages of developing organizational plans to effectively utilize AI technology, several healthcare institutions have demonstrated return on value by strategically executing key organizational AI readiness

factors illustrated in the literature. Educating healthcare institutions in the early stages of their AI journey about organizational AI readiness frameworks may potentially help healthcare systems navigate the complexities of integrating AI solutions productively and efficiently.

## Acknowledgments

None.

## Funding

None.

## Conflict of interest

The author declares no conflicts of interest.

## Author contributions

This is a single-authored article.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

Not applicable.

## Further disclosure

The opinions expressed herein are those of the author and do not necessarily reflect those of the US Government or any of its agencies. The author is writing book chapters for AI in primary care in collaboration with the American Board of AI in Medicine.

## References

1. American Hospital Association Center for Health Innovation. *AI and the Health Care Workforce: How Hospitals and Health Systems can Use Artificial Intelligence to Build the Health Care Workforce of the Future*. AHA Center for Health Innovation. Available from: https://www.aha.org/system/files/media/file/2019/09/Market_Insights_AI_Workforce_2.pdf [Last accessed on 2025 Jul 17].
2. Berger E, Dries M. *Beyond Hype: Getting the Most Out of Generative AI in Healthcare Today. Bain and Company*; 2023. Available from: https://www.bain.com/insights/getting-the-most-out-of-generative-ai-in-healthcare [Last accessed on 2025 Jul 17].
3. Scottsdale Institute Research Report. *Illuminating Informatics: Exploring Health System Programs' Optimal Future State*. Kirby Partners; 2024. Available from: https://www.kirbypartners.com/wp-content/uploads/2024/05/SI_Kirby-Clinical-Informatics-Report-6_9_24.pdf [Last accessed on 2025 Jul 17].
4. Sriharan A, Sekercioglu N, Mitchell C, *et al*. Leadership for AI transformation in health care organization: Scoping Review. *J Med Internet Res*. 2024;26:e54556.

   doi: 10.2196/54556
5. Roppelt J, Kanbach DK, Kraus S. Artificial intelligence in healthcare institutions: A systematic literature review on influencing factors. *Technol Soc*. 2024;76:102443.

   doi: 10.1016/j.techsoc.2023.102443
6. Randriamiary D. Reframing the Role of leaders navigating the challenges and opportunities of tomorrow's workplace in the age of artificial intelligence. *Int J Innov Sci Res Technol*. 2024;9(1).
7. Felemban H, Sohail M, Kuikar K. Exploring the readiness of organisations to adopt artificial intelligence. *Buildings*. 2024;14(8):2460.

   doi: 10.3390/buildings14082460
8. Frehywot S, Vovides, Y. Contextualizing algorithmic literacy framework for global health workforce education. *AIH*. 2024;2(2):41-6.

   doi: 10.36922/aih.4903
9. Nix M, Onisiforou G, Painter A. *Understanding Healthcare Workers' Confidence in AI: Report 1 of 2. NHS AI Lab and Health Education England*; 2022. Available from: https://digital-transformation.hee.nhs.uk/binaries/content/assets/digital-transformation/dart-ed/understandingconfidenceinai-may22.pdf [Last accessed on 2025 Jul 17].
10. Nix M, Onisiforou G, Painter A. *Developing Healthcare Workers' Confidence in AI*. NHS AI Lab and Health Education England; 2022. Available from: https://digital-transformation.hee.nhs.uk/binaries/content/assets/digital-transformation/dart-ed/developingconfidenceinai-oct2022.pdf [Last accessed on 2025 Jul 17].
11. Madanchian M, Taherdoost H, Vincenti M, Mohamed N. Transforming leadership practices through artificial intelligence. *Proc Comput Sci*. 2024;235(3):2101-2111.

    doi: 10.1016/j.procs.2024.04.199
12. Johnk J, Weibert M, Wyrtki K. Ready or not, AI comes- an interview study of organizational AI readiness factors. *Bus Inf Syst Eng*. 2021;63(1):5-20.

    doi: 10.1007/s12599-020-00676-7
13. Stanford Report. *Stanford's AI Playground Offers a Safe Place to Explore and Experiment*. Stanford University; 2025. Available from: https://news.stanford.edu/stories/2025/01/ai-playground-offers-a-safe-place-to-explore-and-

experiment [Last accessed on 2025 Jul 17].

14. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, *et al*. A framework for the oversight and local deployment of safe and high-quality prediction models. *J Am Med Inform Assoc.* 2022;29(9):1631-1636.
doi: 10.1093/jamia/ocac078

15. Tierney AA, Gayre G, Hoberman B, *et al*. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catal.* 2024;5(3).
doi: 10.1056/CAT.23.0404

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

# Pediatric patient hospital length of stay prediction: A comparative analysis of Bayesian inference and machine learning approaches

**Sarmad Zafar[1]*, Tariq Mahmood[1], Zahra Hoodbhoy[2], and Babar Hasan[3]**

[1]Big Data Analytics Laboratory, Department of Computer Science, School of Mathematics and Computer Science, Institute of Business Administration Karachi, Karachi, Sindh, Pakistan

[2]Department of Paediatrics and Child Health, Medical College, Aga Khan University, Karachi, Sindh, Pakistan

[3]Division of Cardiothoracic Sciences, Division of Cardio-thoracic Sciences, Sindh Institute of Urology and Transplantation, Karachi, Sindh, Pakistan

## Abstract

Predicting patient length of stay (LoS) is crucial for optimizing resource allocation and enhancing healthcare efficiency. However, achieving accurate LoS predictions remains a challenging and complex task. This study presents a non-disease-specific predictive model that integrates machine learning (ML) methods and Bayesian inference techniques to accurately predict hospital LoS using static patient admission data. While traditional statistical regression techniques have been widely used for LoS prediction within hospital settings, this research investigates the capabilities of ML and Bayesian inference algorithms in this context. By leveraging Bayesian inference techniques, our model captures complex relationships within the data and quantifies uncertainty, offering a more nuanced understanding of the outcomes. This methodological approach offers a more comprehensive and probabilistically grounded framework for LoS prediction, allowing more informed decision-making in resource allocation and patient management. Among the evaluated models, extreme boosting and support vector machine regressor models demonstrated the highest efficiency, achieving mean squared logarithmic error (MSLE) values of 0.23 and 0.24, respectively. The Bayesian model also showed competitive performance with an MSLE of 0.25. While it did not outperform other models in terms of error metrics, the Bayesian model's ability to provide additional uncertainty output enhances its utility, offering valuable supplementary information for informed decision-making. This research highlights the potential of ML and Bayesian inference in predicting patient LoS, emphasizing their significance in effective resource allocation and patient care management within the healthcare sector.

***Keywords:*** Length of stay; Machine learning; Predictive model; Bayesian inference; Natural language processing

***Corresponding author:**
Sarmad Zafar
(s.zafar@khi.iba.edu.pk)

**Citation:** Zafar S, Mahmood T, Hoodbhoy Z, Hasan B. Pediatric patient hospital length of stay prediction: A comparative analysis of Bayesian inference and machine learning approaches. *Artif Intell Health*. 2026;3(1):77-87. doi: 10.36922/AIH025160030

**Received:** April 14, 2025

**Revised:** June 26, 2025

**Accepted:** July 7, 2025

**Published online:** July 22, 2025



## 1. Introduction

Predictive modeling has become increasingly prevalent in various domains for forecasting future outcomes and facilitating informed decision-making. In healthcare,

predictive modeling, especially modern machine learning (ML) methods, is widely used to forecast individual patient outcomes using large patient datasets.[1] These techniques have been used to predict numerous critical features such as mortality, readmission probabilities, recommended treatments, and length of stay (LoS).[2-4] Patient LoS is an important metric in healthcare operations, contributing significantly to the efficient flow of patients through hospital systems. Its strategic importance lies in its role in optimizing resource allocation, bed management, specialist scheduling, billing estimation, discharge planning, and overall enhancement of patient satisfaction and operational efficacy.[5,6]

Accurate LoS prediction is an effective tool for addressing challenges in resource planning, capacity management, and staffing. Reliable forecasts of patient discharge dates enable better scheduling of elective admissions and improve hospital bed occupancy management. In addition, accurate LoS predictions can enhance hospital workflow, improve patients' safety, reduce healthcare costs, and optimize resource utilization.[7] LoS is often used as a proxy indicator when direct measurement of certain outcomes is not feasible; for example, it can be used as a proxy for hospital mortality.[8] Moreover, LoS is used to assess illness severity and estimate healthcare resource utilization.[9]

LoS is a complex metric influenced by multiple factors, such as the type of illness, geographic location and season, and individual characteristics including demographics, age, medical complications, and treatment complexity.[10] Conventionally, statistical techniques have dominated LoS predictive modeling.[11,12] These algorithms assume specific relationships between variables and outcomes and often treat variables as independent. Given the complexity of LoS determinants, these assumptions may not hold, leading to limited predictive performance.[13] Hence, there is a growing emphasis on leveraging ML algorithms to improve prediction accuracy. Figure 1 highlights the methods used in the literature for calculating and predicting LoS.[14]

ML algorithms can process and integrate numerous features, capturing complex and non-linear relationships between them to make accurate predictions. Various ML techniques have been used for LoS prediction, including linear regression, multilayer perceptron, random forest, bagging, boosting, and support vector machine.[15-18] While these methods vary in performance, depending on the specific context and data, they offer a flexible framework that often outperforms traditional approaches in capturing the underlying patterns associated with hospital LoS.

Despite their successes, classical ML approaches sometimes fail to deliver satisfactory results or may be unsuitable due to specific circumstances, available data, desired outcomes, and personal preferences. Consequently, Bayesian inference is used as an alternative to frequentist approaches because it allows the incorporation of prior knowledge during model training, helping to mitigate limitations posed by small or imperfect datasets.[19] In addition, Bayesian methods provide insights into the model's confidence in each prediction by quantifying uncertainty – distinct from the confidence intervals in frequentist methods – and by detecting areas with insufficient training data.[20]

Bayesian methods, unlike classical ML methods, leverage probability distributions for model parameters, enabling uncertainty estimation. This capability allows Bayesian models to quantify uncertainties associated with predicted outputs, offering a more nuanced understanding of the reliability of predictions.[21] In healthcare, if Bayesian predictive models demonstrate comparable performance to frequentist models, they are often preferred due to their ability to provide uncertainty estimates.[22]

The existing body of literature predominantly relies on conventional regression methods or frequently incorporates ML approaches such as random forest and support vector machines for LoS prediction. In this research, we aim to bridge this gap by introducing Bayesian inference-based techniques. The primary objective is to enhance predictive accuracy through improved point estimates while simultaneously strengthening the reliability of uncertainty quantification. In addition, this study compares the performance of Bayesian and ML approaches to evaluate their relative effectiveness in LoS prediction.

## 2. Data and methods

### 2.1. Study setting and data variables

This study utilized data sourced from Aga Khan University Hospital, Karachi, a distinguished not-for-profit tertiary healthcare facility. Aga Khan University Hospital serves as a leading hospital in Pakistan, catering to a highly diverse patient population presenting with a wide range of health conditions from across the country and abroad. The study focused on pediatric patients aged less than 18 years admitted as inpatients between January 01, 2015, and November 30, 2019. Records pertaining to inpatients admitted for elective procedures, surgeries, or diagnostic examinations were omitted from the analysis. The dataset consisted of records from approximately 22,106 pediatric patients, encompassing a range of attributes including demographic and health-related parameters, clinical admission information, such as the initial level of care classification (general, intensive, special, or isolation care), and other relevant administrative data. Socio-demographic attributes include age, gender, address, and financial class.

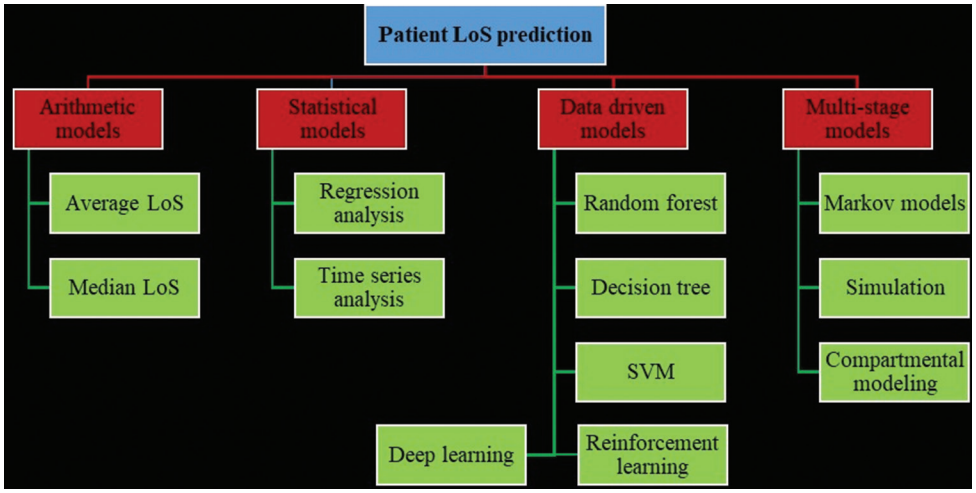



**Figure 1.** Length of stay prediction methods
Abbreviations: LoS: Length of stay; SVM: Support vector machine.

LoS in hospitals tends to increase when patients have comorbidities, which are additional medical conditions existing alongside the primary ailment.[23,24] In our study, we used the visit reason feature to create a comorbidity indicator. Each patient was assigned an integer value: "1" if they had no comorbidities or a single disease, and higher values (e.g., ≥2) if they had multiple comorbidities. Table 1 presents an overview of all variables utilized in this study, including their brief descriptions and data types. The main outcome variable was the LoS for admitted patients, measured as a continuous numeric variable with a wide distribution of durations. Notably, within the pediatric patient dataset examined, approximately 95% of cases exhibited LoS within 15 days. Instances of LoS surpassing 20 days were deemed outliers due to their infrequency and were consequently eliminated from the dataset.

**Table 1. Feature description**

| Field name | Description | Data type |
|---|---|---|
| Patient ID | Unique numeric identifier assigned to each patient | Integer |
| Address | District-level location of patient residence | String (open-text entry) |
| Street | Detailed residential location, including street and neighborhood | |
| Financial class | Classification of financial coverage or funding source | Categorical (5 categories) |
| Admission date | Calendar date on which the patient was admitted | Date (DD/MM/YYYY) |
| Discharge date | Calendar date on which the patient was discharged | Date (DD/MM/YYYY) |
| Visit reason | The initial reason for seeking care, as assessed by the emergency room consultant | String (open-text entry) |
| Length of stay | Total duration of hospitalization in days | Integer |
| Admission source | Indicates if admission was emergency-based or scheduled | Categorical (2 categories) |
| Date of birth | Patient calendar date of birth, used to derive patient age | Date (DD/MM/YYYY) |
| Admission care type | Accommodation type assigned upon admission (e.g., ward and private room) | Categorical (3 levels) |
| Admission care level | Classification of care intensity (e.g., general and intensive) | Categorical (4 levels) |
| Gender | Biological sex of the patient | Categorical (Male/Female) |

## 2.2. Evaluation measures

In this study, we evaluated the predictive model using internal validation, dividing the dataset into training and testing subsets. We randomly partitioned patient data, allocating three-fourths for training and one-fourth for testing. Model performance was primarily assessed using the mean squared log error (MSLE), defined in Equation I.

$$MSLE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(log(1+y_i) - log(1+\hat{y}_i)\right)^2 \quad \text{(I)}$$

The MSLE metric is considered a primary evaluation metric due to its ability to penalize errors proportionally to the size of predictions, making it suitable for our applications. For example, if the actual LoS is two days but predicted as 3 days, and another case has an LoS of 15 days but is predicted as 16.5 days, MSLE accounts for the relative significance of errors. Thus, a 1-day error in the shorter stay is treated as more significant than the same absolute

error in the longer stay. For a comprehensive evaluation, we also reported other metrics such as mean absolute error (MAE), mean square error, and root mean square error (RMSE). In this study, Python (version 3.12.4) served as the primary tool for data analysis, feature engineering, and model estimation, chosen for its support for ML, Bayesian inference, and natural language processing (NLP)-related tasks.

### 2.3. Pre-processing and feature engineering

In healthcare, clinical histories often consist of unstructured text, serving as a primary source of information regarding patients' diagnoses, medications, symptoms, and other clinical factors.[25] In our study, we utilized the named entity recognition (NER) technique to process two unstructured features: visit reason and address. The visit reason comprises textual statements of potential diagnoses recorded by attending physicians in the Emergency Room, while the address feature contains detailed textual descriptions of patients' locations.

To convert unstructured features into usable data for ML, we employed a systematic pre-processing approach followed by NER. Initially, lemmatization was applied to standardize word forms and enhance consistency across the dataset. Subsequently, part-of-speech tagging was applied to filter nouns, adjectives, and prepositions, which are indicative of named entities in clinical texts. Further refining steps included spelling error correction using specialized dictionaries and the removal of redundant phrases to streamline the dataset. In addition, medical abbreviations were expanded to ensure coherence. For the address feature, NER involved error correction in street, district, and area names, followed by the application of a location dictionary specific to Pakistan districts, particularly compiled to ensure accuracy. Overall, our NER methodology integrated linguistic analysis techniques with domain-specific knowledge to accurately extract named entities from clinical text, facilitating the creation of usable features for subsequent ML applications. In our study, we utilized the visit reason feature to derive a comorbidity indicator, assigning integer values to quantify the complexity of their medical conditions.

### 2.4. ML

This study evaluated and compared a range of ML models to identify the most suitable model for our problem, taking into account that variations in dataset characteristics and outcome variables can significantly influence model performance. We tested six standard regression models, encompassing seven different algorithms commonly used in the field[26-28] (Table 2). Patient data were randomly split into training (75%) and testing (25%) sets, and each algorithm was assessed using three metrics: MSLE, MAE, and RMSE.[26] Optimal hyperparameter configurations for each model were determined using an exhaustive grid search over a predefined list of hyperparameters (Table S1). For each hyperparameter combination, the testing error was recorded, and the lowest error was reported in the results.

**Table 2. Machine learning models brief description**

| Classification model | Description |
|---|---|
| Linear regression | A foundational algorithm used for modeling the relationship between input features and a continuous target variable. Fitting a linear equation to the training data enables predictions of the target variable based on new input data, making it a fundamental technique for regression tasks in predictive modeling. Algorithm: multivariate linear regression |
| Decision tree | A decision tree structures decisions as branches and outcomes as leaves, where internal nodes correspond to feature-based queries. It recursively partitions the data space to predict output values based on input feature thresholds. Algorithm: decision tree regressor |
| Bagging | Bagging, or bootstrap aggregation, builds multiple independent models on varied subsets of data. The aggregation of these models, typically decision trees, enhances generalization and minimizes overfitting. Algorithm: random forest regressor |
| Boosting | Boosting constructs a strong predictor by sequentially training weak learners, each compensating for the errors of its predecessors. Greater weight is given to previously mispredicted instances to refine subsequent models. Algorithm: extreme gradient boosting, Adaptive boosting |
| Nearest neighbor | The Nearest neighbor approach operates by identifying a set number of closest data points to a given point and predicting its value based on the average or weighted average of these neighboring points. Algorithm: K-nearest neighbor regression |
| Support vectors | This method aims to find the optimal separating hyperplane that maximizes the margin between data points and the decision boundary. Kernel functions can be applied to capture non-linear relationships in higher-dimensional spaces. Algorithm: support vector regression |

### 2.4. Bayesian modeling

This study also utilized a Bayesian framework to predict the LoS, integrating both prior knowledge and observed data. We started by specifying a probabilistic model for the LoS, assuming it follows a distribution conditioned on parameters. Let $L$ represent the LoS, and $\theta$ denotes the parameters governing its distribution. The likelihood function captures the probability of observing a specific LoS given the model parameters. Incorporating prior knowledge, we assigned a prior distribution $P(\theta)$ over

the parameters. Then, using Bayes' theorem, the posterior distribution of the parameters given the observed data D is obtained, as shown in Equation II:

$$p(\theta \mid D) = \frac{p(D \mid \theta) \times p(\theta)}{p(D)} \quad \text{(II)}$$

where P(D|θ) is the likelihood function, P(θ) is the prior distribution, and P(D) is the marginal likelihood. To obtain the posterior distribution of the LoS P(L|D), we integrated all possible parameter values, as shown in Equation III:

$$p(L \mid D) = \int \left[ p(L \mid \theta, D) \times p(\theta \mid D) \right] d\theta \quad \text{(III)}$$

This integral represents the predictive distribution of the LoS given the observed data. However, obtaining the exact form of the posterior distribution can be computationally challenging, especially for complex models. To address this, we employed the Markov Chain Monte Carlo method, a powerful class of algorithms, for efficient sampling from the posterior distribution.

We assumed non-informative densities as prior distributions for the coefficients, and tested multiple weakly informative priors for the coefficients, including chi-squared, Gamma, bounded normal, and Poisson distributions, with different parameter values. Using a wide prior expanded the search space, increasing the possibility of finding an optimal solution. Given that LoS follows a skewed distribution, the Gamma prior combined with a Poisson likelihood yielded the best results. Other skewed distributions, like Gamma and chi-squared, were also tested.[29] Bayesian multiple linear and non-linear models were trained using different distributions, parameters, and numbers of features to evaluate their predictive performance.

#### 2.4.1. Linear model (single feature)

First, we tested the Bayesian linear regression model with different non-informative prior distributions with one regressor. The first step of the Bayesian prediction is to choose/compute the prior probability distribution and likelihood distribution. Our first model is simple, with a conjugate normal prior with a mean of five and a standard deviation of 10 (Figure S1). The model is specified as below in Equations IV-VII.

$$\alpha, \beta = N(5,10) \quad \text{(IV)}$$

$$\sigma = \text{HC}(4) \quad \text{(V)}$$

$$\mu_i = \alpha + \beta x_i \quad \text{(VI)}$$

$$L|\mu_i, \sigma \sim N(\mu_i, \sigma) \quad \text{(VII)}$$

α and β are the prior distributions for the intercept and variable coefficient, respectively. The standard deviation of the likelihood function was modeled using the Half-Cauchy distribution throughout this study.

#### 2.4.2. Linear model (multiple features)

Similarly, the regression scenario was generalized in several ways using multiple variable regression settings, where the mean of a continuous response was written as a linear function of several predictor variables. Similar to a simple linear regression model, a multiple linear regression model assumes an observation-specific mean $\mu_i$ for the $i^{th}$ response variable Yi (Equation VIII).

$$L|\mu_i, \sigma \sim N(\mu_i, \sigma), i = 1....n \quad \text{(VIII)}$$

In addition, it assumes that the mean of Yi is $\mu_i$, a linear function of all predictors. With four predictors, it is written as in Equations IX-XI.

$$\mu_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} \quad \text{(IX)}$$

$$\alpha, \beta_r = \text{N}(5,10) \quad \text{(X)}$$

$$\sigma = HC(4) \quad \text{(XI)}$$

where $x_{i,1}$ is a predictor, like age and location, for observation i, and β = (β1, β2,⋯, βr) is a vector of unknown regression parameters (coefficients), shared among all observations. Equations X and XI are the prior distribution and standard deviation distribution, respectively.

#### 2.4.3. Non-linear model

While linear models assume a straight line relationship between the input and output data, a non-linear approach is more suitable when variables exhibit a curved relationship. There are many ways to model a curved relationship between two variables, including using higher-order polynomial terms, such as x squared, x cubed, or basic geometry or trigonometry functions, such as exponential or cosine functions. Similar to the linear model, in a non-linear model, an additional term with squares and cubes of the predictor variable was added. This only altered Equation XII while keeping the others unchanged.

$$\mu_i = \alpha + \beta_1 x_{i,1} + \beta_{11} x_{i,1}^{\ 2} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} \quad \text{(XII)}$$

## 3. Results

Table 3 summarizes the general findings from the patient data. The LoS exhibits a skewed distribution, with a mean of 3.82 days and a median of 2 days. Approximately 95% of patients have a LoS of <15 days, and over 82% of patients have a LoS of <5 days. The average age of patients admitted to the hospital is 4.2 years, which also displays a skewed distribution, with 75% of patients being under seven years old. Seasonal patterns are evident in the data, as observed

**Table 3. Statistical findings**

| Variable | Values (*n*=22,106) |
|---|---|
| Length of stay (LoS), days (median [IQR]) | 2 (4) |
| LoS <5 days | 82% (18,114) |
| LoS <10 days | 92% (20,406) |
| LoS <15 days | 95% (21,056) |
| Sex | |
| Male | 60% (13,283) |
| Female | 40% (8823) |
| Clinician prediction | |
| True prediction | 64.7% (14,304) |
| LoS (median [IQR]) | 1.6 [1.0, 2.4] |
| False prediction | 35.3% (78,02) |
| LoS (median [IQR]) | 3.9 [2.6, 7.0] |
| Age, years (mean, standard deviation) | 4.2 (5.0) |
| Hospital admission (months) | |
| January | 8.5% (1878), mean: 3.8 |
| February | 7.7% (1702), mean: 4.2 |
| March | 7.6% (1676), mean: 4.1 |
| April | 7.5% (1665), mean: 4.1 |
| May | 7.9% (1747), mean: 3.9 |
| June | 5.9% (1316), mean: 4.5 |
| July | 8.5% (1898), mean: 3.8 |
| August | 9.1% (2002), mean: 3.6 |
| September | 9.5% (2103), mean: 3.8 |
| October | 9.2% (2027), mean: 3.7 |
| November | 8.1% (1799), mean: 3.4 |
| December | 9.0% (1900), mean: 3.2 |

Note: Data are presented as percentages (numbers), unless specified otherwise.

from the monthly distribution of patient admissions. In addition to these general findings, we examined the independent associations between various factors and LoS. We included seven features, and their importance was evaluated using a random forest model, as shown in Figure 2. The results indicated that age, location, and type of admission care have a more significant impact on LoS compared to other features.

The gradient boosting regressor demonstrated the best performance in terms of MSLE, closely followed by extreme gradient boosting. The support vector regressor achieved the lowest MAE, with values below 1. Notably, there is a significant disparity between the mean and median absolute error, attributable to the skewed distribution of the LoS (Table 4).

Our investigation into predicting LoS using Bayesian models yielded several noteworthy findings. Initially, we

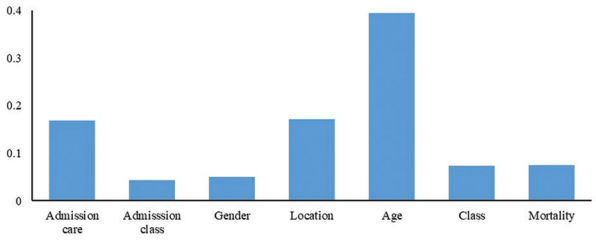



**Figure 2.** Significance of the random forest-based feature

examined the performance of various Bayesian models, each focused on a single predictor based on the results of a random forest feature importance analysis. Surprisingly, despite age being the most significant predictor in ML, we did not achieve satisfactory results when using it as the sole predictor. However, a linear model incorporating admission care as the predictor, with normal priors and a Poisson likelihood, emerged as the most effective combination (Table 5). Expanding our analysis, we integrated multiple predictors to construct Bayesian models. Utilizing the top five predictors identified by the random forest feature selector – age, admission care, location, gender, and mortality – we computed various combinations to predict LoS. Notably, a linear model incorporating age, admission care, location, and gender produced the most promising results (Table 5). While age stood out as the most significant predictor, our findings revealed that admission care level outperformed age when assuming a linear relationship with LoS in the Bayesian framework. This discrepancy prompted us to explore non-linear models by introducing squared, cubed, and exponential transformations of age. Subsequent analysis demonstrated that incorporating non-linearity in age prediction led to improved results compared to linear models (Table 5).

The optimal configuration among Bayesian models included four predictors: age, admission care, location, and gender, utilizing Gamma prior distributions and Poisson likelihood functions. This Bayesian model yielded results comparable to those of ML algorithms, achieving an MSLE of 0.25. Furthermore, the Bayesian approach offered robust performance across various error measures, underscoring its utility in healthcare decision-making. The incorporation of uncertainty information through predictive distributions enhanced the applicability of results within decision-theoretic frameworks, thereby augmenting the power and generalization of our findings.

## 4. Discussion

### 4.1. Key findings

The landscape of healthcare analytics is ever-evolving, largely driven by the exponential growth of patient data

**Table 4. Predictive performance of machine learning models**

| Models | Mean squared log error | Root mean standard error | Mean absolute error | Median absolute error |
|---|---|---|---|---|
| Linear regression | 0.27 | 2.39 | 1.68 | 1.31 |
| Decision tree | 0.24 | 2.22 | 1.53 | 1.25 |
| Random forest | 0.27 | 2.35 | 1.61 | 1.08 |
| K-nearest neighbor | 0.27 | 2.38 | 1.62 | 1.00 |
| Support vector machine | 0.24 | 2.38 | 1.45 | 0.96 |
| Extreme gradient boosting | 0.23 | 2.16 | 1.47 | 1.06 |
| Adaptive boosting | 0.29 | 2.28 | 1.72 | 1.74 |

**Table 5. Predictive performance of Bayesian models**

| Model | Predictors | Distribution prior/likelihood | Mean squared log error | Root mean standard error | Mean absolute error |
|---|---|---|---|---|---|
| Linear regression | | | | | |
| Linear model-0 (with outliers) | 0 | Normal/Normal | 0.42 | 4.39 | 2.47 |
| Linear model-1 | 0 | Normal/Normal | 0.3 | 2.52 | 1.76 |
| Linear model-2 | 0 | Gamma/Normal | 0.3 | 2.52 | 1.75 |
| Linear model-3 | 0 | Gamma/Gamma | 0.3 | 2.52 | 1.75 |
| Linear model-4 | 0 | Chi-squared/Chi-squared | 0.3 | 2.52 | 1.76 |
| Linear model-5 | 0 | Chi-squared/Poisson | 0.3 | 2.51 | 1.74 |
| Linear model-6 | 2 | Normal/Poisson | 0.3 | 2.48 | 1.73 |
| Linear model-7 | 1 | Normal/Poisson | 0.25 | 2.21 | 1.56 |
| Linear model-8 | 3 | Normal/Poisson | 0.3 | 2.49 | 1.75 |
| Linear model-9 | 2 | Normal/Chi-squared | 0.29 | 2.45 | 1.73 |
| Linear model-10 | 1 | Normal/Chi-squared | 0.27 | 2.35 | 1.65 |
| Multivariate linear regression (MLR) | | | | | |
| MLR model-1 | 0, 2 | Normal/Normal | 0.44 | 4.74 | 2.55 |
| MLR model-2 | 0, 2 | Chi-squared/Chi-squared | 0.4 | 4.76 | 2.36 |
| MLR model-3 | 1, 2 | Normal/Poisson | 0.26 | 2.27 | 1.58 |
| MLR model-4 | 0, 1 | Normal/Poisson | 0.26 | 2.24 | 1.57 |
| MLR model-5 | 0, 1, 2 | Chi-squared/Poisson | 0.24 | 2.27 | 1.57 |
| MLR model-6 | 0,1, 2 | Gamma, normal, normal/Poisson | 0.25 | 2.32 | 1.59 |
| MLR model-7 | 0, 1, 2 | Chi-squared, normal, normal/Poisson | 0.25 | 2.28 | 1.57 |
| MLR model-8 | 0, 1, 2 | Chi-squared, normal, normal/Gamma | | | |
| MLR model-9 | 0, 1, 2, 4 | Normal/Normal | 0.25 | 2.22 | 1.55 |
| MLR model-10 | 0, 1, 2, 4 | Chi-squared/Chi-squared | 0.26 | 2.23 | 1.6 |
| MLR model-11 | 0, 1, 2, 4 | Chi-squared/Poisson | 0.25 | 2.21 | 1.55 |
| MLR model-12 | 0, 1, 2, 4 | Gamma/Poisson | 0.25 | 2.20 | 1.54 |
| MLR model-13 | 0, 1, 2, 4 | Gamma/Poisson | 0.25 | 2.24 | 1.56 |
| MLR model-14 | 0, 1, 2, 3, 4 | Chi-squared/Poisson | 0.26 | 2.28 | 1.59 |
| Non-linear regression (NLR) | | | | | |
| NLR model-1 | 0, 02 | Chi-squared, Gamma/Normal | 0.3 | 2.4 | 1.71 |
| NLR model-2 | 0, $0^2$, 1, 2 | Chi-squared/Normal | 0.26 | 2.3 | 1.59 |
| NLR model-3 | 0, $0^2$, 1, 2 | Chi-squared, Gamma, Gamma/Poisson | 0.25 | 2.26 | 1.56 |

*(Cont'd...)*

**Table 5.** ***(Continued)***

| Model | Predictors | Distribution prior/likelihood | Mean squared log error | Root mean standard error | Mean absolute error |
|---|---|---|---|---|---|
| NLR model-4 | 0, $0^2$, 1, 2 | Chi-squared, normal, Gamma/Chi-squared | 0.26 | 2.27 | 1.61 |
| NLR model-5 | 0, $0^3$, 1, 2 | Chi-squared, Gamma, Gamma/Poisson | 0.26 | 2.26 | 1.58 |
| NLR model-5 | 0, $0^2$, 1, $1^2$, 2 | Chi-squared/Normal | 0.28 | 2.66 | 1.73 |

Note: Predictor coding: 0: Age; 1: Admission care; 2: Location; 3: Mortality; 4: Gender.

and advancements in predictive methodologies. This study aimed to predict hospital LoS, a critical factor in healthcare resource management and patient care planning. We discovered that both Bayesian and ML models have good predictive power. Particularly, boosting models achieved the lowest MSLE, corroborating existing literature that highlights their superior regression performance. Conversely, support vector regression proved to be the most robust and widely applicable model, given its consistent performance across multiple evaluation metrics. While Bayesian models did not outperform ML algorithms in terms of error measures, they provided a useful measure of uncertainty. This is particularly helpful for medical professionals as the model can identify predictions that have a higher level of uncertainty, thus increasing effectiveness and reliability when making decisions.

### 4.2. Comparison with similar research

Our research adds to the growing body of literature on the prediction of hospital stay, offering a broader scope by utilizing a large pediatric dataset rather than focusing on a specific disease subgroup. The superior performance of ensemble methods like extreme gradient boosting aligns with other studies, highlighting its ability to model complex and non-linear relationships in healthcare data. The use of Bayesian modeling aims to fill a gap in existing literature by exploring its potential to complement traditional ML models with the added feature of uncertainty estimation. In addition, the integration of NER for feature engineering is a significant methodological contribution, improving the models' predictability while integrating recent advancements in NLP within healthcare. The comparative analysis in Table 6 reveals comparable performance scores in predicting LoS. Our approach stands out for its broader applicability and superior performance across various regression models and metrics.

### 4.3. Strengths and limitations

This study is distinguished by its comprehensive evaluation of predictive models, incorporating both traditional ML techniques and Bayesian inference methods. The use of Bayesian models for LoS prediction offers the added advantage of uncertainty quantification. By processing unstructured clinical text, NER strengthens the predictive power of our models, contributing to the broader field of NLP within healthcare. However, the study is limited by constraints inherent in the dataset, such as the lack of detailed patient medical conditions, relying instead on initial physician diagnoses. The absence of variables such as temperature, blood pressure, or other physiological indicators restricts the depth of analysis and predictive accuracy.

### 4.4. Implications and future actions

The findings of this study include several practical implications for hospital administrators, healthcare policymakers, and clinical practitioners aiming to improve operational efficiency and patient care. The predictive capabilities of both ML and Bayesian models have the potential for integrating such tools into operational hospital management systems. Accurate prediction of LoS can significantly enhance bed occupancy planning, reduce bottlenecks in patient flow, and optimize resource distribution across departments, particularly in resource-constrained settings like Pakistan.

Our study also paves the way for future research, particularly in developing hybrid models that combine the strengths of ML and Bayesian approaches. Moreover, the study highlights the value of transforming unstructured clinical text into structured features through NLP. Hospitals should prioritize digitization and structured documentation of clinical records to facilitate future predictive analytics applications. Building robust data infrastructures and standardizing terminology used in reason for visit and diagnostic notes would enhance model accuracy.

On a broader policy level, healthcare authorities should consider investing in model development and validation for other key outcomes, including readmissions and mortality. The integration of Bayesian frameworks, which allow uncertainty quantification, also presents opportunities for embedding probabilistic reasoning into clinical decision-support systems, offering a cautious and more reliable approach to automation in healthcare.

Finally, future work should aim to expand data collection to include real-time physiological metrics, laboratory

**Table 6. Summary of results of prior studies**

| Study | Method used | Results | Data size |
|---|---|---|---|
| Turgeman *et al.*[18] | Regression tree (Cubist) model | MAE: 1.0, $R^2$: 0.79 | 20,321 |
| Liu *et al.*[24] | Linear regression | MSE: 0.029, $R^2$: 0.146 | 155,474 |
| Lee *et al.*[30] | Multivariable regression model | MAE: 7.6, RMSE: 11 | 22,824 |
| Muhlestein *et al.*[31] | Gradient boosted trees, SVM, others | RMSLE: 0.631 | 41,222 |
| Medeiros *et al.*[32] | Regression techniques, ML techniques | RMSE: 2.5 – 4.26 | 23,551 |
| Alsinglawi *et al.*[33] | Gradient boosting, random forest, DNN | MAE: 2.0, $R^2$ 0.81, | 61,532 |
| Zolbanin *et al.*[34] | DNN | MAE: 1.239, RMSE: 2.063<br>$R^2$: 0.613 | 86,338 |
| Fang *et al.*[35] | Bayesian neural network | MAE: 1.955, $R^2$: 0.098 | 200,000 |
| Muhlestein *et al.*[31] | ML algorithms | RMLSE: 0.661 | 41,222 |
| Abdurrab *et al.*[36] | Bayesian regression versus ML regressors | RMSE 3.36, MAE 1.98 | 5,636 |
| Lequertier *et al.*[37] | Embeddings+FFNN (deep learning) | Accuracy 0.944 | 515,199 |
| Hu *et al.*[38] | Random forest, XGBoost, SVM, DNN (ML meta-analysis) | RMSE~5.8 – 7.0<br>$R$ ~0.10 – 0.38 | 10,700,000 |
| Rocheteau *et al.*[39] | TPC (deep time-series CNN) | MAE 1.55 – 2.28 | MIMIC-IV database |

Abbreviations: CNN: Convolutional neural network; DNN: Deep neural network; FFNN: Feed forward neural network; MAE: Mean absolute error; ML: Machine learning; MSE: Mean squared error; RMSE: Root mean squared error; RMSLE: Root mean squared log error; SVM: Support vector machine; TPC: Temporal pointwise convolutional; XGBoost: Extreme gradient boosting.

results, and longitudinal comorbidity information. These additional layers of data would likely improve predictive accuracy.

## 5. Conclusion

This study underscores the potential of both ML and Bayesian models in predicting hospital LoS, with each approach offering unique strengths. ML models, particularly boosting algorithms, have an excellent predictive accuracy, while Bayesian models offer valuable insights into prediction uncertainty. Integrating these models into healthcare systems could significantly improve resource management and patient care. Future research should explore hybrid approaches and incorporate more detailed patient data to further enhance predictive capabilities.

## Acknowledgments

We extend our sincere gratitude to the management of Aga Khan University Hospital for providing the data to conduct this research. We are also thankful for their valuable subject expertise, which greatly contributed to the success of our study.

## Funding

None.

## Conflict of interest

The authors declare that they have no competing interests.

## Author contributions

*Conceptualization:* Tariq Mahmood, Babar Hasan, Zahra Hoodbhoy

*Formal analysis:* Sarmad Zafar, Tariq Mahmood, Zahra Hoodbhoy

*Investigation:* Sarmad Zafar, Tariq Mahmood, Zahra Hoodbhoy

*Methodology:* Sarmad Zafar, Tariq Mahmood, Babar Hasan

*Writing–original draft:* Sarmad Zafar, Tariq Mahmood

*Writing–review & editing:* All authors

## Ethics approval and consent to participate

This study is retrospective in nature, utilizing historical data from patients admitted to the hospital, with all procedures conducted as part of routine care following established clinical practices. The data provided by Aga Khan University Hospital is anonymized to protect patient privacy, aligning with ethical guidelines and ensuring confidentiality. The hospital has obtained consent from all patients to use their data, without disclosing any identifying information, for research purposes.

## Consent for publication

The Aga Khan University Hospital has obtained consent from all patients to use their data, without disclosing any identifying information, for research purposes, including the publication of results.

## Availability of data

The data used for this research was provided by Aga Khan University Hospital under a confidentiality agreement. Data will be shared upon request to the corresponding author, subject to the permission from Aga Khan University Hospital and the signing of a Non-Disclosure Agreement (NDA) if required.

## References

1. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*. 2017;5:8869-8879.

   doi: 10.1109/access.2017.2694446

2. Jain R, Singh M, Rao AR, Garg R. Predicting hospital length of stay using machine learning on a large open health dataset. *BMC Health Serv Res*. 2024;24(1):860.

   doi: 10.1186/s12913-024-11238-y

3. Bopche R, Gustad LT, Afset JE, Ehrnström B, Damås JK, Nytrø Ø. In-hospital mortality, readmission, and prolonged length of stay risk prediction leveraging historical electronic patient records. *JAMIA Open*. 2024;7(3):ooae074.

   doi: 10.1093/jamiaopen/ooae074

4. Kothinti RR. Deep learning in healthcare: Transforming disease diagnosis, personalized treatment, and clinical decision-making through AI-driven innovations. *World J Adv Res Rev*. 2024;24(2):2841-2856.

   doi: 10.30574/wjarr.2024.24.2.3435

5. Van Houdenhoven M, Nguyen DT, Eijkemans MJ, *et al*. Optimizing intensive care capacity using individual length-of-stay prediction models. *Crit Care*. 2007;11(2):R42.

   doi: 10.1186/cc5730

6. Barnes SL, Hamrock E, Toerper MF, Siddiqui S, Levin SR. Real-time prediction of inpatient length of stay for discharge prioritization. *J Am Med Inform Assoc*. 2016;23(e1):e2-e10.

   doi: 10.1093/jamia/ocv106

7. Mahyoub MA, Dougherty K, Yadav RR, Berio-Dorta R, Shukla A. Development and validation of a machine learning model integrated with the clinical workflow for inpatient discharge date prediction. *Front Digit Health*. 2024;6:1455446.

   doi: 10.3389/fdgth.2024.1455446

8. Wessman T, Ärnlöv J, Carlsson AC, *et al*. The association between length of stay in the emergency department and short-term mortality. *Intern Emerg Med*. 2021;17(1):233-240.

   doi: 10.1007/s11739-021-02783-z

9. Arabi Y, Venkatesh S, Haddad S, Al Shimemeri A, Al Malik S. A prospective study of prolonged stay in the intensive care unit: Predictors and impact on resource utilization. *Int J Qual Health Care*. 2002;14(5):403-410.

   doi: 10.1093/intqhc/14.5.403

10. Gruenberg DA, Shelton W, Rose SL, Rutter AE, Socaris S, McGee G. Factors influencing length of stay in the intensive care unit. *Am J Crit Care*. 2019;15(5):502-509.

    doi: 10.4037/ajcc2006.15.5.502

11. Gustafson DH. Length of stay: Prediction and explanation. *Health Serv Res*. 2025;3(1):12-34.

12. Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S. Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. *PLoS One*. 2018;13(4):e0195901.

    doi: 10.1371/journal.pone.0195901

13. Awad A, Bader-El-Den M, McNicholas J. Patient length of stay and mortality prediction: A survey. *Health Serv Manag Res*. 2017;30(2):105-120.

    doi: 10.1177/0951484817696212

14. Stone K, Zwiggelaar R, Jones P, Mac Parthaláin N. A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*. 2022;1(4):e0000017.

    doi: 10.1371/journal.pdig.0000017

15. Morton A, Marzban E, Giannoulis G, Patel A, Aparasu R, Kakadiaris IA. *A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term in-Hospital Length of Stay among Diabetic Patients*. United States: IEEE Xplore; 2014. p. 428-43.

    doi: 10.1109/ICMLA.2014.76

16. Ma F, Yu L, Ye L, Yao DD, Zhuang W. Length-of-stay prediction for pediatric patients with respiratory diseases using decision tree methods. *IEEE J Biomed Health Inform*. 2020;24(9):2651-2662.

    doi: 10.1109/jbhi.2020.2973285

17. Goh KH, Wang L, Yeow AYK, *et al*. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun*. 2021;12(1):711.

    doi: 10.1038/s41467-021-20910-4

18. Turgeman L, May JH, Sciulli R. Insights from a machine learning model for predicting the hospital length of stay (LOS) at the time of admission. *Exp Syst Appl*. 2017;78:376-385.

    doi: 10.1016/j.eswa.2017.02.023

19. Fortuin V. Priors in bayesian deep learning: A review. *Int Statis Rev*. 2022;90:563-591.

    doi: 10.1111/insr.12502

20. Austin PC, Naylor CD, Tu JV. A comparison of a bayesian vs. A frequentist method for profiling hospital performance. *J Eval Clin Pract*. 2001;7(1):35-45.

doi: 10.1046/j.1365-2753.2001.00261.x

21. Abdullah AA, Hassan MM, Mustafa YT. A review on bayesian deep learning in healthcare: Applications and challenges. *IEEE Access*. 2022;10:36538-36562.

    doi: 10.1109/access.2022.3163384

22. Ruhe D, Cinà G, Tonutti M, Bruin D, Elbers P. *Bayesian Modelling in Practice: Using Uncertainty to Improve Trustworthiness in Medical Applications*; 2019. Available from: https://arxiv.org/abs/1906.08619 [Last accessed on 2025 Jul 10].

23. Yang Y, Yang KS, Hsann YM, Lim V, Ong BC. The effect of comorbidity and age on hospital mortality and length of stay in patients with sepsis. *J Crit Care*. 2010;25(3):398-405.

    doi: 10.1016/j.jcrc.2009.09.001

24. Liu V, Kipnis P, Gould MK, Escobar GJ. Length of stay predictions: Improvements through the use of automated laboratory and comorbidity variables. *Med Care*. 2010;48(8):739-744.

    doi: 10.1097/mlr.0b013e3181e359f3

25. Durango MC, Torres-Silva EA, Orozco-Duque A. Named entity recognition in electronic health records: A methodological review. *Healthc Inform Res*. 2023;29(4):286-300.

    doi: 10.4258/hir.2023.29.4.286

26. Mitchell T. *Machine Learning*; 1997. Available from: https://www.cs.cmu.edu/~tom/files/machinelearningtommitchell.pdf [Last accessed on 2025 Jul 21].

27. Bishop CM. *Pattern Recognition and Machine Learning*. Berlin: Springer; 2006.

28. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res*. 2014;15:3133-3181.

29. Williford E, Haley V, McNutt LA, Lazariu V. Dealing with highly skewed hospital length of stay distributions: The use of Gamma mixture models to study delivery hospitalizations. *PLoS One*. 2020;15(4):e0231825.

    doi: 10.1371/journal.pone.0231825

30. Lee H, Bennett MV, Schulman J, Gould JB, Profit J. Estimating length of stay by patient type in the neonatal intensive care unit. *Am J Perinatol*. 2016;33(8):751-757.

    doi: 10.1055/s-0036-1572433

31. Muhlestein WE, Akagi DS, Davies JM, Chambless LB. Predicting inpatient length of stay after brain tumor surgery: Developing machine learning ensembles to improve predictive performance. *Neurosurgery*. 2018;85(3):384-393.

    doi: 10.1093/neuros/nyy343

32. Medeiros NB, Fogliatto FS, Rocha MK, Tortorella GL. Forecasting the length-of-stay of pediatric patients in hospitals: A scoping review. *BMC Health Serv Res*. 2021;21(1):938.

    doi: 10.1186/s12913-021-06912-4

33. Alsinglawi B, Alnajjar F, Mubin O, *et al*. *Predicting Length of Stay for Cardiovascular Hospitalizations in the Intensive Care Unit: Machine Learning Approach*. United States: IEEE Xplore; 2020. p. 5442-5445

    doi: 10.1109/EMBC44109.2020.9175889

34. Zolbanin HM, Davazdahemami B, Delen D, Zadeh AH. Data analytics for the sustainable use of resources in hospitals: Predicting the length of stay for patients with chronic diseases. *Inform Manag*. 2020;59:103282.

    doi: 10.1016/j.im.2020.103282

35. Fang J, Zhu J, Zhang X. Prediction of length of stay on the intensive care unit based on bayesian neural network. *J Phys Conf Ser*. 2020;1631(1):012089.

    doi: 10.1088/1742-6596/1631/1/012089

36. Abdurrab I, Mahmood T, Sheikh S, *et al*. Predicting the length of stay of cardiac patients based on pre-operative variables-bayesian models vs. Machine learning models. *Healthcare (Basel)*. 2024;12(2):249.

    doi: 10.3390/healthcare12020249

37. Lequertier V, Wang T, Fondrevelle J, Augusto V, Polazzi S, Duclos A. Length of stay prediction with standardized hospital data from acute and emergency care using a deep neural network. *Med Care*. 2024;62(4):225-234.

    doi: 10.1097/mlr.0000000000001975

38. Hu Z, Qiu H, Wang L, Shen M. Network analytics and machine learning for predicting length of stay in elderly patients with chronic diseases at point of admission. *BMC Med Inform Decis Mak*. 2022;22(1):62.

    doi: 10.1186/s12911-022-01802-z

39. Rocheteau E, Liò P, Hyland SL. *Temporal Pointwise Convolutional Networks for Length of Stay Prediction in the Intensive Care Unit*. United States: Cornell University; 2021.

    doi: 10.1145/3450439.3451860

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

# M2Echem: A multilevel dual encoder-based model for predicting organic chemistry reactions

**Linxing Zhu[1†], Jing Wang[1†], Jiashuang Huang[1], Yifan Jiang[2*], and Shu Jiang[1*]**

[1]Department of Computer Science, School of Artificial Intelligence and Computer Science, Nantong University, Nantong, Jiangsu, China

[2]Department of Electrical and Computer Engineering, State Key Laboratory of Internet of Things for Smart City, Faculty of Science and Technology, University of Macau, Macau, China

## Abstract

Chemical reaction prediction is a vital application of artificial intelligence. While Transformer models are widely used for this task, they often overlook deeper-level semantic information. In addition, the traditional Transformer model suffers from a decline in prediction performance and shows poor generalization when faced with different representations of the same molecule. To address these challenges, we propose a dual encoder-based reaction prediction method tailored for multilevel organic chemistry. Our approach began with the introduction of synergistic dual-encoder architecture: The atomic encoder focused on inter-atomic attention weights. In contrast, the molecular encoder employed a molecular maximum dimension reduction algorithm to identify key chemical features. We then performed multilevel feature fusion by combining the outputs from both the atomic and molecular encoders. Finally, we applied an optimized contrast loss to enhance the model's robustness. The results indicated that this method outperformed existing models across all four datasets, significantly improving generalization performance and contributing to advancements in artificial intelligence-driven drug development and research.

***Keywords:*** Forward reaction prediction; Multilevel feature fusion; Machine learning; Simplified molecular input line entry system code; Transformer

*†These authors contributed equally to this work.*

***Corresponding authors:**
Yifan Jiang
(yc27495@umac.mo)
Shu Jiang
(jshmjs45@ntu.edu.cn)

**Citation:** Zhu L, Wang J, Huang J, Jiang Y, Jiang S. M2Echem: A multilevel dual encoder-based model for predicting organic chemistry reactions. *Artif Intell Health*. 2026;3(1):88-103. doi: 10.36922/AIH025260058

**Received:** June 26, 2025

**1st revised:** July 15, 2025

**2nd revised:** July 21, 2025

**Accepted:** July 23, 2025

**Published online:** August 5, 2025



## 1. Introduction

As early as the 1960s, organic chemists attempted to use modern computers for planning synthetic pathways[1]. However, these early efforts relied heavily on reaction templates[1-3]—manual rules developed by experts—or other forms of chemical knowledge.[4-6] Such approaches struggled to address the full complexity of organic chemistry prediction problems.

In the 1980s, the Simplified Molecular Input Line Entry System (SMILES)[7] was developed to represent the structure of chemical compounds, allowing computers to manage chemical reactions in a linear format. Consequently, chemical reactions could be

regarded as a specific language and modeled symbolically for computer-assisted reaction prediction. Many text processing methods have been utilized in chemical reaction prediction, including the sequence-to-sequence model,[8] with the Transformer model[9] gaining widespread use in neural machine translation.

Traditional chemical reaction prediction methods use SMILES representations to segment chemical formulas at the atomic level and apply the Transformer model.[10] However, these approaches can only capture attention weights between individual atoms, thereby overlooking important intermolecular interactions. Moreover, the Transformer model struggles with handling long sequences,[11,12] which frequently occur in chemical reaction text.

We analyzed the vocabulary size and sequence length of natural language and chemical reaction texts, highlighting their significant differences (Table 1). The units in the table are based on words. Chemical reaction texts exhibit longer sequences than natural language. While attention is calculated between any two tokens, long sequences can lead to insufficient encoding. In natural language processing, missing parts of text often have minimal effect on comprehension. In contrast, missing or altered structures in a chemical molecular formula may severely compromise the overall representation.

In this study, we propose a dual-encoder method for multilevel organic chemical reaction prediction—M2Echem—to address the problems of limited attention to molecular interactions, insufficient encoding, and poor generalization ability of the Transformer model. First, the Transformer model is enhanced using two encoders to simultaneously construct representations of atoms and molecules in the input chemical formula. The atomic encoder utilizes the self-attention mechanism of the Transformer model to capture the interrelationships between atoms. In contrast, the molecular encoder employs the molecular maximum dimension reduction algorithm to generate molecular embeddings. The outputs of the atomic and molecular encoders undergo multilevel feature fusion, and the fused feature representations are input into the decoder. Finally, the optimized loss function enhances the model's understanding of different SMILES representations.

**Table 1. Comparison between natural language and chemical reaction text**

| Dataset | Corpus size | Vocabulary size | Length |
|---|---|---|---|
| WMT17-en[a] | 4,004,240 | 40,394 | 28.44 |
| WMT17-en[b] | 1,104,577 | 35,483 | 28.43 |
| USPTO (original)[c] | 479,035 | 448 | 375.00 |
| Single reactant | 12,581 | 175 | 87.51 |
| Multiple reactants | 1,091,996 | 574 | 114.17 |

Notes: [a]WMIT17-en refers to the English corpus from the Second Conference on Machine Translation, [b]WMT17-en (random) refers to randomly selected sentences from WMT17-en for a fair comparison, and [c]USPTO (canonical) refers to the canonical version of the tokenized chemical dataset USPTO-Jin.
Abbreviation: USPTO: United States Patent and Trademark Office.

The main contributions of this study are as follows:

(i) The proposed model utilizes molecular and atomic encoders to extract information at both molecular and atomic levels. Key features in molecules are identified using a molecular maximum dimension reduction algorithm. Compared to traditional methods, this approach effectively captures hierarchical correlations between atoms and molecules, thereby improving the encoding capabilities of the Transformer model for processing long-sequence data.

(ii) A fusion layer with automatic weight updating is proposed for multilevel feature fusion. This layer uses linear concatenation to fuse the outputs of the atomic and molecular encoders. It also incorporates batch normalization and a softmax activation function to obtain weight parameters. Unlike simple concatenation, this design enables deeper extraction of complementary information and promotes the integration of multiple features.

(iii) The proposed model generates augmented SMILES representations with implicit positive and negative labels for contrastive learning. Dimension compression is then applied to retain all the essential features required for this learning process. The contrastive learning loss is combined with the cross-entropy loss function, enabling automatic updates of the weight parameters.

## 2. Related work

This section includes three parts, focusing on essential methods relevant to this study, such as chemical product prediction methods, multilevel feature fusion approaches, and comparative learning.

### 2.1. Chemical product prediction methods

Large language models effectively capture long-distance dependencies through a self-attention mechanism, enabling stronger feature representation and parallel computing capabilities.[13-15] These models have shown excellent performance in predicting organic chemistry outcomes.[16] However, their performance is influenced by both the quality and quantity of training data. Therefore, extracting more accurate and detailed chemical information is essential for improving model predictions,

particularly when computational resources and chemical datasets are limited.

Researchers have developed various methods[17,18] to enhance or extend the SMILES language. Lo *et al.*[19] proposed SELFIES disambiguation to obtain valid SMILES molecules. Ucak *et al.*[20] proposed atom-in-SMILES disambiguation, where the model learns chemical information within the radius of covalent bonding between atoms to reduce molecular labeling duplication. While these methods provide in-depth chemical insights and enhance the accuracy of chemical reaction predictions, the large word lists used during model training often overlook the overall information contained within the molecule.

In contrast, models can learn more chemical information through pre-training and fine-tuning phases or multi-task training. For example, Wu *et al.*[21] proposed the knowledge-enabled language representation model, which acquires chemical information through atomic feature prediction, molecular feature prediction, and comparison learning. Chen and Jung[22] introduced the LocalRetro model, which utilizes mechanisms of local reactivity and global attention. Liu *et al.*[23] developed the MolXPT model, which combines scientific text with SMILES representations of chemical molecules for pre-training to improve the performance. Lu and Zhang[24] created the T5Chem model, which takes advantage of mutual learning among related tasks. Although the models extract deeper information, they all employ a single encoder-decoder architecture, which makes it difficult to optimize the extraction of information on different features simultaneously.

### 2.2. Multilevel feature fusion approaches

Multilevel feature fusion methods create a unified representation by mapping unimodal representations into a shared semantic subspace, allowing for the integration of multimodal features.[25] Joint representations can be categorized into two types: Feature-level fusion and model-level fusion.

Feature-level fusion integrates features from different modalities into a unified representation. For instance, Ma *et al.*[26] proposed an early fusion unified encoder model, Flat-Transformer, which concatenates context and source sentence features in the same space. Zhang *et al.*[27] proposed the multi-grained Bidirectional Encoder Representations from Transformers (BERT) model, which utilizes a dual encoder to extract information from chemical SMILES sequences. The two feature matrices generated by the encoders are combined and fed into a decoder, enhancing the performance of natural language understanding tasks. However, directly fusing features from different modalities does not effectively capture complex relationships.

Model-level fusion refers to the simultaneous processing of inputs from several models. Zhu *et al.*[28] enhanced machine translation performance by integrating BERT features into the encoder and decoder layers of the neural machine translation model using the attention mechanism. This approach often requires vital computational resources and longer training times.

### 2.3. Comparative learning

Several studies have shown that machine learning-based natural language processing models are vulnerable to minor disturbances.[29,30] To prevent the model from being affected by synonyms and different SMILES representations, contrastive learning can help models identify semantic similarities among different SMILES sequences representing the same chemical formula.

Gao *et al.*[31] proposed the Simple Contrastive Learning of Sentence Embeddings framework based on BERT, which generates positive samples in an unsupervised manner using dropout and optimizes contrastive loss to enhance sentence embeddings. Wu *et al.*[21] proposed that the knowledge-enabled language representation model enhances the dataset by dividing positive and negative samples, which are then input into the model for contrastive learning. While these methods generate pairs of positive and negative samples that retain semantic information, the samples are typically augmented independently, and feature dimensionality reduction using token classification is commonly employed in BERT models. In contrast, Chen *et al.*[32] proposed a contrastive learning loss mechanism that obtains local and global losses through average pooling-based feature dimensionality reduction for scientific literature-related work generation tasks. Improving the relevance of generated text requires sampling negative examples from numerous non-references, as average pooling can lead to information loss.

## 3. Methods

This section introduces a multilevel dual encoder model—M2Echem—designed for predicting organic chemistry reactions. The M2Echem model is built upon the T5chem framework, as shown in Figure 1. Compared to T5chem, the M2Echem model incorporated three significant modifications: A feature extraction module, a multilevel feature fusion module, and a fused loss function module.

### 3.1. Feature extraction module

The M2Echem model employed character-level tokenization to segment reaction SMILES into individual alphabet letters, digits, or special symbols. The processed

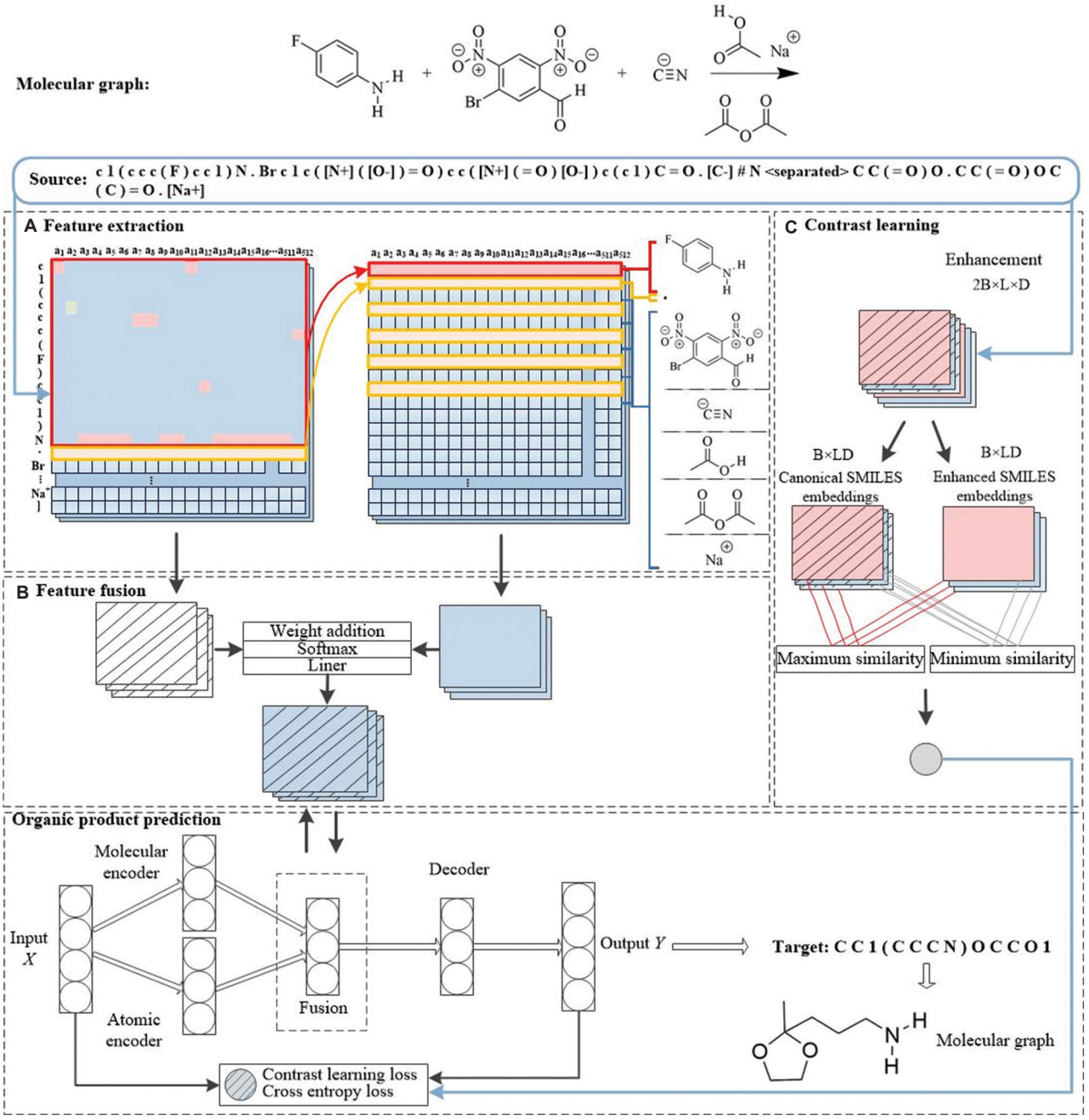



**Figure 1.** Model diagram of M2Echem. (A) Feature extraction: extraction of atomic-level and molecular-level features. (B) Feature fusion: Multi-level fusion of atomic and molecular characteristic information. (C) Contrastive learning: fusion of contrastive learning loss and cross-entropy loss from different representations of the same SMILES to train the model.
Abbreviation: SMILES: Simplified molecular input line entry system.

data were then input into the atomic and molecular encoders to extract features, with the two encoders not sharing weights.

Atomic encoding used the multi-head attention mechanism to project the query vector $Q$, the key vector $K$, and the value vector $V$ $h$ times to dimensions $d_q$, $d_k$, and $d_v$, respectively. The attention function was then executed in parallel to produce output values of dimension $d_v$. The final results were generated by concatenating these output values and projecting them again. The scaled dot-product attention was computed using the formula in Equation I:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{(I)}$$

Multi-head attention combined multiple scaled dot-product attention operations, as defined by Equations II and III:

$$Atom\ (H) = Concant\ (head_1,\ldots,\ head_h)W^p \quad \text{(II)}$$

$$head_i = Attention\ (HW_1^O, HW_1^K, HW_1^V) \quad \text{(III)}$$

Where H is an input representation or the hidden state in the encoder.

The method for extracting features from the molecular encoder is illustrated in Figure 1A, while the molecular maximum dimension reduction algorithm is outlined in Algorithm 1. Following Algorithm 1, feature fragments were embedded within the molecules of each chemical reaction system. Dimensional reduction was then conducted using a maximum aggregation operation alongside conditional structural parsing, enabling directional extraction and enhanced characterization of key features. The resulting point numbers and special symbols were retained to construct a molecular embedding representation. Finally, the molecular embeddings and the new filling matrix were fed into the encoder within the multi-head attention mechanism.

Specifically, the input to the model consisted of a SMILES sequence, in which different reactants were separated by a period (.) and reactants were separated from reagents using a greater-than symbol (>). For example, if the SMILES input is "CC.CCO>NCO</s>," Algorithm 1 first maps this SMILES sequence to an embedding matrix of size (1 × 11 × 256). It then identifies the positions of specific markers within the sequence: The fill position is at 1, the "." is at 2, the ">" position is at 6, and "</s>" is at 10. Using these position numbers, the embedding matrices for each reactant and reagent were obtained, resulting in matrices of sizes (2 × 256) for the first reactant, (3 × 256) for the second reactant, and (3 × 256) for the reagent. The maximum values of these features were reduced to a matrix of size (1 × 256), which encapsulated the important molecular information. Finally, the reactants, reagents, and characteristic symbols were combined into a new size embedding matrix of (6 × 256). This approach differs from traditional pooling-based methods through its integration of conditional structural parsing and a residual maximization mechanism.

### 3.2. Multilevel feature fusion module

The outputs from the molecular and atomic encoders were further processed through context-gating mechanisms.[33]

**Algorithm 1: Molecular maximum dimension reduction algorithm**

Input: Dataset *Z*, dictionary corresponding to the source dataset *n*.

Output: The embedding matrix of the final molecular-atomic combination *U'*

```
src_dict=n;
for i=0 to Z - 1 do
  Extract dot_p and sep_p according to src_dict; // dot_p represents dots, and sep_p stands for a greater-than symbol.
  Extract last_src, pad_p; // last_src is the final value in the list, and pad_p indicates whether the sentence has padding.
  if sep_p.numel() > 0 then
    if dot_p.numel() > 0 then
      v=dot_p [-1];
      for j, dot in enumerate (dot_p) do
        if dot < sep_p and dot==v then
          The embedding between dot and sep_p and sep_p and last_src is maximum;
        else if start < sep_p < dot then
          sep_p is maximum before and after embedding;
        else
          Handle embedding between start and dot;
        end
        start=dot + 1;
        Residual embedding is maximum;
      end
    else
      Handles embedding between > and last_src;
    end
  else
    if dot_p.numel() > 0 then
      for each dot in dot_p do
        Handles embedding between start and dot;
      end
      Handles embedding between start and last_src;
    End
    Finalize and store U' for the sample;
End
```

This enabled the dynamic modulation of atomic-level feature representations $y_a$ and molecular-level context embeddings $f_a$, facilitating fine-grained interactions between hierarchical feature spaces. The design of the multilevel feature fusion module is illustrated in Figure 1B, with the accompanying formulas provided in Equations IV-VI:

$$L = W \times Concant\ (y_a, f_a) \quad \text{(IV)}$$

$$\lambda_a = softmax\ (L, din = -1) \quad (V)$$

$$\bar{y}_{a=}\lambda_a y_a + (1-\lambda_a) f_a \quad (IV)$$

Where W represents the parameter matrix and $y_a$ denotes the final hidden representation of the fusion of the two encoders.

### 3.3. Fused loss functions module

The fused loss functions module introduces a novel combination of cross-entropy loss[34] and contrastive loss.

For the cross-entropy component, let the target labels be represented as $y = [y_1, y_2,\ldots, y_N]$, and the probability distribution predicted by the model as $Y = [Y_1, Y_2,\ldots, Y_N]$. Given that the predicted labels in the SMILES sequence span multiple categories, the multi-category cross-entropy loss was adopted. The loss is calculated using the following formula in Equations VII and VIII:

$$\bar{Y} = \ln(\sigma(Y,a)) \quad (VII)$$

$$L_{con} = -\sum_{i=1}^{N} y_i^{bi} \log(\bar{Y}^{bi}) \quad (VIII)$$

Where $b$ is the target class of sample, $\bar{Y}$ denotes the predicted probability of the processed tensor from the i-th sample on the target label $y_i$, and $N$ represents the number of samples in a batch.

Contrastive loss learning focuses on enhancing consistency among positive samples while reducing similarity among negative samples in the representation space. The contrastive learning approach used in this study is illustrated in Figure 1C. When the model's data loader retains the original samples, the sample size doubles using the molecular order exchange enhancement method. In this process, a positive sample pair consists of one standard and one enhancement sequence, both of which are labeled accordingly.

In addition, the sequence tensor comprises three dimensions: Batch size, sample length, and feature dimension. Before conducting contrastive learning, batch flattening was applied to merge the sample length and feature dimension into a single dimension while preserving the integrity of the batch dimension. The relevant formulas are given in Equations IX and X:

$$\cos(\tau_a, \tau_b) = \frac{\tau_a \cdot \tau_b}{\tau_a \tau_b} \quad (IX)$$

$$L_{contrast} = -\frac{1}{N}\sum_{n=1}^{N} \log \frac{exp(\cos(\tau_{n,c}, \tau_{n,d}))/\zeta}{exp(cos(\tau_{n,c}, \tau_{n,d}))/\zeta + \sum_{m \epsilon D_n} exp(\cos(\tau_{n,c}, \tau_{n,m}))/\zeta} \quad (X)$$

Where:

(i) cosrepresents the cosine similarity function.
(ii) $\tau_{n,c}$ represents the embedding of the canonical SMILES for molecule $n$.
(iii) $\tau_{n,d}$ refers to the SMILES sequence enhanced from the canonical SMILES.
(iv) $\tau_{n,m}$ represents the embedding of the negative samples.
(v) $D_n$ represents the set of molecules in the batch, excluding the canonical SMILES.

Figure 2 shows that a temperature parameter (ζ) of 0.05 yields the best product prediction results. Therefore, ζ was fixed at 0.05 for all experiments.

For the overall loss function, after introducing the two individual loss functions, the fused loss function $L_z$ is defined as in Equations XI and XII:

$$W_s = \frac{E_\alpha}{E_{total}} \quad (XI)$$

$$L_z = W_s L_{con} + (1-W_s) L_{contrast} \quad (XII)$$

The value of $E_a$ increases with the number of training epochs, while $E_{total}$ was fixed at 500. During both the early and late stages of training, the model adaptively adjusted the weights assigned to two loss functions. In the early phase, greater emphasis was placed on learning the fundamental features captured by the $L_{con}$ loss function. In the later phase, the focus shifted to fine-tuning the details of the $L_{contrast}$ loss function to improve the flexibility and effectiveness of the training process.

## 4. Experiments

### 4.1. Datasets

The datasets, as shown in Table 2, were derived from two different sources: The chemical journals with high impact factors (CJHIF) dataset, a large-scale compilation of reactions extracted from reports in high-impact chemical

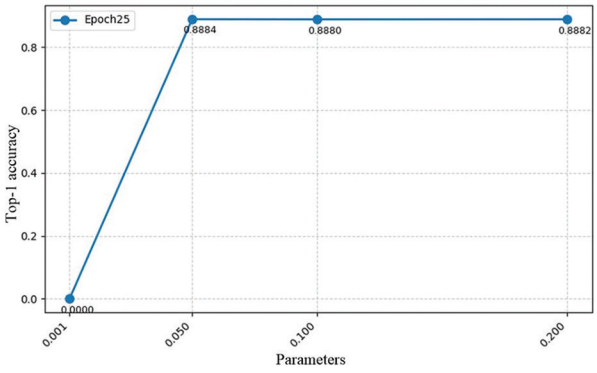



**Figure 2.** M2Echem prediction results under varying temperature parameter (ζ) at Epoch 25

**Table 2. Number of reactions across four datasets**

| Dataset | Training set | Validation set | Testing set | Total |
|---|---|---|---|---|
| USPTO-50k | 40,029 | 5,004 | 5,004 | 50,037 |
| USPTO-Jin | 409,035 | 30,000 | 40,000 | 479,035 |
| USPTO-Schwaller | 902,581 | 50,131 | 50,258 | 1,002,970 |
| CJHIF | 2,894,430 | 3,000 | 3,000 | 2,900,430 |

Abbreviations: CJHIF: Chemical journals with high impact factors; USPTO: United States Patent and Trademark Office.

journals,[35] and the United States Patent and Trademark Office (USPTO) dataset, a public collection of chemical reactions mined from United States patent grants spanning from 1976 to September 2016.[36]

(i) USPTO-50k: This dataset contains 50,037 reactions categorized into 10 different types.[37] Training, validation, and testing sets comprised 40,029, 5,004, and 5,004 reactions, respectively.

(ii) USPTO-Jin: Jin *et al.*[38] extracted data from the USPTO dataset and retained 479,035 reactions without stereochemical information or atom-mapping. They combined the reagents and reactants into a source string, delimiting them with a period (.).

(iii) USPTO-Schwaller: Schwaller *et al.*[8] removed duplicates and non-canonicalizable items, ultimately retaining 1,002,970 reactions that contained only a single product.

(iv) CJHIF: Since the reagent information in the raw CJHIF data was provided by name rather than in the SMILES representation, we utilized the PUG-View application programming interface to search for reagents in PubChem and obtained their SMILES expressions. For reagents that were not found, we excluded them. Subsequently, we converted the reactions in the CJHIF dataset to follow the reactants > reagents form. The target dataset contained the corresponding products. Finally, we retained 638,597 reactions that were normalizable and non-duplicated, and selected 3,000 reactions for each validation and test set.

### 4.2. Related parameter settings

M2Echem was developed using Python 3.7, RDKit version 2022.9.5 (Novartis, Basel), and Hugging Face Transformers version 4.10.2 (Thomas, America). The model architecture includes encoders and a decoder consisting of four identical layers, with a multi-head attention layer employing eight attention heads. The hidden dimension was set to 256, and the intermediate feed-forward layer size was set to 2,048. The Adam optimizer was used, and a beam search size of five was implemented. The batch size, total number of training epochs, and training equipment were configured to 8, 30, and A6000, respectively.

### 4.3. Evaluation metrics

Three commonly used evaluation metrics were considered: Bilingual evaluation understudy (BLEU),[39] accuracy (Top-ς), and *t*-test ($\left|\hat{T}\right|$). These are standard external methods for evaluating models.

The BLEU metric measures similarity between sentences by calculating the maximum matches of *n*-grams from the target sequence within the prediction sequence. The formula for *n*-grams is presented in Equation XIII:

$$T_n = \frac{\sum_i^E \sum_k^K min(M_k(c_i), max_{j\epsilon\omega} M_k(s_{i,j}))}{\sum_i^E \sum_k^K min(M_k(c_i))} \quad \text{(XIII)}$$

Where ω refers to the number of target sequences associated with the same chemical reaction text. *E* denotes the collection of target sequences in the dataset, *k* is the *k*-th phrase, $M_k(c_i)$ depicts the number of times the *k*-th phrase occurs in the prediction sequence, while $M_k(S_{i,j})$ represents its occurrences in the target sequence. We used a 4-gram BLEU calculation, and its formula is presented in Equation XIV:

$$BLEU = \frac{1}{N} \sum_{i=1}^N T_n \quad \text{(XIV)}$$

Top-ς accuracy gauges the percentage of predictions with the correct label among the Top-ς results. Higher accuracy generally corresponds to improved prediction performance. The calculation formula is presented in Equation XV:

$$Top-\varsigma = \frac{1}{N} \sum_{i=1}^N \pi(s_i, c_{i,s}) \quad \text{(XV)}$$

Where π (·) indicates a value of 1 if the target sequence matches the prediction sequence; otherwise, it indicates a value of 0. $c_{i,s}$ represents the topresults in the Top-ς results in the prediction results, with $\varsigma\epsilon[1, 2, 3, 4, 5]$.

$\left|\hat{T}\right|$ assesses whether the difference between the two group means is significantly greater than the random error. The greater the value, the more significant the difference. The formula is given in Equation XVI:

$$\left|\hat{T}\right| = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2(\frac{1}{n_1} + \frac{1}{n_2})}} \quad \text{(XVI)}$$

Where $\bar{X}_1$ and $\bar{X}_2$ represent the sample means, while $S^2$ denotes the variance.

### 4.4. Comparative analysis

Table 3 presents the Top-1, Top-3, and Top-5 accuracy rates, BLEU scores, $|\hat{T}|$ values, and *p*-values for the baseline and M2Echem models across four datasets, with the best results indicated by superscripted lowercase "a" ([a]). If the average values of the two groups are the same, the probability of obtaining the observed *p*-value was calculated. If this probability is <0.05, the difference is considered statistically significant and is marked with an asterisk (*). If this probability is <0.001, it is considered highly statistically significant and is marked with a double asterisk (**). "Baseline" refers to the T5chem model, which utilizes a single encoder-decoder architecture and optimizes using cross-entropy loss.

Based on the experimental results, M2Echem outperformed the baseline model by the following margins on four different datasets: It surpassed the baseline by 1.3, 1.3, 1.5, and 0.36 points across four metrics on the USPTO-50k dataset. These results indicate that the fusion loss mitigates small-sample overfitting. For the USPTO-Jin dataset, M2Echem outperformed the baseline model by 0.3, 0.1, and 0.07 points across three metrics. These findings suggest that the model enhances information complementarity and improves prediction performance by capturing atomic details, key molecular features, and implementing a fusion strategy. In addition, for the USPTO-Schwaller dataset, the proposed model surpassed the baseline model by 0.3, 0.3, 0.4, and 0.21 points across four metrics. Finally, for the CJHIF dataset, the proposed model exceeded the baseline by 4.1, 5.0, 5.0, and 4.1 points across four metrics. Both USPTO-Schwaller and CJHIF datasets provide stereochemical information. The improved experimental results demonstrate the effective use of stereochemical information in model design. Moreover, the independent sample $|\hat{T}|$ demonstrated that the prediction accuracy of the M2Echem model on the four datasets was significantly better than that of the baseline model, with $p<0.05$. This indicates a statistically significant difference, suggesting that our method achieved improved prediction accuracy.

To assess the effectiveness of the proposed model for product prediction, we compared it with several existing models, including S2S,[8] WLDN5,[22] Fairseq,[40] Molecular Transformer,[10] and T5chem. As shown in Table 4, the M2Echem model significantly outperformed these traditional models in both accuracy and BLEU scores on the USPTO-Jin and USPTO-Schwaller datasets. In addition, we conducted a *t*-test to compare the validation accuracy during the training process of the M2Echem model and several traditional models, including T5chem, Molecular Transformer, Fairseq, WLDN5, and S2S, resulting in *p*-values of 0.035010, 0.033011, 0.000091, 0.000042, and 0.000039, respectively. These *p*-values were well below the conventional significance threshold of 0.05, confirming that the accuracy improvement achieved by the M2Echem model is statistically significant.

### 4.5. Ablation experiments

To analyze the role of the molecular maximum dimension reduction algorithm and the fused loss method, ablation experiments were conducted on three datasets, and the results are presented in Table 5. The label "MAX" refers to the molecular maximum dimension reduction algorithm, while "NTL" represents the fused loss method.

Based on Table 5, the molecular maximum dimension reduction algorithm plays a crucial role in enabling the model

**Table 3. Baseline and M2Echem models' performances across four datasets**

| Dataset | Model | Accuracy (%) | | | BLEU | $\lvert\hat{T}\rvert$ | *p*-value | Time (h) |
|---|---|---|---|---|---|---|---|---|
| | | Top-1 | Top-3 | Top-5 | | | | |
| USPTO-50k | Baseline | 40.60 | 61.50 | 68.40 | 88.18 | 2.601 | 0.01093* | 3.40 |
| | M2Echem | 41.90[a] | 62.80[a] | 69.90[a] | 88.54[a] | | | 3.92 |
| USPTO-Jin | Baseline | 89.40 | 95.10 | 96.10 | 98.29 | 2.108 | 0.03501* | 5.31 |
| | M2Echem | 89.70[a] | 95.10[a] | 96.20[a] | 98.36[a] | | | 7.08 |
| USPTO-Schwaller | Baseline | 77.00 | 85.80 | 87.60 | 94.45 | 2.191 | 0.03452* | 35.27 |
| | M2Echem | 77.30[a] | 86.10[a] | 88.00[a] | 94.66[a] | | | 36.45 |
| CJHIF | Baseline | 56.20 | 69.10 | 73.40 | 87.81 | 4.503 | 0.00001** | 119.01 |
| | M2Echem | 60.30[a] | 74.10[a] | 78.40[a] | 90.22[a] | | | 120.00 |

Notes: Superscripted lowercase "a" ([a]) indicates the best results. An asterisk (*) represents a statistically significant difference at $p<0.05$, whereas a double asterisk (**) denotes a highly statistically significant difference at $p<0.001$.
Abbreviations: BLEU: Bilingual evaluation understudy; CJHIF: Chemical journals with high impact factors; USPTO: United States Patent and Trademark Office.

**Table 4. Comparison of the M2Echem model with several existing models**

| Model | USPTO-Jin | | | | | USPTO-Schwaller | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy (%) | | | BLEU | $p$ ($10^{-5}$) | Accuracy (%) | | | BLEU | $p$ ($10^{-5}$) |
| | Top-1 | Top-3 | Top-5 | | | Top-1 | Top-3 | Top-5 | | |
| S2S | 80.30 | 86.20 | 87.50 | - | 3.9** | 65.40 | 74.10 | - | - | 27.1** |
| WLDN5 | 80.60 | - | 93.40 | - | 4.2** | - | - | - | - | - |
| Fairseq | 82.42 | 89.85 | 90.74 | 90.74 | 9.1** | 69.69 | 77.33 | 78.92 | 92.50 | 52.3** |
| Molecular Transformer | 88.80 | 92.60 | 94.40 | 96.27 | 3301.1* | 76.17 | 82.86 | 83.69 | 85.17 | 3107.5* |
| T5chem | 89.40 | 95.10 | 96.10 | 98.29 | 3501.0* | 77.00 | 85.80 | 87.60 | 94.45 | 3452.0* |
| M2Echem | 89.70[a] | 95.10 | 96.20[a] | 98.36[a] | - | 77.30[a] | 86.10[a] | 88.00[a] | 94.66[a] | - |

Notes: Superscripted lowercase "a" ([a]) indicates the best results. An asterisk (*) represents a statistically significant difference at $p<0.05$, whereas a double asterisk (**) denotes a highly statistically significant difference at $p<0.001$.
Abbreviations: BLEU: Bilingual evaluation understudy; CJHIF: Chemical journals with high impact factors; USPTO: United States Patent and Trademark Office.

**Table 5. Results of the ablation experiments**

| Dataset | Molecular maximum dimension reduction algorithm method | Fused loss method | Accuracy (%) | | | BLEU |
|---|---|---|---|---|---|---|
| | | | Top-1 | Top-3 | Top-5 | |
| USPTO-Jin | / | / | 89.40 | 95.10 | 96.10 | 98.29 |
| | MAX | / | 89.60 | 95.10 | 96.10 | 98.30 |
| | MAX | NTL | 89.70 | 95.10 | 96.20 | 98.36 |
| USPTO-Schwaller | / | / | 77.00 | 85.80 | 87.60 | 94.45 |
| | MAX | / | 77.20 | 85.80 | 87.60 | 94.53 |
| | MAX | NTL | 77.30 | 86.10 | 88.00 | 94.66 |
| CJHIF | / | / | 56.20 | 69.10 | 73.40 | 87.81 |
| | MAX | / | 58.80 | 71.80 | 76.10 | 89.23 |
| | MAX | NTL | 60.30 | 74.10 | 74.10 | 90.22 |

Notes: "MAX" refers to the molecular maximum dimension reduction algorithm, while "NTL" represents the fused loss method.
Abbreviations: BLEU: Bilingual evaluation understudy; CJHIF: Chemical journals with high impact factors; USPTO: United States Patent and Trademark Office.

to learn key molecular features, thereby improving the accuracy of chemical product predictions. When this algorithm is removed, there is a noticeable decrease in evaluation metrics across all three datasets. In addition, the fused loss method contributed positively to training the M2Echem model. Removing this component negatively affects the evaluation metrics across the three datasets. The results of the ablation experiments demonstrate that the molecular maximum dimension reduction algorithm and the fused loss method together enhance the proposed model's performance.

### 4.6. Model understanding of different SMILES representations

A molecule represented by "C\C(COC1CCCCO1)=C/I," which was not included in the pre-training model, was selected to generate four different SMILES using non-standard SMILES data augmentation.[41,42] Figure 3 illustrates the generation of atom embeddings using both the baseline and M2Echem models, visualized through $t$-distributed stochastic neighbor embedding. The baseline model exhibited a broad and scattered distribution of atoms at Epoch 1 as the model had not yet been sufficiently trained. By Epoch 30, the baseline model gradually learnt to interpret the different SMILES representations of the same molecule. In contrast, the M2Echem model at Epoch 30 produced a more aggregated distribution of the different SMILES of the same molecule and was able to distinguish between iodine atoms in different chemical environments while maintaining their similarity. Therefore, the M2Echem model demonstrated superior capability in recognizing the same atomic labeling in different SMILES at Epochs 1 and 30, while the baseline showed weaker performance.

### 4.7. Model understanding of long SMILES embeddings

Figure 4 illustrates the embedding of Tanimoto coefficients[43] derived from three different SMILES

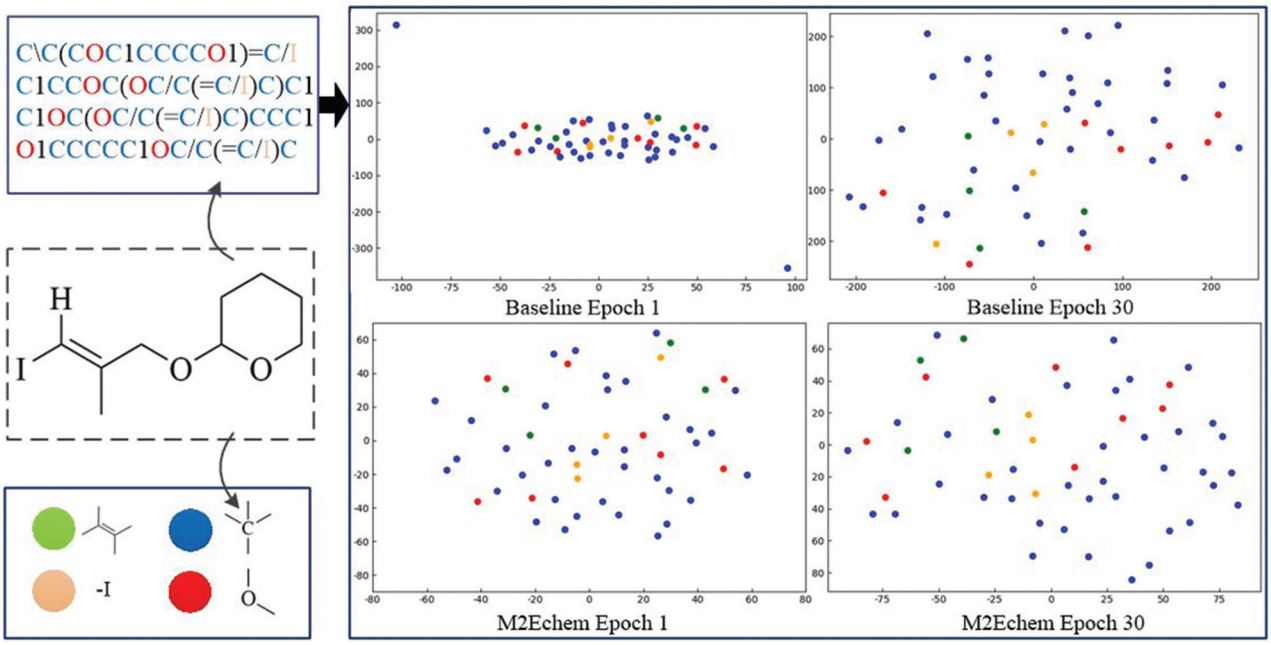



**Figure 3.** Visualization of atom embeddings for different SMILES formats using *t*-distributed stochastic neighbor embedding

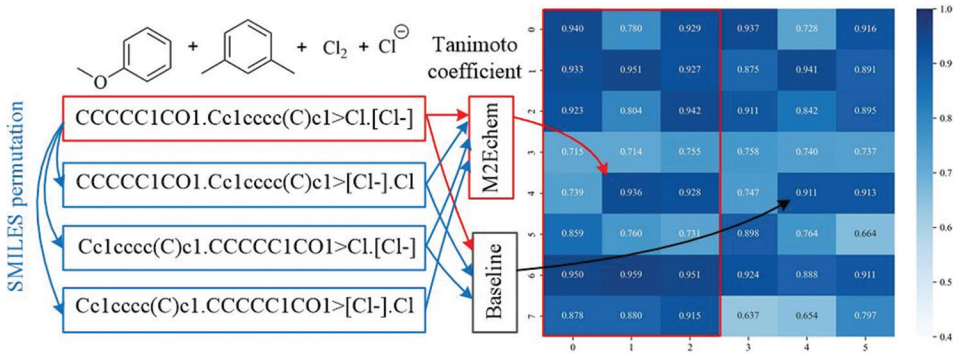



**Figure 4.** Tanimoto coefficient embeddings for reactants. The left panel (red box) displays the Tanimoto coefficients for M2Echem, while the right panel displays those for the baseline model.

embeddings of eight molecules that were not included in the pre-trained dataset, along with the corresponding canonical SMILES embeddings. The left panel (red box) displays the Tanimoto coefficients for M2Echem, while the right panel displays those for the baseline model. The Tanimoto coefficient approaches 1 as the SMILES embeddings increase in similarity. In Table 6, "Tanimoto coefficient 1" indicates the average value for different SMILES representations of the same molecule in the M2Echem model, while "Tanimoto coefficient 2" refers to the corresponding average in the baseline model. The graph and table illustrate the performance of both models at Epoch 30, and the vertical coordinates numbered 0–7 in Figure 4 correspond to the SMILES serial numbers 0–7 in Table 6. These findings suggest that, for SMILES sequences ranging in length from 32 to 90, the proposed model demonstrated greater embedding similarity compared to the baseline model.

To assess the model's ability to understand long sequences, we categorized three datasets according to sequence length: 0–55 as short sequences, 55–110 as medium sequences, and those longer than 110 as long sequences. Table 7 presents the accuracy rates and BLEU scores for these three sequence-length categories across the three datasets on different models. The molecular maximum dimension reduction algorithm demonstrated significant benefits for long sequences in the USPTO-Jin dataset, showing improvements of 3.15 points in BLEU

**Table 6. Sequence lengths and Tanimoto coefficients for eight simplified molecular input line entry system embeddings**

| SMILES | Sequence length | Tanimoto coefficient 1 | Tanimoto coefficient 2 |
|---|---|---|---|
| 0 | 32 | 0.883065[a] | 0.860441 |
| 1 | 86 | 0.937071[a] | 0.902324 |
| 2 | 47 | 0.889812[a] | 0.882767 |
| 3 | 50 | 0.728019 | 0.744914[a] |
| 4 | 59 | 0.867746[a] | 0.744914 |
| 5 | 90 | 0.783501[a] | 0.775470 |
| 6 | 85 | 0.953346[a] | 0.907687 |
| 7 | 57 | 0.890992[a] | 0.696146 |

Note: Superscripted lowercase "a" ([a]) indicates that the average embedding values for different SMILES of the same molecule exhibit high similarity.
Abbreviation: SMILE: Simplified molecular input line entry system embeddings.

score and 2.13 points in Top-1 accuracy compared to the baseline. The M2Echem model provided consistent performance across all three sequence lengths in the datasets. In particular, for the USPTO-Jin dataset, the Top-3 accuracy for long sequences increased by 2.13 points, while the BLEU score and accuracy for long sequences in the CJHIF dataset improved by 1.35, 2.94, 5.95, and 7.28 points, respectively, compared to the baseline.

### 4.8. Model performance in product prediction

Reaction types were randomly selected from the test set to predict organic chemical products using both the M2Echem and the baseline model. The results are illustrated in Figure 5, where red markings indicate changes in core atoms or atomic groups. The organic reactions in Figures 5A and 5B are both substitution reactions, and although the reactions are simple, the baseline model predicted incorrect products, thereby violating the principle of elemental conservation. The baseline model shown in Figure 5A predicted products with fewer oxygen atoms, while the model in Figure 5B predicted products with fewer carbon and fluorine atoms.

In addition, Figure 5C depicts a mono-disubstituted reduction reaction in which an NH group first undergoes acid-base neutralization with formic acid to form NOOCH, and then NOOCH is reduced to $NCH_3$. This transformation is correctly predicted by the M2Echem model, whereas the baseline model generates unreasonable products. In addition, Figure 5D illustrates a nucleophilic substitution reaction in which negatively charged carbon atoms attack positively charged carbon atoms. This interaction disrupts the C-Cu coordination bond and the C-Br bond, leading to the formation of the final product.

However, the predictions made by the baseline model were incorrect. Figure 5E shows a displacement reaction. The product predicted by the M2Echem model was close to the correct answer, whereas the baseline model fails to account for the metal sodium atom. These findings suggest that the model presented in this study effectively predicts the outcomes of fundamental reaction types, including displacement and substitution.

### 4.9. Attention weight

Two SMILES sequences were randomly selected from the USPTO-Jin test set. Figure 6 illustrates the relationship between attention weights for reactant-product mappings under both the baseline model and the M2Echem model for the two selected SMILES sequences.

In the attention matrix, the horizontal axis represents the reactants and reagents, while the vertical axis represents the products. Figure 6A presents the attention maps for both the baseline and M2Echem models during a reaction in which the aldehyde group "-COH" is reduced to an alcohol. In comparison to the baseline model, the M2Echem model demonstrated a more continuous and concentrated focus within the blue box. This suggests that the M2Echem model can accurately and consistently capture the mapping relationship between reactants and products, and that there are fewer noise points in the non-core areas. Figure 6B illustrates an esterification reaction. The M2Echem model better identified the esterification process involving acid dehydroxylation and alcohol dehydrogenation within the first blue box.

### 4.10. Computational efficiency of the model

The time and space complexity analyses of the baseline and M2Echem models is as follows, where the longest sequence length is L, the number of heads in multi-head attention is H, the dimension of the hidden layer is $d_{hid}$, the dimension of the feedforward layer is $d_{ff}$, the batch size is B, the number of layers is $L_{layer}$, the total number of samples is N, the average number of points in the sequence is P, and the model parameters are denoted as $\vartheta$.

The baseline model consists of a single encoder-decoder structure. The time complexity of the encoder primarily arises from the multi-head attention mechanism and the feedforward network, which have time complexities of $O(d_{hid}HL^2)$ and $O(d_{hid}d_{ff}L)$, respectively. Therefore, the overall time complexity of the encoder is $O(BL_{layer}d_{hid}HL^2)$, and that of the decoder is $O(L_{layer}L^3d_{hid}H)$. The total time and space complexities of the baseline model are $O(L_{layer}d_{hid}H(B/L+1)L^3)$ and $O(2\vartheta + BLd_{hid})$, respectively.

**Table 7. Accuracy rates and bilingual evaluation understudy scores across three sequence-length categories for different models on three datasets**

| Dataset | Sample size | Molecular maximum dimension reduction algorithm method | Fused loss method | Accuracy (%) | | | BLEU |
|---|---|---|---|---|---|---|---|
| | | | | Top-1 | Top-3 | Top-5 | |
| USPTO-Jin-pre | 31,225 | / | / | 88.43 | 94.59 | 95.67 | 98.02 |
| | | MAX | / | 88.62 94.56 95.73 98.02 | 94.56 | 95.73 [a]nd 98.02 | 98.02 |
| | | MAX | NTL | 88.62 | 94.60[a] | 95.83[a] | 98.12[a] |
| USPTO-Jin-mid | 8,728 | / | / | 92.99 | 96.77 | 97.49[a] | 99.26 |
| | | MAX | / | 92.98 | 96.84 | 97.39 | 99.23 |
| | | MAX | NTL | 93.27[a] | 96.91[a] | 97.48 | 99.27[a] |
| USPTO-Jin-last | 47 | / | / | 82.98 | 91.49 | 91.49 | 95.32 |
| | | MAX | / | 85.11[a] | 91.49 | 93.62 | 98.47 |
| | | MAX | NTL | 76.60 | 93.62[a] | 93.62[a] | 97.42[a] |
| USPTO-Schwaller-pre | 37,052 | / | / | 76.40 | 85.02 | 87.13 | 93.61 |
| | | MAX | / | 76.34 | 85.29 | 87.18 | 93.61 |
| | | MAX | NTL | 76.96[a] | 85.71[a] | 87.75[a] | 93.93[a] |
| USPTO-Schwaller-mid | 12,845 | / | / | 78.74 | 87.24 | 89.16 | 96.89 |
| | | MAX | / | 79.17[a] | 87.44 | 89.05 | 96.86 |
| | | MAX | NTL | 79.05 | 87.68[a] | 89.27[a] | 96.94[a] |
| USPTO-Schwaller-last | 350 | / | / | 60.94 | 72.85 | 75.35 | 93.70 |
| | | MAX | / | 61.11 | 73.71[a] | 75.56 | 93.86 |
| | | MAX | NTL | 61.43[a] | 72.22 | 76.29[a] | 94.19[a] |
| CJHIF-pre | 2,025 | / | / | 57.63 | 70.67 | 75.11 | 87.31 |
| | | MAX | / | 59.51 | 72.94 | 77.58 | 88.96 |
| | | MAX | NTL | 61.33[a] | 75.37[a] | 79.62[a] | 89.84[a] |
| CJHIF-mid | 732 | / | / | 49.32 | 62.84 | 67.21 | 88.19 |
| | | MAX | / | 53.96 | 66.94 | 70.77 | 89.84 |
| | | MAX | NTL | 55.12[a] | 68.70[a] | 73.27[a] | 91.16[a] |
| CJHIF-last | 127 | / | / | 50.19 | 62.20 | 64.57 | 89.39 |
| | | MAX | / | 51.18 | 64.57 | 68.50 | 90.09 |
| | | MAX | NTL | 53.33[a] | 68.15[a] | 71.85[a] | 90.74[a] |

Notes: "MAX" refers to the molecular maximum dimension reduction algorithm, while "NTL" represents the fused loss method. Superscripted lowercase "a" ([a]) indicates that the three models yield the best results on the same dataset within the same sequence length category.
Abbreviations: BLEU: Bilingual evaluation understudy; CJHIF: Chemical journals with high impact factors; USPTO: United States Patent and Trademark Office.

In contrast, the M2Echem model introduced additional components, including a molecular encoder, an algorithm, multi-level feature fusion, and an optimization loss, which have time complexities of $O(2BL_{layer}d_{hid}HL^2)$, $O(2NP^2)$, $O(2Nd_{hid}B)$, and $O(2d_{hid}B)$, respectively. The space complexities for these added components are $O(2BLd_{hid})$, $O(NL)$, and $O(2Bd_{hid})$. As a result, the total time and space complexities for the M2Echem model are $O(2BL_{layer}d_{hid}HL^2 + 2NP^2 + 2NBd_{hid} + 2L_{layer}d_{hid}HL^3)$, and $O(4\vartheta + 2BLd_{hid} + NL + 2Bd_{hid})$, respectively.

In this study, both models utilized eight multi-head attention heads, a hidden dimension of 256, four layers, and a feedforward layer dimension of 2,048. Consequently, as shown in Table 8, the time and space complexities for the baseline model are $O(8192BL^2+8192L^3)$ and $O(2\vartheta + 256B)$, while the time and space complexities for the M2Echem model are $O(2NP^2 + 512NBP^2 + 16384BL^2 + 16384L^3)$ and $O(4\vartheta + 512BL + NL + 512B)$, respectively.

The time and space complexities of the M2Echem model were approximately twice those of the baseline model. Furthermore, training time records (Table 3) indicated that the training time of the M2Echem model is 1–3 h longer than that of the baseline model. Nevertheless, as shown in Figure 7, M2Echem achieved

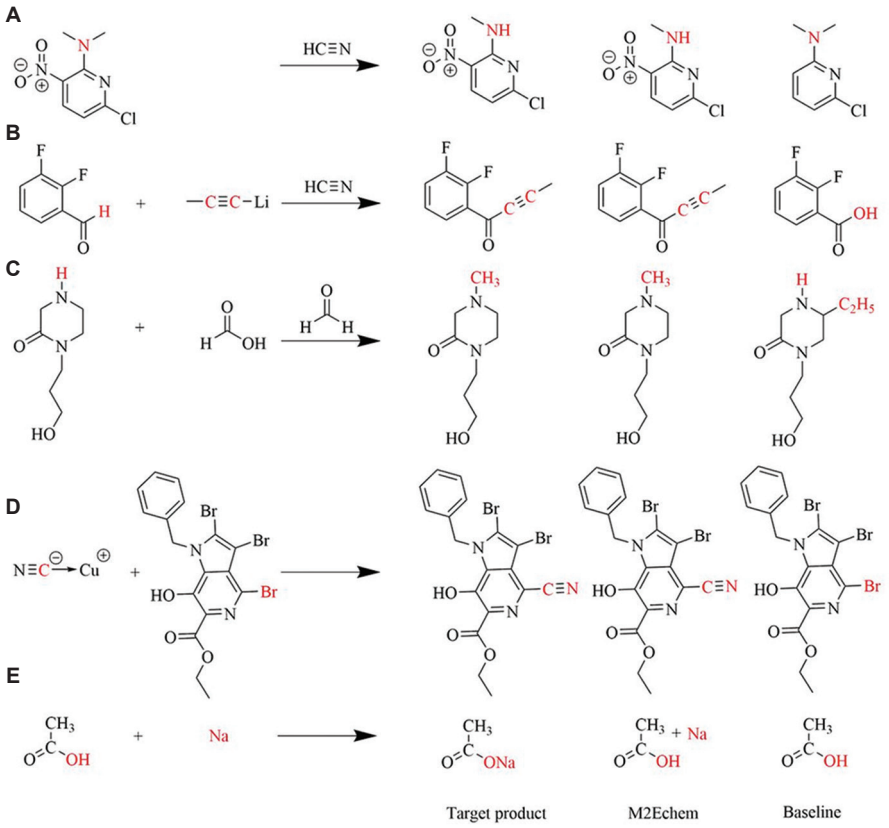



**Figure 5.** (A-E) Prediction results of the M2Echem model versus the baseline model across various reaction types

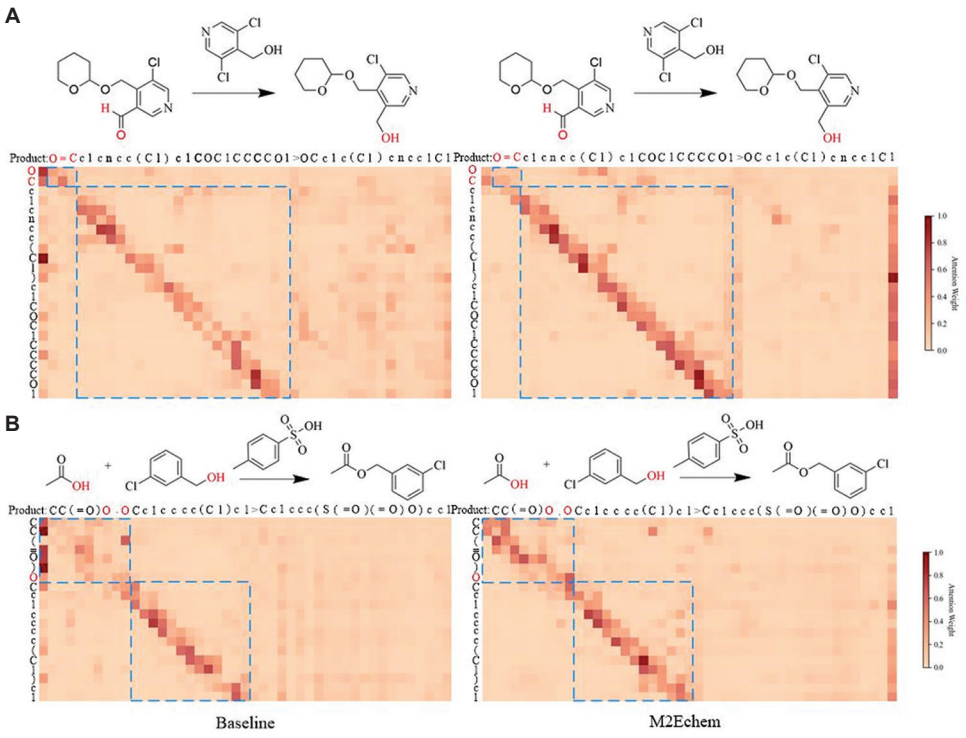



**Figure 6.** (A and B) Comparative analysis of attention weights for reactant-product interactions in chemical reactions between the baseline and M2Echem models

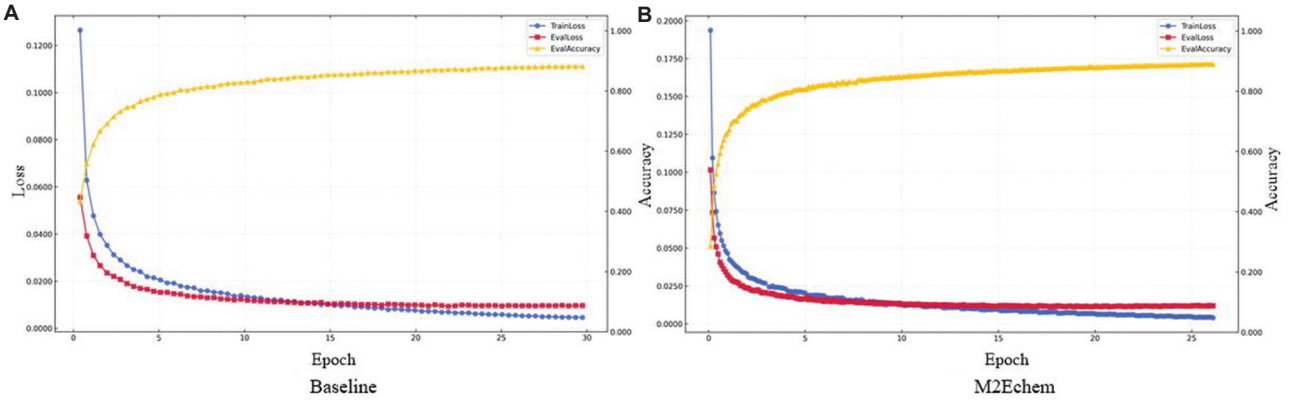



**Figure 7.** (A and B) Convergence curves for the baseline and M2Echem models

**Table 8. Time and space complexities of the baseline and M2Echem models**

| Model | Time complexity | Space complexity | Model parameters |
|---|---|---|---|
| Baseline | O ($8192BL^2$+$8192L^3$) O ($8192BL^2$+$8192L^3$) | O (2ϑ + 256BL) | 24.2 million |
| M2Echem | O ($2NP^2$+$512NBP^2$+ $16384BL^2$+$16384L^3$) | O (4ϑ + 512BL+ NL+512B) | 37.8 million |

faster convergence in product prediction. Although the time and space complexities of our model increased, the model extracted richer features and converged more rapidly during training, thereby improving prediction accuracy for product prediction.

## 5. Conclusion

The M2Echem model was developed to address complex organic chemistry forward prediction tasks, enhancing the performance of the T5chem model. This model utilizes molecular and atomic encoders to extract crucial feature information from atoms and molecules, enabling multi-level feature fusion that is subsequently fed into the decoder. The fusion loss component helps the model learn SMILES similarity more effectively. The findings demonstrate that the M2Echem model surpasses the baseline model across all four datasets. It extracts more comprehensive feature information, improves generalization across different representations of the same molecular structure, enhances the encoding of long sequence features, and contributes to artificial intelligence-driven drug development and research. However, we also noted that the model's effectiveness declines when dealing with complex or rare reaction types. Future research should focus on exploring a broader range of feature extraction methods and on improving the model's generalization ability for complex or rare reaction types.

## Acknowledgments

None.

## Funding

This research was funded by the National Natural Science Foundation of China (62406153, 62471259, and 62371261), the General Program of the Natural Science Research of Higher Education of Jiangsu Province (23KJB520031), and the Postgraduate Research and Practice Innovation Program of Jiangsu Province (SJCX25_2007).

## Conflict of interest

Jiashuang Huang is the Youth Editorial Board Member of this journal, but was not in any way involved in the editorial and peer-review process conducted for this paper, directly or indirectly. Separately, other authors declared that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

## Author contributions

*Conceptualization:* Shu Jiang, Yifan Jiang
*Formal analysis:* Linxing Zhu
*Investigation:* Yifan Jiang, Jiashuang Huang
*Methodology:* Shu Jiang, Jing Wang, Linxing Zhu
*Writing–original draft:* Jing Wang
*Writing–review & editing:* All authors

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

The data are publicly accessible in an open repository. The USPTO-50k dataset can be found at https://yzhang.hpc.nyu.edu/T5Chem/, while the USPTO-Schwaller and USPTO–Jin datasets are available at https://ibm.ent.box.com/v/ReactionSeq2SeqDataset. However, the CJHIF dataset is proprietary and cannot be accessed due to commercial restrictions.

## References


1. Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses. *Science*. 1969;166(3902):178-192.

   doi: 10.1126/science.166.3902.178

2. Satoh H, Funatsu K. Further development of a reaction generator in the SOPHIA system for organic reaction prediction. Knowledge-guided addition of suitable atoms and/or atomic groups to product skeleton. *J Chem Inform Comput Sci*. 1996;36(2):173-184.

   doi: 10.1021/ci950058a

3. Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF. Prediction of organic reaction outcomes using machine learning. *ACS Cent Sci*. 2017;3(5):434-443.

   doi: 10.1021/acscentsci.7b00064

4. Duvenaud DK, Maclaurin D, Iparraguirre J, *et al*. Convolutional networks on graphs for learning molecular fingerprints. In: *Advances in Neural Information Processing Systems 28*. California: Curran Associates, Inc.; 2015. p. 2224-2232.

5. Raccuglia P, Elbert KC, Adler PD, *et al*. Machine-learning-assisted materials discovery using failed experiments. *Nature*. 2016;533(7601):73-76.

   doi: 10.1038/nature17439

6. Segler MH, Waller MP. Modelling chemical reasoning to predict and invent reactions. *Chemistry*. 2017;23(25):6118-6128.

   doi: 10.1002/chem.201604556

7. Weininger DJ. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Infrom Comput Sci*. 1988;28(1):31-36.

   doi: 10.1021/ci00057a005

8. Schwaller P, Gaudin T, Lanyi D, Bekas C, Laino TJ. "Found in translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem Sci*. 2018;9(28):6091-6098.

   doi: 10.1039/c8sc02339e

9. Vaswani A, Shazeer N, Parmar N, *et al*. *Attention is All you Need*. Vol. 30. United States: Cornell University; 2017.

10. Schwaller P, Laino T, Gaudin T, *et al*. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent Sci*. 2019;5(9):1572-1583.

    doi: 10.1021/acscentsci.9b00576

11. Tang G, Müller M, Rios A, Sennrich RJ. *Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures*. Pennsylvania: Association for Computational Linguistics; 2018.

12. Wu F, Fan A, Baevski A, Dauphin YN, Auli MJ. *Pay Less Attention with Lightweight and Dynamic Convolutions*. United States: Cornell University; 2019.

13. Schwaller P, Probst D, Vaucher AC, *et al*. Mapping the space of chemical reactions using attention-based neural networks. *Nat Mach Intell*. 2021;3(2):144-152.

    doi: 10.1038/s42256-020-00284-w

14. Mellah Y, Kocaman V, Haq HU, Talby D. Efficient schema-less text-to-SQL conversion using large language models. *AIH*. 2024;1(2):96-106.

    doi: 10.36922/aih.2661

15. Mumtaz U, Ahmed A, Mumtaz S. LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties. *AIH*. 2024;1(2):16-28.

    doi: 10.36922/aih.2558

16. Bran AM, Schwaller P. Transformers and large language models for chemistry and drug discovery. In: *Drug Development Supported by Informatics*. Berlin: Springer; 2024. p. 143-163.

17. Leon M, Perezhohin Y, Peres F, Popovič A, Castelli M. Comparing SMILES and SELFIES tokenization for enhanced chemical language modeling. *Sci Rep*. 2024;14(1):25016.

    doi: 10.1038/s41598-024-76440-8

18. Xiong J, Zhang W, Wang Y, *et al*. Bridging chemistry and artificial intelligence by a reaction description language. *Nat Mach Intell*. 2025;7(5):782-793.

    doi: 10.1038/s42256-025-01032-8

19. Lo A, Pollice R, Nigam A, White AD, Krenn M, Aspuru-Guzik AJ. Recent advances in the self-referencing embedded strings (SELFIES) library. *Dig Discov*. 2023;2(4):897-908.

    doi: 10.1039/D3DD00044C

20. Ucak UV, Ashyrmamatov I, Lee J. Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization. *J Cheminform*. 2023;15(1):55.

    doi: 10.1186/s13321-023-00725-9

21. Wu Z, Jiang D, Wang J, *et al*. Knowledge-based BERT: A method to extract molecular features like computational chemists. *Brief Bioinform*. 2022;23(3):bbac131.

    doi: 10.1093/bib/bbac131

22. Chen S, Jung Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*.

2021;1(10):1612-1620.

doi: 10.1021/jacsau.1c00246

23. Liu Z, Zhang W, Xia Y, *et al*. *Molxpt: Wrapping Molecules with Text for Generative Pre-Training*. United States: Association for Computational Linguistics; 2023.

24. Lu J, Zhang Y. Unified deep learning model for multitask reaction predictions with explanation. *J Chem Inform Model*. 2022;62(6):1376-1387.

doi: 10.1021/acs.jcim.1c01467

25. Guo W, Wang J, Wang S. Deep multimodal representation learning: A survey. *IEEE Access*. 2019;7:63373-63394.

doi: 10.1109/ACCESS.2019.2916887

26. Ma S, Zhang D, Zhou M. *A Simple and Effective Unified Encoder for Document-Level Machine Translation*. United States: Association for Computational Linguistics; 2020. p. 3505-3511.

27. Zhang X, Li P, Li H. *AMBERT: A Pre-Trained Language Model with Multi-Grained Tokenization*. United States: Cornell University; 2020.

28. Zhu J, Xia Y, Wu L, *et al*. *Incorporating Bert into Neural Machine Translation*. Cornell University; 2020.

29. Jin D, Jin Z, Zhou JT, Szolovits P. *Is Bert Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment*. United States: Cornell University; 2020. p. 8018-8025.

30. Singh S, Shingatgeri V, Srivastava P. Revolutionizing new drug discovery: Harnessing AI and machine learning to overcome traditional challenges and accelerate targeted therapies. *AIH*. 2024;2(2):29-40.

doi: 10.36922/aih.4423

31. Gao T, Yao X, Chen D. *Simcse: Simple Contrastive Learning of Sentence Embeddings*. United States: Cornell University; 2021.

32. Chen X, Alamro H, Li M, *et al*. Target-Aware Abstractive Related Work Generation with Contrastive Learning. In: SIGIR '22: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*; 2022. p. 373-383.

33. Miculicich L, Ram D, Pappas N, Henderson J. *Document-Level Neural Machine Translation with Hierarchical Attention Networks*. Belgium: Association for Computational Linguistics; 2018.

34. Mao A, Mohri M, Zhong Y. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. In: *Proceedings of Machine Learning Research PMLR*; 2023. p. 23803-23828.

35. Jiang S, Zhang Z, Zhao H, *et al*. When SMILES smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. *IEEE Access*. 2021;9:85071-85083.

doi: 10.1109/ACCESS.2021.3083838

36. Lowe DJ. *Chemical Reactions from US Patents (1976-Sep2016)*; 2017.

37. Liu B, Ramsundar B, Kawthekar P, *et al*. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent Sci*. 2017;3(10):1103-1113.

doi: 10.1021/acscentsci.7b00303

38. Jin W, Coley C, Barzilay R, Jaakkola TJ. *Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network*. Vol. 30. United States: Cornell University; 2017.

39. Papineni K, Roukos S, Ward T, Zhu WJ. *Bleu: A Method for Automatic Evaluation of Machine Translation*. USA: Association for Computational Linguistics; 2002. p. 311-318.

40. Wang T. *Research on Chemical Reaction Prediction Model Based on Fairseq*. United States: IEEE; 2021. p. 167-171.

41. Bjerrum EJ. *SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules*. United States: Cornell University; 2017.

42. Tetko IV, Karpov P, Van Deursen R, Godin G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat Commun*. 2020;11(1):5575.

doi: 10.1038/s41467-020-19266-y

43. Khalifa AA, Haranczyk M, Holliday J. Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. *J Chem Inf Model*. 2009;49(5):1193-1201.

doi: 10.1021/ci8004644

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

# EpilepsyLLM: Fine-tuning large language models for Japanese epilepsy knowledge representation

**Xuyang Zhao[1,2,3], Qibin Zhao[4], and Toshihisa Tanaka[5]***

[1]Medical Science Data-driven Mathematics Team, RIKEN Center for Interdisciplinary Theoretical and Mathematical Sciences, Yokohama, Kanagawa, Japan

[2]Medical Data Mathematical Reasoning Special Team, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan

[3]Department of Artificial Intelligence Medicine, Chiba University, Chiba, Japan

[4]Tensor Learning Team, RIKEN Center for Advanced Intelligence Project, Chuo, Tokyo, Japan

[5]Department of Electrical Engineering and Computer Science, Tokyo University of Agriculture and Technology, Koganei, Tokyo, Japan

## Abstract

With massive training data and sufficient computing resources, large language models (LLMs) have demonstrated impressive capabilities. These models can rapidly respond to questions in almost all domains and are capable of retrieving, synthesizing, and summarizing information. The capabilities demonstrated by LLMs can enhance our livelihood and foster innovation. Nonetheless, in some professional domains, the focus is not only on response speed but also on higher requirements for response reliability. For example, in the medical domain, the reliability of information provided by the model poses a great risk to subsequent diagnosis and treatment, especially when the language is not English. In specific domains, domain-specific knowledge can be used to refine pre-trained LLMs to improve their performance in specific tasks. In this study, we aimed to build an LLM for epilepsy, called EpilepsyLLM. We constructed an epilepsy knowledge dataset in Japanese for LLM fine-tuning, and the dataset contained basic information on epilepsy, common treatment methods and drugs, and important notes on patients' lives. Using the constructed dataset, we refined several different pre-trained models with supervised learning. In the evaluation process, we applied multiple metrics to measure the reliability of the LLMs' output. The experimental results highlighted that the fine-tuned EpilepsyLLM can provide more reliable and specialized epilepsy responses.

***Keywords:*** Epilepsy; Large language models; Domain-specific; Fine-tuning

***Corresponding author:**
Toshihisa Tanaka
(tanakat@cc.tuat.ac.jp)

**Citation:** Zhao X, Zhao Q, Tanaka T. EpilepsyLLM: Fine-tuning large language models for Japanese epilepsy knowledge representation. *Artif Intell Health*. 2026;3(1):104-115. doi: 10.36922/AIH025180042

**Received:** May 3, 2025

**Revised:** August 5, 2025

**Accepted:** August 14, 2025

**Published online:** September 8, 2025



## 1. Introduction

In recent years, large language models (LLMs) have demonstrated remarkable advances across a broad range of natural language processing (NLP) tasks. Their ability to understand instructions, generate human-like responses, and generalize across domains has significantly transformed the landscape of artificial intelligence (AI). These models have consistently set new benchmarks in widely recognized NLP evaluation tasks, showcasing their superior reasoning, comprehension, and generative capabilities. A majority of high-performing LLMs are built upon the transformer architecture.[1] This

design enables efficient parallel processing and the capture of long-range dependencies in text. The success of LLMs is attributed not only to architectural innovations but also to the availability of massive training corpora and extensive computational resources, enabling these models to scale to hundreds of billions of parameters. One of the most influential series in this domain is OpenAI's Generative Pre-trained Transformer (GPT) line, which includes GPT-1,[2] a foundational model that explored unsupervised learning on large-scale corpora; GPT-2,[3] known for its fluency and coherence; and GPT-3,[4] which introduced few-shot and zero-shot learning capabilities with its 175 billion parameters. InstructGPT[5] further refined the inherent capabilities by incorporating reinforcement learning from human feedback, allowing it to align better with human instructions. The most recent iteration, GPT-4,[6] pushes the boundaries further with enhanced multimodal capabilities and improved alignment with human preferences. However, the architectural details, training datasets, and optimization strategies for these models remain proprietary, limiting transparency and reproducibility in the academic community. In contrast, Meta's Large Language Model Meta AI (LLaMA)[7] adopts a more open approach. Released specifically for research purposes, the LLaMA family includes models with 7B, 13B, 30B, and 65B parameters. Despite having fewer parameters than some commercial models, LLaMA exhibits strong performance across standard benchmarks. For example, LLaMA-13B surpasses GPT-3 on several tasks, while being significantly more parameter-efficient. The largest model, LLaMA-65B, achieves performance on par with other models, such as Chinchilla[8] and PaLM-540B,[9] both of which are far larger in size. Advancements in LLMs demonstrated that with efficient training, massive amounts of data can yield higher performance. The growing availability of LLMs has opened new possibilities for real-world applications, including education, healthcare, creative writing, and human–computer interaction. The rapid progress and accessibility of LLMs continue to inspire innovations, suggesting a transformative impact on society and daily life.

LLMs have found increasingly diverse applications in the medical field, showcasing their ability to assist in a wide range of healthcare-related tasks. Among the most prominent applications are medical licensing examination,[10-12] diagnostic support,[13,14] patient communication,[15,16] and medical education,[17-19] and their impact extends well beyond these areas. However, alongside these promising developments, the growing reliance on LLMs also introduces a range of risks and ethical concerns that must be addressed.[20-22] One major issue is the risk of misinformation, as LLMs can generate plausible-sounding but factually incorrect responses, which could lead to misdiagnosis or inappropriate clinical decisions if not properly supervised. Furthermore, the black-box nature of most models makes it difficult to understand how conclusions are reached, raising issues of accountability and trust in clinical environments.

To further advance the performance and accessibility of LLMs, Stanford Alpaca[23,24] was introduced, demonstrating an efficient, scalable, and low-cost method for fine-tuning LLMs to follow instructions more effectively. The Alpaca is built upon the LLaMA-7B model, and through a carefully curated fine-tuning process, it achieved performance that is qualitatively comparable to GPT-3.5 (text-davinci-003). In addition, the Alpaca collects a small dataset of 175 human-written instruction-output pairs. These examples represent a diverse set of tasks designed to probe a model's ability to understand and carry out various instructions, ranging from summarization and translation to reasoning and creative writing. Rather than relying on massive manual annotation, Alpaca uses GPT-3.5 (text-davinci-003) to automatically generate additional instruction-following examples, scaling the dataset to 52,000 unique prompts and responses.

LLMs have demonstrated remarkable capabilities across a wide range of general NLP tasks, achieving state-of-the-art performance in areas such as question answering, summarization, translation, and dialogue generation. However, when these models are applied to highly specialized domains, such as medicine, law, and finance, their performance often diminishes. The reduced accuracy and reliability stem from the fact that general-purpose LLMs are typically trained on broad, heterogeneous corpora that lack the depth and nuance required for domain-specific expertise. To address this limitation, one effective method is domain-specific fine-tuning, where a general LLM is further trained or adapted using targeted corpora relevant to a particular professional field. In the medical domain, this approach has yielded significant improvements in performance across various medical NLP benchmarks and real-world clinical tasks. In the medical field, some of these medical-specific LLMs include PubMedBERT,[25] BioLinkBert,[26] BioMedLM,[27] BioGPT,[28] Med-PlaM,[29] ClinicalGPT,[30] PMC-LLaMA,[31] and ChatDoctor.[32] In addition, these LLMs typically incorporate various types of medical knowledge, including patient-physician dialogue transcripts, PubMed abstracts, full-text articles from PubMed Central (PMC), and deidentified electronic health records, to further enhance model understanding in clinical settings. These models demonstrate the effectiveness of incorporating structured and unstructured domain-specific data into the training pipeline. With the help of medical knowledge, fine-

tuned LLMs achieve better performance in medical tasks. For instance, in the United States Medical Licensing Examination (USMLE), MedPaLM 2 achieved the highest score of 86.5.

While existing medical LLMs have demonstrated impressive capabilities in handling a broad range of medical tasks, their success has largely been confined to English and generalized medical knowledge. These models focus on English-based medical reasoning, question answering, and literature comprehension, and they often fall short when applied to non-English languages or specialized subdomains of medicine. In this study, we present the EpilepsyLLM, a domain-specific LLM designed to focus exclusively on epilepsy and operate primarily in the Japanese language. Epilepsy is one of the most prevalent neurological disorders worldwide,[33] affecting millions of individuals across age groups. It manifests through a variety of seizure types, including tonic rigidity, myoclonic jerks, and atonic seizures, each of which can significantly impair a patient's quality of life.[33-36] The burden of epilepsy is particularly acute in pediatric patients, where the condition may hinder cognitive development, behavioral stability, and social integration.[37] Treatment for epilepsy generally begins with medication, which helps control seizures in many patients. However, a substantial subset of individuals suffer from refractory (drug-resistant) epilepsy, for which surgical intervention, such as resective surgery or neurostimulation, is considered. Despite medical or surgical treatment, epilepsy patients often live with numerous daily life restrictions, including avoiding seizure triggers, adhering to medication schedules, navigating driving limitations, and managing social stigma.[38] Given these challenges, EpilepsyLLM aims to serve as a specialized medical assistant capable of understanding, generating, and interpreting Japanese texts related to epilepsy. Its applications may span from assisting clinicians with diagnosis and treatment planning to supporting patients and caregivers in understanding disease management and improving communication. By concentrating on a single neurological disease and embracing linguistic diversity, EpilepsyLLM offers a promising direction in the field of medical AI, highlighting the potential for more precise, localized, and equitable healthcare solutions.

The application of LLMs to the medical field requires a high level of professionalism and domain accuracy. Unlike general-purpose use cases, medical applications demand high precision, factual correctness, and a nuanced understanding of clinical terminology and practices. Errors in model outputs can have serious consequences, which makes the incorporation of expert-level medical knowledge essential for safe and effective deployment. To address these needs in the context of epilepsy, we constructed a fine-tuning dataset composed of high-quality, domain-specific knowledge collected from publicly available Japanese resources on the internet. These sources are reliable epilepsy-focused content. The collected data were transformed into instruction-following demonstrations. As a base model, we employed two pre-trained models, LLaMA[7] and LLM-jp[39] (a Japanese language foundation model known for its linguistic alignment with Japanese text). LLM-jp provided a strong foundation for our study, given that its vocabulary, tokenizer, and pre-training corpus are optimized for the Japanese language, making it highly suitable for our targeted application. Our experiments demonstrated that the LLMs fine-tuned with our curated epilepsy-specific dataset significantly outperformed other baseline models. EpilepsyLLM presents more professional and reliable answers when faced with epilepsy knowledge. The experimental results also confirmed that by using more domain-specific knowledge to fine-tune the LLMs, the performance of the model in the particular domain can be significantly enhanced.

The proposed EpilepsyLLM holds potential for application across several clinically relevant areas. Clinical decision support is one promising area where the model could assist healthcare providers by synthesizing guidelines, summarizing treatment options, or generating initial responses to structured clinical queries. In the context of patient education, the system could help produce tailored, accessible explanations of medical information—particularly valuable for underserved or linguistically diverse populations. In addition, the model could serve in medical scribe and documentation assistance, helping clinicians to structure or translate free-text notes into more formal documentation based on high-level prompts or queries. Despite the model's potential, some issues should also be noted in its prospective clinical applications, such as misinformation, overreliance, and the lack of transparency, thereby warranting further research in these areas.

## 2. Methods

In this study, we leveraged domain-specific knowledge related to epilepsy to fine-tune LLMs to enhance their performance in epilepsy-focused medical applications. Moreover, we focused on optimizing the model's performance in Japanese, ensuring that the resulting language model can offer high-quality, linguistically and culturally tailored support for Japanese-speaking medical professionals, researchers, and patients. The overview of EpilepsyLLM is displayed in Figure 1.

### 2.1. Epilepsy knowledge dataset

To construct a high-quality dataset suitable for fine-tuning a domain-specific model, we systematically collected

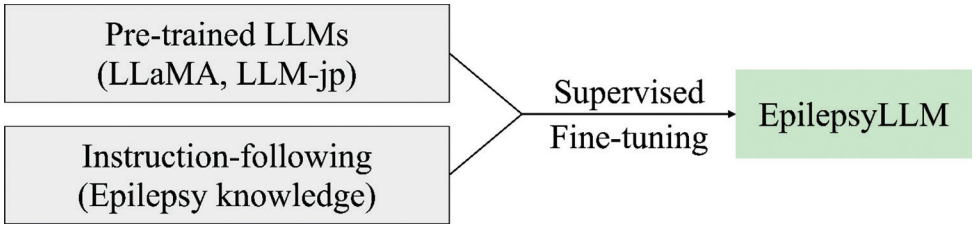



**Figure 1.** The overview of EpilepsyLLM
Abbreviations: LLMs: Large language models; LLaMA: Large Language Model Meta Artificial Intelligence.

relevant information from reputable public sources focused on epilepsy. These sources included official websites operated by medical associations, hospitals, and pharmaceutical companies. We aimed to ensure that the curated dataset is both reliable and medically accurate. This dataset served as a foundational resource for fine-tuning models tailored to epilepsy, enhancing their ability to understand, analyze, and generate clinically meaningful text. The specific websites used in the dataset collection process are listed below:

(i) Japan Epilepsy Association[1]: The association conducts research and nationwide campaigns aimed at promoting social understanding of epilepsy, providing social support for people suffering from epilepsy, and improving epilepsy policies.
(ii) Epilepsy Information Center[2]: The website belongs to the National Epilepsy Center, National Hospital Organization (NHO), Shizuoka Institute of Epilepsy and Neurological Disorders, which is the largest epilepsy center in Japan in terms of both patient volume and number of specialized clinicians.
(iii) Tenkan Net[3]: The website, developed by Alfresa Pharma Corporation, collects basic information about epilepsy, including basic diagnosis and treatment plans, routine examinations, commonly used drugs and their side effects, and lifestyle precautions.

After data collection, we conducted a systematic and uniform preprocessing and curation process to ensure data quality and ethical compliance. The curation pipeline involved several key steps. First, we standardized the formatting by normalizing whitespace, punctuation, and special characters to ensure consistency across the dataset. Next, we applied language-specific cleaning procedures to correct encoding issues and remove non-textual elements, such as HTML tags or formatting artifacts. To address privacy concerns and adhere to ethical research standards, we implemented a filtering step to identify and remove any personally identifiable information. This included names, physical addresses, phone numbers, email addresses, and other sensitive information that could potentially identify the individuals. The resulting dataset maintained the integrity of the original content while safeguarding user privacy and improving overall data usability for downstream NLP tasks.

Data collected from the three independent websites were divided into training and testing datasets after preprocessing. Content from the Japan Epilepsy Association and Epilepsy Information Center was used to construct the training dataset. For the testing dataset, we curated 24 representative instruction–response pairs exclusively from Tenkan Net, which were not included in the training dataset. This separation ensured an independent evaluation and allowed us to better assess the model's ability to generalize to unseen material within the same domain.

The dataset from Alpaca[23] was constructed with a focus on instruction-following demonstrations, where each data sample consisted of a prompt and a corresponding response designed to guide model behavior. Some representative examples from the dataset are displayed in Table 1. For fine-tuning the model, more than 200 carefully curated samples were utilized. These samples covered a diverse range of instructions and domains to ensure that the model could effectively generalize to various tasks. In addition, a separate subset comprising 24 distinct samples was selected and reserved exclusively for testing. This test set was used to evaluate the model's performance on unseen instruction–response pairs, providing a measure of its generalization capabilities after fine-tuning.

## 2.2. Pre-trained models

### *2.2.1. LLaMA*

The LLaMA[7] family of models, introduced by Meta AI, represents a series of foundational LLMs designed to advance research in NLP tasks. Unlike many commercially released models, LLaMA models are explicitly intended for academic and research purposes, and these models are made accessible to researchers under a non-commercial license. LLaMA models are trained on publicly available

[1] https://www.jea-net.jp/
[2] https://shizuokamind.hosp.go.jp/epilepsy-info/
[3] https://www.alfresa-pharma.co.jp/general/tenkan/faq/index.html

**Table 1. Examples of an epilepsy dataset**

| Question | Answer |
|---|---|
| てんかんはどんな病気ですか？(What is epilepsy?) | てんかん発作は大脳神経細胞の過剰な亢奮（てんかん発射）によって生じます。てんかん発作を繰り返し、慢性に経過する大脳の病気がてんかんです。ただ1 回だけの発作や高熱のときだけに起こる発作はてんかんとは区別されます。また、脳炎の急性期や頭部外傷の直後に起こったけいれん発作も、それだけであれば、てんかんとは区別されます。慢性に経過することがてんかんの特徴です。てんかん発作は大脳の機能が種々の形で表出されるので、発作症状はきわめて多彩です。しかし、ひとりの患者の発作症状はほぼ一定しています。なぜてんかんになるのか、なぜてんかん発作が起こるかは十分にはわかっていません。大脳の神経細胞ネットワークの亢奮と抑制の微妙なバランスの破綻に求める学説が有力です。てんかん発作を起こしやすい下 地（素因）と大脳の機能を不安定にする脳損傷（外因）とが複雑に絡み合ってんかんが発病すると推測されています。(Epileptic seizures are caused by excessive activation of cerebral nerve cells [epileptic discharge]. Epilepsy is a chronic disease of the cerebrum that causes repeated epileptic seizures. Seizures that occur only once or only during high fever are not considered epilepsy. Seizures that occur during the acute phase of encephalitis or immediately after head trauma are also not considered epilepsy. The chronicity of epilepsy is a characteristic of epilepsy. Since epileptic seizures express the functions of the cerebrum in various ways, the symptoms of the seizures are extremely diverse. However, the symptoms of seizures in a single patient are almost constant. It is not fully understood why epilepsy occurs. The prevailing theory is that it is due to the breakdown of the delicate balance between activation and inhibition of the cerebral nerve cell network. It is speculated that epilepsy develops when a complex combination of a predisposing factor [predisposing factor] that makes the patient susceptible to epileptic seizures and brain damage [external factor] that destabilizes the functions of the cerebrum is involved.) |
| てんかん発作が起こることで、脳にダメージを受けますか？(Do epileptic seizures cause brain damage?) | てんかん重積状態という特別な状態をのぞけば、通常、一度の発作で脳がダメージを受けることはありません。てんかん重積状態とは、30 分以上発作がとまらなかったり、ひとつの発作が終わった後、意識が完全に回復する前に次の発作が起こることを 30 分以上繰り返す状態を指し、注射など薬物で発作をとめる必要のある緊急事態です。ただし、長い経過を見ますと、記憶障害、知能低下や行動障害、精神医学的問題を合併することがありま す。これらは発作を繰り返した影響や、過ぎた薬物の服用、心理的要因、てんかんの原因となっているもと の病気の影響など種々の要因が関与しています。てんかんが難治に経過する場合は多種類の抗てんかん薬による治療はできるだけ避け、場合によってはてんかん外科手術を早めに検討したほうがいいことがあります。また、てんかんではありませんが、乳幼児期に熱性けいれんなどのけいれんが長時間持続すると、側頭葉 の内側にある海馬などの萎縮が起こり、それが数年から 10 年後に側頭葉てんかんの原因になることが知られています。(Except for a special condition called status epilepticus, a single seizure does not usually damage the brain. Status epilepticus refers to a state in which a seizure does not stop for more than 30 min, or a seizure occurs repeatedly for more than 30 min after one seizure has ended, before consciousness is fully restored. This is an emergency situation in which the seizures must be stopped with drugs such as injections. However, over the long term, memory disorders, intellectual disability, behavioral disorders, and psychiatric problems may occur. These are caused by various factors, such as the effects of repeated seizures, past drug use, psychological factors, and the effects of the underlying disease that causes epilepsy. If epilepsy is intractable, it is best to avoid treatment with multiple types of antiepileptic drugs as much as possible, and in some cases, it may be better to consider epilepsy surgery early. Although not classified as epilepsy, prolonged convulsions in infancy, such as febrile seizures, can lead to atrophy of the hippocampus and other inner temporal lobe structures, which is known to result in temporal lobe epilepsy several to 10 years later.) |
| 高齢で発病するてんかんの特徴を教えてください。(What are the characteristics of epilepsy that develop in older adults?) | てんかんの多くは小児期に発病します。2008 年2 月～ 11 月に当院（静岡てんかん・神経医療センター）でてんかんと診断された928 名のうち、751 名（80.9 %）が20 歳までに発病していました。20 歳を過ぎるとてんかんの発病率は次第に低下していきますが、60 歳を越えると再び増加に転ずるといわれています。脳血管障害など、高齢になって新たに生ずる脳の器質的な障害を背 景に、てんかんが発病しやすくなると考えられています。当院の統計でも高齢になるとわずかに発病率が増え、15 名（1.6 %）が60 歳以降に発病していました。推定病因は脳梗塞2、脳膿瘍2、脳腫瘍 1、脳外傷1、長年の大酒1 名で、他は不明でした。このように 器質性の病因が約半数に認められるのは、てんかん全体から見れば多い数字です。一方、熱性けいれんの既往や、てんかんまたは熱性けいれんの家族歴をもつ例は1 名もありませんでした。診断は15 名全例が症候性部分てんかんで、うち4 名は側頭葉に焦点を認めました。発作型は全般性強直間代発作が 10 、複雑部分発作が 7、単純部分発作が5、非けいれん性てんかん重積状態の疑いが2 名でした（重複あり）。発作頻度は年単位 5、月単位 7、週単位 2、日単位1 名でした。ほぼ全例が結婚してお子さんもあり、お仕事もしてこられた方々です。一般に、高齢発病のてんかんは薬物治療で比較的容易に発作がコントロールされることが多いといわれています。高齢者では血中濃度が上昇しやすいので、服用量を決定する際には注意が必要です。(Most epilepsy develops in childhood. Of the 928 people diagnosed with epilepsy at our hospital [Shizuoka Epilepsy and Neurological Medical Center] between February and November 2008, 751 [80.9%] had developed the disease by the age of 20. The incidence rate of epilepsy gradually decreases after the age of 20, but it is said to increase again after the age of 60. It is thought that epilepsy is more likely to develop against the background of organic brain disorders that occur in old age, such as cerebrovascular disorders. According to our hospital's statistics, the incidence rate increases slightly with age, with 15 people [1.6%] developing the disease after the age of 60. The suspected causes were cerebral infarction [two cases], brain abscess [two cases], brain tumor [one case], brain trauma [one case], and long-term heavy drinking [one case], and the causes of the remaining cases were unknown. As such, the fact that organic causes are found in about half of the cases is a high number compared to the overall number of cases of epilepsy. On the other hand, none of the patients had a history of febrile convulsions or a family history of epilepsy or febrile convulsions. All 15 patients were diagnosed with symptomatic partial epilepsy, of whom four had a focal seizure in the temporal lobe. The seizure types were generalized tonic-clonic seizures in 10 patients, complex partial seizures in seven patients, simple partial seizures in five patients, and suspected non-convulsive status epilepticus in two patients [with some overlapping]. The frequency of seizures was annual in five patients, monthly in seven patients, weekly in two patients, and daily in one patient. Almost all patients were married, had children, and were working. In general, it is said that seizures in elderly patients with epilepsy can be relatively easy to control with drug therapy. Drug use led to an increase in blood levels of the elderly, so caution is required when determining the dosage.) |

Note: Information is provided in Japanese along with its English translation.

datasets, emphasizing transparency and reproducibility in their development. They span a range of sizes, from 7 billion to 65 billion parameters, allowing researchers to study scaling laws and model behavior across different model capacities. Notably, LLaMA models achieve competitive or superior performance compared to larger proprietary models by optimizing data quality and training methodology rather than merely increasing model size. Due to their open-access nature and strong performance, LLaMA models have quickly become a foundation for further research into instruction-tuned models, alignment, domain adaptation, and low-resource language modeling. Several instruction-following models, such as Alpaca, are built by fine-tuning smaller LLaMA checkpoints with specially constructed instruction datasets.

#### *2.2.2. LLM-jp*

The LLM-jp model is an initiative focused on developing LLMs specifically optimized for the Japanese language. Recognizing that many widely used LLMs are predominantly trained in English and multilingual datasets with limited Japanese content, LLM-jp aims to address the gap by creating models that better capture the linguistic nuances and specific syntax of Japanese. LLM-jp models are typically trained on extensive corpora consisting of high-quality Japanese text sourced from books, news articles, web data, and other publicly available materials. The model is open access to promote research and development within the Japanese AI community. A distinguishing feature of LLM-jp is its careful curation of datasets to ensure broad coverage across different domains while maintaining linguistic richness and correctness. In addition, instruction-tuning and fine-tuning on Japanese-specific tasks are integral to the project, making the models more suitable for downstream applications, such as summarization, translation, question answering, and dialogue generation in Japanese. LLM-jp plays an essential role in enabling the creation of high-quality, culturally aligned AI systems for Japanese users and contributes to the broader goal of linguistic diversity and inclusivity in AI development.

### 2.3. LLM fine-tuning

Fine-tuning an LLM model involves adapting a general pre-trained model to work better on specific tasks or domains by continuing the training process with new data. In this study, two open-source models, LLaMA and LLM-jp, were used as base models. The training data used in LLaMA and LLM-jp are presented in Table 2. In the LLaMA training, most of the training data comes from English. In contrast, Alpaca was used as the fine-tuning dataset, and a Japanese version of Alpaca was also used for fine-tuning to improve the model's Japanese capabilities. LLM-jp training included datasets with more Japanese corpus, and three different Japanese fine-tuning datasets were used for model fine-tuning.

**Table 2. Training and fine-tuning datasets**

| Dataset | Content | Language |
|---|---|---|
| LLaMA training | Common Crawl | English |
| | C4 | English |
| | GitHub | English |
| | Wikipedia | 20 languages |
| | Books | English |
| | arXiv | English |
| | Stack Exchange | English |
| Alpaca | Instruction-following | English |
| Alpaca (Japanese) | Instruction-following | Japanese (translated from Alpaca) |
| LLM-jp training | mC4 | Japanese |
| | Wikipedia | Japanese |
| | Pile | English |
| | Wikipedia | English |
| | Stack (code) | English |
| Jaster | Instruction-following | Japanese |
| Dolly (Japanese) | Instruction-following | Japanese (translated from Dolly) |
| OASST (Japanese) | Instruction-following | Japanese (translated from OASST) |
| Epilepsy dataset | Epilepsy knowledge | Japanese |

Abbreviation: OASST: OpenAssistant Conversations Dataset Jaster: j + asterisk.

## 3. Results

Herein, we present a comprehensive evaluation of the proposed fine-tuning method through a series of controlled experiments. The primary objective was to assess the impact of domain- and language-specific fine-tuning on the performance of LLMs in epilepsy-related tasks. Since we had limited computing resources (four 80 GB A100s), for LLaMA experiments, LLaMA (7B) was used as the base model for fine-tuning, while larger models LLaMA (13B) and LLaMA (30B) were used directly for inference. To achieve general-purpose instruction-following capabilities, we fine-tuned LLaMA-7B using the Alpaca dataset, a widely used synthetic instruction-tuning dataset based on OpenAI's text-davinci-003. In addition, our main focus was to fine-tune the model using our Japanese language epilepsy dataset, which contained domain-specific medical knowledge tailored to epilepsy. Since the epilepsy dataset is collected in Japanese, to verify the impact of language

on LLM performance, a translated version of the Alpaca dataset was also used for fine-tuning. The English-Alpaca dataset was translated using ChatGPT, while the Japanese-Alpaca dataset was obtained from Github[4]. Double fine-tuning was also conducted: First using the Alpaca or Japanese Alpaca dataset, followed by further fine-tuning using the epilepsy dataset.

For the LLM-jp experiments, we conducted both fine-tuning and inference evaluations using models from the LLM-jp family, which are pre-trained primarily on Japanese language data and optimized for Japanese language tasks. We selected the LLM-jp (1.3B) model as the base model for fine-tuning due to its manageable size and compatibility with our computational constraints. This model was fine-tuned using our curated Japanese epilepsy dataset, allowing us to evaluate the impact of domain-specific training on a smaller-scale Japanese language LLM. In addition, we also evaluated the LLM-jp (13B) model without additional fine-tuning, serving as a larger-scale baseline for comparison. To further explore the performance of existing instruction-tuned models in Japanese, we included three publicly released fine-tuned variants of LLM-jp (13B) provided by the LLM-jp initiative: LLM-jp-13B-instruct-full-jaster (fine-tuned using the JASTER instruction dataset), LLM-jp-13B-instruct-full-jaster-dolly-oasst (fine-tuned using a merged instruction dataset combining JASTER, Dolly, and OpenAssistant [OASST]), and LLM-jp-13B-instruct-full-dolly-oasst (fine-tuned on the combined Dolly and OpenAssistant datasets, without JASTER). Dolly[40] is the Japanese translation of databricks-dolly-15k, and OASST[41] is the Japanese translation of the English subset of OASST. The outputs from different models are presented in Table 3. The fine-tuned model generated responses that are more aligned with the intended clinical context and included essential contextual details. In contrast, the LLM-jp without fine-tuning was overly brief and omitted critical information.

In the evaluation phase, we adopted a comprehensive set of four widely recognized metrics to assess the performance of the LLMs on epilepsy-related tasks. Specifically, the evaluation metrics included: (i) Bilingual Evaluation Understudy (BLEU),[42] a precision-based metric that measures the overlap between the generated text and the reference text by calculating n-gram matches; (ii) Metric for Evaluation of Translation with Explicit ORdering (METEOR),[43] which extends beyond simple n-gram overlap by incorporating synonym matching, stemming, and word order penalties, making it more sensitive to linguistic variations and semantic meaning; (iii) Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L),[44] which focuses on the longest common subsequence between the generated and reference texts, emphasizing both the sequence and recall aspects, and providing insights into how much essential information is preserved; (iv) Semantic Propositional Image Caption Evaluation (SPICE),[45] which evaluates the semantic content of the generated text by parsing it into scene graphs and comparing the relationships and objects described, offering a conceptual-level evaluation particularly valuable for specialized and technical domains.

The experimental results are displayed in Table 4, demonstrating that fine-tuning the base LLM with epilepsy-specific knowledge data significantly improves performance across all evaluation metrics. Fine-tuning allows the models to better capture specialized terminologies and contexts required for epilepsy-related tasks. Among the models, LLM-jp (1.3B) achieved the highest performance despite having the smallest number of parameters. This is most likely associated with the model's extensive use of Japanese language data during pre-training, which enhanced its ability to understand and

[4] https://github.com/masa3141/japanese-alpaca-lora/tree/main

**Table 3. An example of different model outputs**

| Model | Output |
|---|---|
| Test data | スポーツは、身体を丈夫にして抗てんかん薬の副作用を出にくくしますので、してかまいません。ただし、発作が万が一おきた場合に致命的になるようなスポーツ（スキューバダイビング、ロッククライミング、など）は避けましょう。球技や陸上競技は問題ありません。(Sports are fine, as they strengthen the body and reduce the side effects of antiepileptic drugs. However, sports in which a seizure could be fatal (scuba diving, rock climbing, etc.) should be avoided. Ball games and track and field are fine.) |
| LLM-jp (1.3B) | 運動が可能です。(It is possible to exercise.) |
| LLM-jp (1.3B) and epilepsy data | てんかんのある人が、時々、運動をすることは問題ありません。精神活動が、脳と身体の健康に良いからです。しかし、急な動作は危険ですので、慎むようにしてください。また、激しいスポーツは発作の誘発になりますので、中止するべきです。(It is okay for people with epilepsy to exercise occasionally, as mental activity is good for the brain and body. However, sudden movements can be dangerous, so they should be avoided. Also, strenuous sports can trigger seizures, so they should be avoided.) |

Note: Output refers to the model's answer to the question: スポーツをしてもいいですか。 (Can I play sports?).

**Table 4. Large language model (LLM) performance in epilepsy-related tasks**

| Model | First fine-tuning | Second fine-tuning | These are dimensionless scores (pure numbers) | | | | |
|---|---|---|---|---|---|---|---|
| | | | BLEU | METEOR | ROUGE-L | SPICE | Mean |
| BioMedLM | - | - | 0.0058 | 0.0091 | 0.0092 | 0.0000 | 0.0060 |
| LLaMA (7B) | - | - | 0.0173 | 0.0237 | 0.0234 | 0.0069 | 0.0178 |
| LLaMA (7B) | Epilepsy data | - | 0.2256* | 0.1836* | 0.2820* | 0.1045 | 0.1989* |
| LLaMA (7B) | Alpaca | - | 0.0273 | 0.0418 | 0.0639 | 0.0439 | 0.0442 |
| LLaMA (7B) | Alpaca | Epilepsy data | 0.1701 | 0.1705 | 0.2347 | 0.1070 | 0.1706 |
| LLaMA (7B) | Alpaca (Japanese) | - | 0.1637 | 0.1380 | 0.2217 | 0.0876 | 0.1528 |
| LLaMA (7B) | Alpaca (Japanese) | Epilepsy data | 0.2037 | 0.1678 | 0.2668 | 0.1308* | 0.1923 |
| LLaMA (13B) | - | - | 0.0281 | 0.0559 | 0.0417 | 0.0057 | 0.0328 |
| LLaMA (30B) | - | - | 0.0281 | 0.0572 | 0.0417 | 0.0057 | 0.0332 |
| LLM-jp (1.3B) | - | - | 0.1418 | 0.1793 | 0.1805 | 0.0144 | 0.1290 |
| LLM-jp (1.3B) | Epilepsy data | - | 0.2351* | 0.2314* | 0.2631* | 0.0727* | 0.2006* |
| LLM-jp (13B) | - | - | 0.1673 | 0.2102 | 0.2010 | 0.0198 | 0.1496 |
| LLM-jp (13B) | Jaster | - | 0.0004 | 0.0192 | 0.0174 | 0.0160 | 0.0132 |
| LLM-jp (13B) | Dolly (Japanese) | - | 0.0880 | 0.0891 | 0.1421 | 0.0647 | 0.0960 |
| LLM-jp (13B) | Jaster; Dolly (Japanese) | - | 0.0712 | 0.0889 | 0.1295 | 0.0712 | 0.0902 |

Note: *Indicates the best performance in each metric for LLaMA or LLM-jp.
Abbreviations: BLEU: Bilingual Evaluation Understudy; METEOR: Metric for Evaluation of Translation with Explicit ORdering; ROGUE-L: Recall-Oriented Understudy for Gisting Evaluation; SPICE: Semantic Propositional Image Caption Evaluation.

adapt when fine-tuned on Japanese epilepsy datasets. These results highlight the importance of both domain-specific knowledge and language alignment in improving LLM performance on specialized non-English tasks. Although these evaluation methods can provide useful insights into surface-level overlap and some aspects of semantic similarity, they may not fully capture clinical relevance or factual adequacy in medical contexts. In this study, we did not conduct a formal human evaluation due to resource constraints, but we recognize its value—particularly for assessing factual correctness, clinical appropriateness, and usability in real-world scenarios—and plan to implement it in future studies.

## 4. Discussion

While LLMs excel in a wide range of general tasks, their performance often falls short in specialized professional domains. This gap is particularly evident in tasks requiring domain-specific knowledge, where responses may lack the required depth, accuracy, and professionalism. Furthermore, these limitations become even more pronounced in non-English language tasks, where linguistic nuances and domain-specific terminology are not always well captured by models primarily pre-trained on English corpora. To address these challenges and enhance the reliability and professionalism of LLM-generated responses, recent research has explored fine-tuning pre-trained models using domain-specific knowledge bases. By integrating specialized expertise into the model, it becomes possible to significantly improve its performance on tasks requiring deep understanding within a specific field. In this study, we focused on the domain of epilepsy, aiming to fine-tune a pre-trained LLM using highly granular and specialized knowledge related to epilepsy. Unlike general domain adaptation, we focused on fine-tuning at a more detailed, technical level to ensure that the model can handle complex epilepsy-related inquiries with greater precision. Importantly, our fine-tuning efforts were conducted using non-English language resources to further bridge the linguistic and professional knowledge gaps. Through this approach, we aimed to develop a language model that delivers highly accurate, professional, and contextually appropriate responses in specialized epilepsy-related tasks, even in non-English settings.

In our experiments, we selected two different open-source pre-trained LLMs as base models for fine-tuning. The fine-tuning datasets were meticulously curated from a variety of publicly available websites, focusing specifically on sources that provide high-quality, domain-specific knowledge related to epilepsy. This approach ensured that the fine-tuning corpus was both comprehensive and rich in specialized content. Among the two models evaluated, the Japanese language model LLM-jp (1.3B) demonstrated the highest overall performance following

fine-tuning. After incorporating detailed epilepsy knowledge, LLM-jp exhibited significant improvements in generating accurate, contextually appropriate, and professional responses to epilepsy-related queries in Japanese. These results highlight the effectiveness of leveraging fine-tuned domain knowledge to enhance the specialized capabilities of language models in non-English settings. In contrast, BioMedLM, a medical-domain pre-trained model originally optimized for English-language biomedical tasks, did not achieve comparable performance levels when evaluated on Japanese epilepsy tasks. The primary limitation stemmed from BioMedLM's lack of robust support for the Japanese language. Despite its strong foundation in biomedical knowledge, the absence of multilingual capabilities, particularly in Japanese, restricted its ability to adapt effectively to non-English fine-tuning, resulting in less professional and less accurate outputs compared to LLM-jp. These experimental findings underscore the critical importance of both language compatibility and domain specificity when fine-tuning LLMs for specialized tasks in non-English languages.

For LLaMA models without fine-tuning, increasing the parameters (from 7B to 30B) did not yield significant performance improvement. For baseline models without fine-tuning, we observed that simply increasing the model size did not significantly improve performance on Japanese epilepsy-related tasks. Despite the larger parameter count, the models struggled to understand and generate accurate responses to specialized epilepsy-related questions in Japanese. This limitation can be attributed to two main factors. First, the original pre-training corpus of LLaMA models contains very limited professional medical knowledge, particularly with respect to the field of epilepsy. Second, the proportion of Japanese language data included in the pre-training dataset is extremely small, which further restricts the model's ability to handle non-English, domain-specific inquiries effectively. As a result, across all model sizes, the baseline LLaMA models demonstrated inadequate comprehension and response quality when evaluated on Japanese epilepsy test datasets. However, after fine-tuning the LLaMA-7B model using a carefully curated dataset focused on Japanese language epilepsy knowledge, we observed a substantial performance improvement. The fine-tuned model exhibited a significantly enhanced understanding of epilepsy-specific terminology, clinical concepts, and contextual nuances within the Japanese language. This indicates that targeted domain-specific fine-tuning can dramatically compensate for the deficiencies of the original pre-trained model, even without requiring a larger parameter size. These findings highlight that, for specialized non-English applications, fine-tuning with high-quality, domain-specific datasets is far more critical than merely scaling up the model size. A well-focused fine-tuning strategy can enable even smaller models like LLaMA-7B to achieve strong task-specific performance, outperforming larger but unfine-tuned counterparts. The Alpaca dataset effectively improved the LLaMA performance in general tasks;[23] in the epilepsy task, it also resulted in a slight performance gain. However, the second fine-tuning using the epilepsy dataset did not achieve the highest performance. By using the Japanese-translated Alpaca and epilepsy datasets, the performance of the twice fine-tuned model was improved.

The LLM-jp model, which was pre-trained using a substantial amount of Japanese language data, naturally demonstrated strong baseline performance on a variety of Japanese language tasks. Its robust handling of Japanese text gives it a considerable advantage compared to models primarily trained on English or multilingual corpora with a lower proportion of Japanese content. Building on this strong foundation, we conducted fine-tuning of LLM-jp (1.3B) using a carefully curated dataset focused specifically on epilepsy-related knowledge in Japanese. The results revealed a significant performance improvement: the evaluation metric increased from 0.129 (pre-fine-tuning) to 0.2006 (post-fine-tuning). This substantial gain highlights how domain-specific fine-tuning can enhance a model's ability to understand and accurately respond to specialized queries, even when starting from an already strong language foundation. These findings further validate the effectiveness of targeted fine-tuning strategies for specialized tasks. In particular, when addressing narrow professional domains such as epilepsy within the broader field of medicine, merely relying on general pre-training is often insufficient, regardless of the model's original language competency. By supplementing the model with domain-specific knowledge during fine-tuning, it is possible to achieve remarkable improvements in task-specific performance, ensuring that the model's outputs are not only linguistically fluent but also professionally accurate and contextually relevant. This result underscores a critical point: for specialized applications, especially in languages other than English, domain-adaptive fine-tuning is essential to bridge the gap between general language understanding and expert-level task performance.

While the performance of LLMs was demonstrated in the study, several notable limitations should be acknowledged, particularly in the context of safety-critical applications such as healthcare. One key limitation is the challenge of knowledge updates. Instruction-tuned models are typically trained on fixed datasets, making it difficult to

incorporate new clinical information without undergoing additional fine-tuning. In fast-evolving domains such as neurology, where diagnostic criteria and treatment protocols for conditions like epilepsy can change, this limitation hinders the model's ability to remain aligned with current clinical best practices.

In addition, there are also issues in their clinical application. LLMs can generate content that may be factually incorrect or misleading, a phenomenon often referred to as hallucination. This risk is especially problematic in clinical settings, where misinformation—even when subtle—can lead to adverse consequences for patient care. Finally, instruction-tuned models suffer from a lack of traceability. These models do not inherently provide mechanisms to link outputs to specific evidence or sources, making it difficult for users to verify the provenance of responses. This opacity can undermine clinician trust and limit the model's utility in settings where explainability and accountability are essential. In order to further integrate the model into clinical use, retrieval-augmented generation (RAG) is a better option for the above problems. RAG can retrieve relevant documents, allowing for easier integration of updated or external knowledge sources. By grounding responses in retrieved content, RAG has the potential to reduce hallucination and improve the traceability of clinical claims, a key limitation of current generative models.

During use, we also face the problem of overreliance. To mitigate overreliance, clinical use of LLMs must be governed by rigorous validation, clear human oversight, and well-defined boundaries of use. High-stakes decisions should remain under the purview of licensed professionals, with LLMs serving to support (and not replace) human expertise. The lack of transparency in LLMs makes it difficult to trace the source or rationale behind specific responses. This raises concerns regarding reproducibility and accountability. Ultimately, the responsible application of LLMs in clinical practice depends on a human-AI collaboration model, continuous performance monitoring, and ethical safeguards to ensure they support rather than undermine clinical judgment.

This study is focused on a specific clinical task (epilepsy) and language (Japanese), which may limit the immediate applicability of the findings to other areas. Nonetheless, the overall architecture and instruction-tuning framework we used were model- and language-agnostic in principle. With appropriate domain-specific data and adaptation, we believe the approach can be extended to other medical specialties and languages, though challenges, such as data availability, terminology alignment, and cultural context, would need to be carefully addressed.

## 5. Conclusion

In this study, we aimed to improve the professionalism and reliability of LLMs in handling epilepsy-related tasks in a non-English language. To achieve this, we fine-tuned pre-trained models using specialized epilepsy knowledge, rather than broad medical datasets commonly used in general-purpose medical LLMs. By focusing on more specific disease knowledge, the model can better understand and respond to professional epilepsy-related questions. Our experimental results revealed that narrowing the domain scope allows for significant performance improvements, even when using a relatively small amount of fine-tuning data. This indicates that targeted fine-tuning with high-quality, domain-specific information is an effective strategy for enhancing LLMs in specialized fields, especially where non-English resources are limited.

## Acknowledgments

None.

## Funding

This work was supported by JST CREST (grant number: JP-MJCR1784).

## Conflict of interest

The authors declare they have no competing interests.

## Author contributions

*Conceptualization:* All authors
*Methodology:* All authors
*Investigation:* Xuyang Zhao
*Writing–original draft:* Xuyang Zhao
*Writing–review & editing:* Qibin Zhao, Toshihisa Tanaka

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

All data used in this paper were collected from publicly accessible websites.

## Further disclosure

EpilepsyLLM is dedicated to the research of LLMs in the medical field. The medical knowledge used in model training and testing is obtained from publicly accessible websites. The response content generated by the model

cannot be guaranteed and cannot be used as a substitute for professional medical treatment.

## References


1. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. Vol. 30. United States: Cornell University; 2017.

   doi: 10.48550/arXiv.1706.03762

2. Radford A, Narasimhan K, Salimans T, Sutskever I. *Improving Language Understanding by Generative Pre-Training*. OpenAI Blog; 2018.

3. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. *Language Models are Unsupervised Multitask Learners*. Vol. 1. OpenAI blog; 2019. p. 9.

4. Brown T, Mann B, Ryder N, *et al*. Language models are few-shot learners. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. United States: Cornell University; 2020. p. 1877-1901.

   doi: 10.48550/arXiv.2005.14165

5. Ouyang L, Wu J, Jiang X, *et al*. Training Language Models to Follow Instructions with Human Feedback. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35. 2022. p. 27730-27744.

   doi: 10.48550/arXiv.2203.02155

6. OpenAI. *GPT-4 Technical Report*; 2023.

   doi: 10.48550/arXiv.2303.08774

7. Touvron H, Lavril T, Izacard G, *et al*. *LLaMA: Open and Efficient Foundation Language Models*. *arXiv:2302 13971*. United States: Cornell University; 2023.

   doi: 10.48550/arXiv.2302.13971

8. Hoffmann J, Borgeaud S, Mensch A, *et al*. *Training Compute-Optimal Large Language Models*. arXiv [Preprint]; 2022.

   doi: 10.48550/arXiv.2203.15556

9. Narang S, Chowdhery A. *Pathways Language Model (Palm): Scaling to 540 Billion Parameters for Breakthrough Performance*. Google AI Blog; 2022.

10. Zong H, Wu R, Cha J, *et al*. Large language models in worldwide medical exams: Platform development and comprehensive analysis. *J Med Int Res*. 2024;26:e66114.

    doi: 10.2196/66114

11. Brin D, Sorin V, Konen E, Nadkarni G, Glicksberg BS, Klang E. *How Large Language Models Perform on the United States Medical Licensing Examination: A Systematic Review*. MedRxiv [Preprint]; 2023. p. 2023-2009.

    doi: 10.1101/2023.09.03.23294842

12. Gilson A, Safranek CW, Huang T, *et al*. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9(1):e45312.

    doi: 10.2196/57594

13. Kim Y, Park C, Jeong H, *et al*. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol 37. 2024. p. 79410-79452.

    doi: 10.48550/arXiv.2404.15155

14. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology--a recent scoping review. *Diagn Pathol*. 2024;19(1):43.

    doi: 10.1186/s13000-024-01464-7

15. Subramanian CR, Yang DA, Khanna R. Enhancing health care communication with large language models-the role, challenges, and future directions. *JAMA Network Open*. 2024;7(3):e240347-e240347.

    doi: 10.1001/jamanetworkopen.2024.0347

16. Mukherjee S, Gamble P, Ausin MS, *et al*. *Polaris: A Safety-Focused LLM Constellation Architecture for Healthcare*. arXiv [Preprint]; 2024.

    doi: 10.48550/arXiv.2403.13313

17. Abd-Alrazaq A, AlSaad R, Alhuwail D, *et al*. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Med Educ*. 2023;9(1):e48291.

    doi: 10.2196/48291

18. Yu H, Zhou J, Li L, *et al*. *AIPatient: Simulating Patients with EHRs and LLM Powered Agentic Workflow*. arXiv [Preprint]; 2024.

    doi: 10.48550/arXiv.2409.18924

19. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ*. 2024;58(11):1276-1285.

    doi: 10.1111/medu.15402

20. Yuan T, He Z, Dong L, *et al*. *R-Judge: Benchmarking Safety Risk Awareness for LLM Agents*. arXiv [Preprint]; 2024.

    doi: 10.48550/arXiv.2401.10019

21. Ong JCL, Chang SYH, William W, *et al*. Medical ethics of large language models in medicine. *NEJM AI*. 2024;1(7):AIra2400038.

    doi: 10.1056/AIra2400038

22. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: A systematic review on large language models (LLMs). *NPJ Digit Med*. 2024;7(1):183.

    doi: 10.1038/s41746-024-01157-x

23. Taori R, Gulrajani I, Zhang T, *et al*. *Alpaca: A Strong,*

*Replicable Instruction-Following Model. Stanford Center for Research on Foundation Models.* Vol. 3. 2023. p. 7. Available from: https://crfm.stanford.edu/2023/03/13/alpaca.html

24. Taori R, Gulrajani I, Zhang T, *et al*. *Stanford Alpaca: An Instruction-following LLaMA model*. GitHub Repository; 2023. Available from: https://github.com/tatsu-lab/stanford_alpaca

25. Gu Y, Tinn R, Cheng H, *et al*. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc*. 2021;3(1):1-23.

doi: 10.1145/3458754

26. Yasunaga M, Leskovec J, Liang P. LinkBERT: Pretraining Language Models with Document Links. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*; 2022. p. 8003-8016.

doi: 10.18653/v1/2022.acl-long.551

27. Venigalla A, Frankle J, Carbin M. *BioMedLM: A Domain-Specific Large Language Model for Biomedical Text*. Vol. 23. United States: MosaicML; 2022. p. 2.

28. Luo R, Sun L, Xia Y, *et al*. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. 2022;23(6):bbac409.

doi: 10.1093/bib/bbac409

29. Tu T, Azizi S, Driess D, *et al*. *Towards Generalist Biomedical AI*. arXiv:2307 14334. United States: Cornell University; 2023.

doi: 10.48550/arXiv.2307.14334

30. Wang G, Yang G, Du Z, Fan L, Li X. *ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation*. arXiv [Preprint]; 2023.

doi: 10.48550/arXiv.2306.09968

31. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: Toward building open-source language models for medicine. *J Am Med Inform Assoc*. 2024;31(9):1833-1843.

doi: 10.1093/jamia/ocae045

32. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *Cureus*. 2023;15(6):e40895.

doi: 10.7759/cureus.40895

33. World Health Organization. *Epilepsy: A Public Health Imperative*. Geneva: World Health Organization; 2019.

34. Annegers JF, Rocca WA, Hauser WA. Causes of epilepsy: contributions of the Rochester epidemiology project. *Mayo Clin Proc*. 1996;71(6):570-575.

doi: 10.4065/71.6.570

35. Shorvon SD. The causes of epilepsy: Changing concepts of etiology of epilepsy over the past 150 years. *Epilepsia*. 2011;52(6):1033-1044.

doi: 10.1111/j.1528-1167.2011.03051.x

36. Korenke GC, Hunneman DH, Eber S, Hanefeld F. Severe encephalopathy with epilepsy in an infant caused by subclinical maternal pernicious anaemia: Case report and review of the literature. *Eur J Pediatr*. 2004;163:196-201.

doi: 10.1007/s00431-004-1402-4

37. Pauschek J, Bernhard MK, Syrbe S, *et al*. Epilepsy in children and adolescents: Disease concepts, practical knowledge, and coping. *Epilepsy Behav*. 2016;59:77-82.

doi: 10.1016/j.yebeh.2016.03.033

38. Unsworth C. Living with epilepsy: Safety during home, leisure and work activities. *Aust Occup Ther J*. 1999;46(3):89-98.

doi: 10.1046/j.1440-1630.1999.00181.x

39. Aizawa A, Aramaki E, Chen B, *et al*. *LLM-jp: A Cross-Organizational Project for the Research and Development of Fully Open Japanese LLMs*. arXiv [Preprint]; 2024.

doi: 10.48550/arXiv.2407.03963

40. Conover M, Hayes M, Mathur A, *et al*. *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*; 2023. Available from: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

41. Köpf A, Kilcher Y, Von Rütte D, *et al*. Openassistant Conversations-Democratizing Large Language Model Alignment. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. Vol 36. 2023. p. 47669-47681.

doi: 10.48550/arXiv.2304.07327

42. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*; 2002. p. 311-318.

doi: 10.3115/1073083.1073135

43. Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*; 2005. p. 65-72.

doi: 10.3115/1626355.1626389

44. Lin CY. ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. USA: Association for Computational Linguistics; 2004. p. 74-81.

45. Anderson P, Fernando B, Johnson M, Gould S. SPICE: Semantic Propositional Image Caption Evaluation. In: *Computer Vision--ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Berlin: Springer; 2016. p. 382-398.

doi: 10.1007/978-3-319-46454-1_24

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

# A bagging ensemble machine learning method for imbalanced data to predict anxiety disorders and analyze risk factors in older people: An observational study

**Jinling Wang[1]*, Michaela Black[1], Debbie Rankin[1], Jonathan Wallace[2], Catherine F. Hughes[3], Leane Hoey[3], Adrian Moore[4], Joshua Tobin[5], Mimi Zhang[5], James Ng[5], Geraldine Horigan[3], Paul Carlin[6], Kevin McCarroll[7], Conal Cunningham[7], Helene McNulty[3], and Anne M. Molloy[8]**

[1]School of Computing, Engineering and Intelligent Systems, Ulster University, Derry-Londonderry, United Kingdom

[2]School of Computing, Ulster University, Jordanstown, United Kingdom

[3]School of Biomedical Sciences, Nutrition Innovation Centre for Food and Health, Ulster University, Coleraine, United Kingdom

[4]School of Geography and Environmental Sciences, Ulster University, Coleraine, United Kingdom

[5]School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

[6]School of Health, Wellbeing and Social Care, The Open University, Belfast, United Kingdom

[7]Mercers Institute for Research on Ageing, St James's Hospital, Dublin, Ireland

[8]School of Medicine, Trinity College Dublin, Dublin, Ireland

***Corresponding author:**
Jinling Wang
(j.wang@ulster.ac.uk)

**Citation:** Wang J, Black M, Rankin D, *et al*. A bagging ensemble machine learning method for imbalanced data to predict anxiety disorders and analyze risk factors in older people: An observational study. *Artif Intell Health*. 2026;3(1):116-137. doi: 10.36922/AIH025070009

**Received:** February 12, 2025

**1st revised:** June 27, 2025

**2nd revised:** July 7, 2025

**Accepted:** July 14, 2025

**Published online:** September 8, 2025



## Abstract

Anxiety disorders (ADs) rank among the most prevalent mental health problems, especially in older people. The high risk and prevalence of ADs underscore the need for effective mental health care. Artificial intelligence has gained popularity in the diagnosis and prediction of medical conditions and diseases, including mental health problems. In this study, we developed an adapted bagging ensemble machine learning system that can be used for the diagnosis and prediction of ADs and can address the challenges posed by extremely imbalanced data from the Trinity-Ulster-Department of Agriculture study. Statistical techniques were used to identify the risk factors for ADs. Feature selection and feature engineering were conducted based on the analysis of biomarker risk factors. Five machine learning methods have been used in the developed system to build weak learner submodels, yielding promising prediction results. Some risk factors were identified. These findings will benefit the early prediction of ADs in our future studies.

***Keywords:*** Anxiety disorder; Bagging ensemble machine learning; Risk factor analysis; Diagnosis; Imbalanced data; Aging

## 1. Introduction

Anxiety disorders (ADs) are one of the most common mental health issues and are characterized by anticipation of future concerns. In the course of our daily lives, instances of anxiety are not uncommon. Mild levels of anxiety can serve to alert us and sharpen our focus on potential dangers in certain situations. ADs encompass a group of conditions, such as excessive anxiety, worry, or fear, persistent worrying on most days for over 6 months without a logical cause, and difficulties in managing these feelings, which profoundly affect normal daily functioning. Symptoms of ADs include headaches, dizziness, muscle tension and aches, bowel problems, sweating, rapid heartbeat, and shortness of breath. This paper focuses on the diagnosis and prediction of ADs, which include five main types: (i) social AD, (ii) obsessive-compulsive disorder, (iii) panic disorder, (iv) generalized AD (GAD), and (v) separation AD. The causes of ADs appear to be multifactorial, including genetic traits and triggers such as traumatic events.[1,2]

Several studies on the prevalence of ADs in the elderly have reported variable results in the incidence rate for people aged 55 years and over, which ranges from 1.2% to 14.2%.[3-6] As the global population ages, the number of people over 60 years of age is expected to exceed 2 billion by 2050.[7] ADs often lead to distress and disability, reducing the overall quality of life. They can even pose a mortality risk in older adults. They are closely linked to cognitive decline, cardiovascular hazards, and other chronic illnesses. The mechanisms underlying anxiety in the elderly are primarily linked to age-related neuropathology and substantial life transitions, including economic hardship, retirement, isolation, and bereavement, that typically occur later in life. The high prevalence of anxiety-related disorders in the population poses a challenge for mental health service providers, who try to provide face-to-face therapy sessions to those who need them in a timely manner.[8] This underscores the importance of developing effective mental health-care strategies.

Machine learning (ML) and artificial intelligence (AI) technologies are becoming increasingly popular in disease diagnosis and are particularly important in the field of mental health, as there is a global shortage of qualified professionals who can handle these issues. The high cost of mental health services and the social stigma associated with these problems often deter people from seeking assistance. If left untreated, ADs can cause functional, mental, physical, cognitive, and social impairments. This, in turn, may result in decreased quality of life, delayed recovery from illness, and increased utilization of medical services.[9] As the availability of more complex health data increases, ML and AI methods are becoming increasingly valuable for analyzing risk factors to facilitate individualized treatment based on a patient's medical condition. Hence, the identification of relevant risk factors and the prediction of the prevalence of ADs among the elderly population will allow health-care providers to develop targeted strategies to reduce the incidence of ADs.

The Trinity-Ulster-Department of Agriculture (TUDA) study (ClinicalTrials.gov identifier: NCT02664584) dataset was used in our analysis. The main contributions of this study include: (i) Exploring and identifying the potential risk factors that contribute to the diagnosis of ADs; (ii) developing a bagging ensemble system for imbalanced data to help in the diagnosis and prediction of ADs; (iii) employing a threshold-moving strategy for prediction making; (iv) identifying appropriate base submodels by comparing the performance of several ML methods employed as weak learners in the system. The specific gap in identifying potential risk factors was addressed. The developed system may serve as a predictor of heightened vulnerability to ADs.

## 2. Related works

ML and AI play an important role in enhancing insights into health care, including mental health care, to support clinical decision-making. With the increasing availability of large amounts of complex data collected from patients in the health-care sector, and the ongoing advancements in computing power, ML can be used to identify illnesses at earlier or prodromal stages. Precision medicine, which involves personalized care and treatments tailored by health-care professionals based on an individual's unique characteristics, utilizes data to uncover knowledge and patterns, enhancing the effectiveness of early interventions.[10] Various factors, including environment, location, population, and medical knowledge, can impact the accuracy of data. Therefore, it is necessary to conduct appropriate preprocessing of the data to facilitate successful decision-making. In health care, ML can be applied in numerous ways. ML can aid health-care providers in predicting disease risks among patients, forecasting the likelihood of hospital re-admission of critically ill patients, and anticipating potential disease outbreaks.[5] Health-care professionals can use ML to help patients in their daily activities, aiming to enhance decision-making processes and minimize errors. Over time, this not only reduces costs but also improves workflow and contributes to the overall well-being of individuals. Many approaches have been developed in the medical and health field. Several review papers within this domain have explored the application of ML and AI in mental health across different domains and highlighted common gaps, trends, and challenges.[11-16]

Ancillon *et al.*[16] conducted a review focusing on the detection and prediction of ADs using various bio-signals and ML methods. The study provided an overview of the advantages and disadvantages of current research efforts, intending to offer guidance for future developments in the diagnosis of ADs. Notably, random forest (RF) and support vector machines (SVM) were two of the most popular ML methods, demonstrating promising performance after being combined with feature selection. Neural networks also achieved good performance and were widely used. The review emphasized the importance of features and highlighted the benefits of integrating multimodal approaches into the context of detecting and predicting ADs.

In their 2018 survey, Khan *et al.*[17] analyzed the mental state of social media users and made a depression prediction. They observed that certain symptoms associated with mental illness could be detectable on Twitter, Facebook, and web forums. They suggested the use of automated methods to identify signs of inactivity and other mental health conditions.

Agarwal *et al.*[18] created a new system designed for the early detection of mental health disorders using social media data, aiming to prevent them from escalating. The system tracked communication patterns on social networks to facilitate the timely identification of mental health issues. The analysis includes preprocessing steps such as stemming and stop word removal, feature extraction, and classification. Ensemble classifiers integrating principles from various models, including classifiers from the Bidirectional Gated Recurrent Unit, Improved Convolution Neural Network (ICNN), and Deep Maxout, were employed. A categorization was performed using the extracted characteristics, resulting in promising performance.

Nemesure *et al.*[19] presented an ensemble approach by combining ML and deep learning to predict psychiatric diseases. The study used electronic health records, including demographic and biometric data from 4184 undergraduate students. The model demonstrated predictive performance on a held-out test set with a sensitivity of 0.66 and a specificity of 0.7. The six most important features identified for predicting GAD were up-to-date vaccinations, control examination needed, the use of other recreational drugs, pre-hypertension or hypertension, systolic blood pressure, and marijuana use. The feature "control examination needed" refers to whether the student needed a follow-up for any reason.

Shen *et al.*[20] proposed a bagging algorithm, termed BPNN-Bagging, that integrates a backpropagation neural network for diagnosing GAD. Neuro-inflammatory biomarkers, specifically cytokines and S100 calcium-binding protein B (S100B), were combined in this approach. The activation of astrocytes and microglia, which are types of glial cells supporting the function of neurons and maintaining homeostasis in the central nervous system, is induced by the production of GAD-related cytokines, while neuronal growth and plasticity can be regulated by using S100B. ML techniques were employed to rank and classify cytokines and S100B, achieving a 94.47% diagnostic accuracy for GAD.

Byeon[21] proposed a stacking ensemble approach designed to identify high-risk older adults for ADs. Base models included RF, SVM, Adaboost, and Light Gradient Boosting (GB) methods, whereas XGBoost was used for the meta-model. He explored different combinations of base models and the meta-model to build stacking models. The results showed that after appropriate selection of the base model, the predictive performance of the stacking ensemble models achieved 87.4% prediction accuracy, 85.1% precision, and 87.4% recall. The highest risk predictors were identified, such as subjective family relations, subjective loneliness, the Self-Esteem Scale, family relationship and dissolution instability, instability in family support and caregiving, subjective frequency of communication with family, and the individual and their family's experience of being a victim of a crime over the past year. This underscores a need for a tool capable of identifying older adults at high risk of developing ADs and managing them effectively.

Henry and Isa[22] proposed an implementation of ensemble methods using the open-sourced mental illness dataset to predict whether IT employees need treatment for mental health. Binary particle swarm optimization (BPSO) was used for feature selection. The performance results of Decision Tree Bagging (DT Bag), Naïve Bayes Bagging (NB Bag), and Logistic Regression Bagging (LR Bag) were presented. NB Bag obtained the highest accuracy performance at 87.86%. Naïve Bayes with BPSO feature selection had 88.44% accuracy. Based on these results, the ensemble methods, such as NB Bag, did not consistently outperform base NB with BPSO in terms of prediction accuracy.

The literature suggests that single models such as RF and SVM, combined with feature selection, can lead to effective diagnosis and prediction of ADs. In ML, ensemble techniques aim to enhance predictive results of models by combining predictions from multiple models, rather than using a single model. This approach reduces variance within a noisy dataset and addresses bias to improve the accuracy of models, while handling bias-variance

trade-offs. Ensemble methods can combine models in two ways: A homogeneous or a heterogeneous ensemble model. A homogeneous ensemble model uses a single-base ML model across all submodels. A heterogeneous ensemble model uses multiple different base ML models for each submodel. The benefits of ensemble learning include increased reliability and stability in predictions. Boosting, bagging, and stacking are the three most popular ensemble models.[23]

In both stacking and bagging (known as bootstrap aggregation), multiple weak learners are trained in parallel. Bagging involves a simple voting mechanism to sum the output of each weak learner to compute the final prediction, typically with each weak learner being of the same type. In contrast, stacking uses a meta-learner trained on the predictions of previous weak learners to output the final prediction, and its weak learners are usually of different types. Stacking ensemble models tends to perform better when the individual models are stacked appropriately, and the designed stacking model, which combines different base models and the meta-model, can achieve the best predictive performance. In both bagging and stacking methods, the input data is randomly sampled with replacement from the original dataset, allowing some instances to be used repeatedly during the training stage.[24] However, boosting learns multiple weak learners sequentially, where each subsequent model assigns more weight to the data points misclassified by the previous weak learners. The weak learners can focus on specific data points and jointly reduce prediction bias.

Although state-of-the-art ML and AI techniques have been used in several studies for mental health problems, more efforts need to be made in this field. AI and ML can be promising solutions for precision medicine tailored to the needs of individual patients.

An analysis of both univariate and multivariate risk factors, conducted on the TUDA dataset, is described in the following sections. The results were then used for feature selection and feature engineering. The structure of the proposed approach is illustrated. The predictive performance, including specificity, sensitivity, accuracy, and Matthew's correlation coefficient (MCC),[25] was compared among adopted weak learners such as RF, SVM, multilayer perceptron (MLP), GB, and Linear regression (LR), and also compared to the results of the base approaches embedded oversampling technique. Efforts were made to develop an adapted bagging ensemble ML method for the prediction of ADs with the extremely imbalanced Hospital Anxiety and Depression Scale (HADs) variable of ADs diagnosis.

## 3. Methods

### 3.1. TUDA dataset

The TUDA cohort consists of detailed sociodemographic, lifestyle, biochemical, clinical, health, and nutritional data on 5186 older people aged between 60 and 102 years who were born on the island of Ireland (Figure 1 for details). Other relevant published works[7,26,27] also provide details regarding this dataset. Conducted between 2008 and 2012, the study recruited participants from general practice clinics or hospital outpatient departments in the Republic of Ireland or Northern Ireland. Standardized protocols were used for sampling, data assessment, data recording, and centralized laboratory analysis across participants. The inclusion of participants without diagnosed dementia and the collection of non-fasting blood samples allowed for the measurement of a broad spectrum of parameters, including hematological profiles, routine biochemistry, and biomarkers of micronutrient status. In addition, from a 90-min interview involving administering a comprehensive health and lifestyle questionnaire, medical and demographic details, as well as information on medication and vitamin supplement use, were collected. Blood pressure, bone health (dual-energy X-ray absorptiometry scans), physiological function tests, and cognitive function tests were also conducted.

The initial dataset contained 701 variables. During preprocessing, variables were grouped into categories including lifestyle, body measures, sociodemographics, diseases, medications, cognitive function, biochemistry, clinical, and nutritional data (Figure 1) based on domain knowledge to facilitate feature selection, feature engineering, and future analysis. Some characteristics of TUDA cohort study participants are summarized in Table 1. The preprocessing and feature selection performed on the original dataset to obtain a subset of data are described in the next subsection.

### 3.2. Preprocessing of the TUDA dataset

Initially, exploratory data analysis was performed, and the dataset was preprocessed following the pipeline shown in Figure 2.

First, in the initial cleansing and exploration phase, variable values were manually checked, identified, and corrected for issues such as spelling mistakes, incorrect units, coding inconsistencies, and invalid values. Duplicate and less relevant variables, as advised by domain experts, were identified and removed. Manual data processing can introduce noise, such as errors and inconsistencies. To mitigate this, a data dictionary was used to maintain consistent definitions of variables and formats. Acceptable

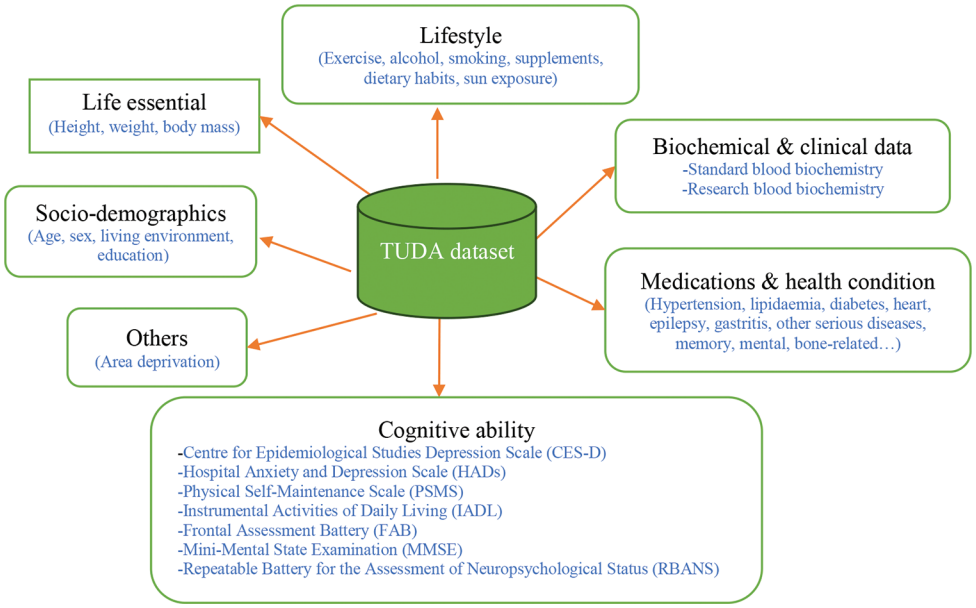



**Figure 1.** The features of the Trinity-Ulster-department of agriculture dataset were grouped into categories based on domain knowledge

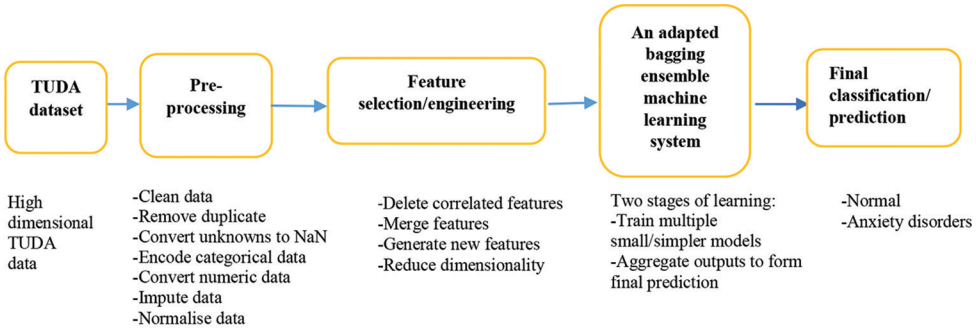



**Figure 2.** Pre-processing and analysis of the Trinity-Ulster-Department of Agriculture dataset with 701 features
Abbreviation: NaN: Undefined value.

values and formats for each variable were defined to prevent data entry errors. In addition, domain experts were invited to help group the data and identify and fix inconsistencies and errors in the dataset.

Second, when dealing with missing values, the aim is to retain as many valuable predictor variables as possible without introducing bias by needing to fill in more missing values. Analyses were performed with the cutoff threshold set at different percentages. By balancing, 10% was set as the cutoff threshold. Where the number of missing values for a specific variable was less than the threshold, the variable was retained; otherwise, it was deleted. The dataset was then split into the data subsets: Real numerical continuous variables and ordinal or nominal categorical variables.

Third, real numeric variables underwent transformation using two methods: The square root and the log transformation. Continuous numeric variables were divided into two groups based on their minimum and maximum values. Variables with a minimum value of zero were square-root processed, whereas those with a minimum value not equal to zero underwent a log transformation.

Next, the nominal subset was filtered for unique values to eliminate duplicates and retain maximal information from variables with high cardinality. Text information was recoded as numeric values according to the requirement defined in the dictionary, and unknown values were converted to an undefined value (NaN). Categorical variables were encoded into numeric values using the one-hot encoding method. As a result of these operations, the cleaned dataset was entirely numerically represented, with some missing values remaining, and ready for analysis to identify characteristics that could affect the fitting of an ML model in subsequent stages.

**Table 1. General characteristics of the 5186 participants on the Trinity-Ulster and Department of Agriculture study**

| Characteristics | Males (*n*=1699) | Females (*n*=3487) |
|---|---|---|
| Age (years), mean (SD) (95% CI) | 73.5 (8.1) (73.1–73.8) | 74.3 (8.3) (74.0–74.6) |
| Education (years)[a], mean (SD) (95% CI) | 16.0 (3.2) (15.9–16.2) | 16.0 (2.9) (15.9–16.1) |
| Health and lifestyle | | |
| BMI (kg/m$^2$), mean (SD) (95% CI) | 28.5 (4.5) (28.3–28.7) | 27.6 (5.8) (27.4–27.8) |
| Waist-to-hip ratio, mean (SD) (95% CI) | 0.97 (0.07) (0.967–0.974) | 0.88 (0.07) (0.877–0.882) |
| Timed up and go (seconds), mean (SD) (95% CI) | 13.8 (9.7) (13.3–14.2) | 14.2 (8.9) (13.9–14.5) |
| Living alone, *n* (%) | 379 (22.3) | 1372 (39.3) |
| Current smoker, *n* (%) | 1185 (69.7) | 1539 (44.1) |
| Current alcohol consumers, *n* (%) | 1099 (64.7) | 1876 (53.8) |
| Past alcohol consumers, *n* (%) | 338 (19.9) | 586 (16.8) |
| Socioeconomically most deprived, *n* (%) | 438 (25.8) | 886 (25.4) |
| Neuropsychiatric assessment | | |
| Depression (CES-D score), mean (SD) (95% CI) | 5.34 (6.8) (5.02–5.66) | 6.41 (7.8) (6.16–6.67) |
| Anxiety (HADs score), mean (SD) (95% CI) | 2.75 (3.5) (2.59–2.92) | 3.39 (3.7) (3.26–3.51) |
| FAB score, mean (SD) (95% CI) | 15.2 (2.7) (15.1–15.3) | 15.2 (2.9) (15.1–15.3) |
| PSMS score, mean (SD) (95% CI) | 23.1 (1.8) (23.1–23.2) | 22.9 (2.0) (22.8–22.9) |
| IADL score, mean (SD) (95% CI) | 24.3 (4.6) (24.1–24.5) | 24.0 (4.1) (23.9–24.2) |
| RBANS score, mean (SD) (95% CI) | 84.7 (15.9) (84.0–85.5) | 85.8 (17.5) (85.2–86.4) |
| Clinical measures | | |
| White cell count (109/L), mean (SD) (95% CI) | 7.15 (3.2) (7.00–7.31) | 7.26 (15.9) (6.74–7.79) |
| Hemoglobin (g/DL), mean (SD) (95% CI) | 14.0 (1.6) (13.9–14.0) | 12.9 (1.3) (12.8–12.9) |
| Mean corpuscular volume (FL)[b], mean (SD) (95% CI) | 90.9 (5.6) (90.7–91.2) | 91.0 (5.5) (90.8–91.2) |
| Platelet count (10$^9$/L), mean (SD) (95% CI) | 233 (65.9) (230.1–236.4) | 269 (71.7) (266.8–271.5) |
| Urea (mmol/L), mean (SD) (95% CI) | 7.32 (3.1) (7.17–7.46) | 6.79 (2.6) (6.70–6.87) |
| Creatinine (μmol/L), mean (SD) (95% CI) | 98.6 (31.7) (97.0–100.1) | 78.9 (24.7) (78.1–79.8) |
| Albumin (g/L), mean (SD) (95% CI) | 41.9 (3.8) (41.7–42.1) | 41.9 (3.5) (41.7–42.0) |
| Gamma GT (U/L), mean (SD) (95% CI) | 43.5 (51.6) (41.1–46.0) | 34.0 (37.4) (32.8–35.3) |
| Sodium (mmol/L), mean (SD) (95% CI) | 139.3 (4.8) (139.1–139.5) | 139.3 (20.6) (138.7–140.0) |
| Potassium (mmol/L), mean (SD) (95% CI) | 4.29 (0.5) (4.27–4.31) | 4.11 (0.4) (4.09–4.12) |
| Calcium (mmol/L), mean (SD) (95% CI) | 2.30 (0.13) (2.29–2.30) | 2.33 (0.14) (2.326–2.335) |
| Phosphate (mmol/L), mean (SD) (95% CI) | 0.96 (0.2) (0.95–0.97) | 1.04 (0.2) (1.04–1.05) |
| Alkaline phosphatase (U/L), mean (SD) (95% CI) | 83.2 (38.3) (81.4–85.0) | 82.0 (34.2) (80.8–83.1) |
| Low-density lipoprotein (mmol/L), mean (SD) (95% CI) | 2.19 (0.8) (2.15–2.23) | 2.56 (0.9) (2.53–2.60) |
| High-density lipoprotein (mmol/L), mean (SD) (95% CI) | 1.25 (0.4) (1.24–1.27) | 1.59 (0.5) (1.57–1.60) |
| Triglycerides (mmol/L), mean (SD) (95% CI) | 1.70 (0.9) (1.66–1.75) | 1.53 (0.8) (1.50–1.56) |
| C-reactive protein (mg/L), mean (SD) (95% CI) | 7.38 (19.2) (6.47–8.29) | 6.71 (16.1) (6.17–7.24) |
| Glycated hemoglobin (%), mean (SD) (95% CI) | 5.97 (0.9) (5.92–6.01) | 5.83 (0.7) (5.81–5.86) |
| Parathyroid hormone (pg/mL), mean (SD) (95% CI) | 45.9 (29.6) (44.5–47.3) | 47.4 (33.7) (46.3–48.5) |
| Glomerular filtration rate (mL/min), mean (SD) (95% CI) | 74.6 (25.9) (73.4–75.8) | 63.7 (23.0) (62.9–64.5) |
| Nutritional biomarkers | | |
| Red blood cell folate (nmol/L), mean (SD) (95% CI) | 1055 (622) (1025–1085) | 1121 (623) (1101–1142) |
| Serum Vitamin B12 (pmol/L), mean (SD) (95% CI) | 268 (173) (260–276) | 300 (221) (293–307) |

*(Cont'd...)*

**Table 1.** ***(Continued)***

| Characteristics | Males (*n*=1699) | Females (*n*=3487) |
|---|---|---|
| Plasma Vitamin B6 (nmol/L), mean (SD) (95% CI) | 70.7 (50.5) (68.3–73.1) | 80.5 (72.9) (78.1–82.9) |
| Riboflavin (EGRac)[c], mean (SD) (95% CI) | 1.35 (0.2) (1.30–1.40) | 1.34 (0.2) (1.30–1.40) |
| Total plasma homocysteine (μmol/L), mean (SD) (95% CI) | 15.5 (6.0) (15.2–15.8) | 14.5 (5.7) (14.3–14.7) |
| Total Vitamin D (nmol/L), mean (SD) (95% CI) | 53.0 (27.9) (51.6–54.3) | 62.1 (32.4) (61.0–63.2) |

Notes: [a]Education refers to the age of stopping formal education. [b]FL: Femtolitre. [c]EGRac: Erythrocyte glutathione reductase activation coefficient, with a higher EGRac value indicating poorer riboflavin status.
Abbreviations: BMI: Body mass index; CES-D: Center for epidemiological studies depression scale; FAB: Frontal assessment battery; HADs: Hospital anxiety and depression scale; IADL: Instrumental activities of daily living; PSMS: Physical self-maintenance scale; RBANS: Repeatable battery for the assessment of neuropsychological status.

Finally, the two data subsets were concatenated, the K-Nearest Neighbors algorithm was used to fill missing values, and the dataset was normalized to the range between 0 and 1 using z-score normalization.

The primary objectives of this study were the diagnosis and prediction of ADs. In an existing work,[28] a self-reported anxiety variable was utilized as the outcome variable where participants reported that they had, at some point in their lifetime, either been diagnosed with anxiety by a doctor or not. The variable encompassed 4064 participants (78.36%) who reported that they did not have an anxiety diagnosis, and 1122 participants (21.64%) who reported that they did have an anxiety diagnosis. This resulted in an approximate 78:22 split in the class of this self-reported outcome variable. This formed the basis for using binary classification models to fit the data. A notable challenge was the inherent imbalance in the outcome variable, as only 22% of the patients had reported an anxiety diagnosis. To address this, synthetic records of the minority class in the training dataset were generated using the standard synthetic minority over-sampling technique (SMOTE) method[29,30] and its extension, the adaptive synthetic sampling (ADASYN) algorithm.[31] The test set was unchanged to preserve the representativeness of the original population. This approach ensures fair comparisons with other methods, as well as reliable predictions on the unseen test set.

In this study, we used the score from the HADs assessment as the outcome variable. HADs are a standard tool designed to measure levels of anxiety and depression in individuals. Scores range from 0 to 21, where a score of 11 or greater indicates ADs, whereas a score <11 is considered normal. Out of 5186 participants, 4918 (94.83%) were diagnosed as not having anxiety, 260 participants (5.01%) had an anxiety diagnosis, and 8 (0.16%) had missing values. The extreme imbalance in the class, with an approximate 95:5 split, hinders the direct construction of models using conventional ML methods. Using oversampling techniques could introduce bias due to the large proportion of records that would need to be oversampled. To address this challenge, we developed a novel diagnosis and prediction system specifically designed to handle imbalanced data.

### 3.3. Feature selection and engineering

Correlation analysis is required when dealing with correlated and multicollinear predictor variables to avoid potentially unstable estimates and redundant predictor variables that do not contribute additional information for developing models.[7] To select predictor variables, we explored correlations between the diagnosis of ADs and predictor variables. In the beginning, risk factor analysis was conducted using an unimputed and unnormalized dataset to investigate associations between an anxiety diagnosis and nominal predictor variables, such as medications, lifestyle factors, and diseases. Depending on the results, some predictor variables were merged, and new predictor variables were created. For example, a new predictor variable (LimitAct) was created by merging "Limithouseholdactivities" and "Limitoutsideactivities" that have a very strong association (Cramer C = 0.6088). The predictor variable "Limithouseholdactivities" has two values: 1 (Yes) and 0 (No), by answering the question "Does the participant limit any household activities because they are afraid they might fall?" The predictor variable "Limitoutsideactivities" has two values: 1 (Yes) and 0 (No), by answering the question "Does the participant limit outside activities because they are frightened, they might fall?" The created predictor variable took the value logical 1 if at least one of the values of these two predictor variables was 1; logical 0 was taken if the values for both are 0. Highly correlated predictor variables that did not add meaningful information for prediction were removed.

The Spearman non-parametric correlation coefficient was used to calculate correlations between numerical predictor variables. Table 2 shows the relationships among the cognitive function numerical predictor variables. Nominal and ordinal variables are common types of categorical variables. Cramer's V, a statistic in the range of

**Table 2. Associations between cognitive function numerical variables**

| | | | | | | |
|---|---|---|---|---|---|---|
| CED-S | 1.00 | | | | | |
| HADs | 0.43 | 1.00 | | | | |
| IADL | −0.20 | −0.03 | 1.00 | | | |
| FAB | −0.16 | −0.05 | 0.44 | 1.00 | | |
| PSMS | −0.17 | −0.07 | 0.59 | 0.30 | 1.00 | |
| RBANS | −0.21 | −0.07 | 0.47 | 0.66 | 0.26 | 1.00 |
| | CED-S | HADs | IADL | FAB | PSMS | RBANS |

Notes: The colors indicate the associations between cognitive function numerical predictive variables. Different colors indicate different levels of intensity.
Abbreviations: CES-D: Center for epidemiological studies depression scale; FAB: Frontal assessment battery; HADs: Hospital anxiety and depression scale; IADL: Instrumental activities of daily living; PSMS: Physical self-maintenance scale; RBANS: Repeatable battery for the assessment of neuropsychological status.

[0, 1], was used to explore the association between these nominal variables. A value of 0 indicates no association between the two variables, whereas a value of 1 indicates a strong association. Cramer's V was calculated as shown in Equation I. Cramer's V was used to identify considerable variation and strong associations between the nominal variables. The results were then used to identify redundant variables, which were subsequently removed from the dataset. For example, the predictor variable "lipid_meds," representing lipidemia medication intake, was removed as it has a strong association with the predictor variable "Hyperlipidemiadiagnosis" (Cramer's V = 0.7168), which represents the diagnosis of hyperlipidemia.

$$\text{Cramer's V} = \sqrt{\frac{\chi^2}{n^{*}\min(c-1, r-1)}} \quad (1)$$

where $\chi^2$ is the Chi-square statistic, n represents the total sample size, r represents the number of rows, and c represents the number of columns. Chi-square tests were used to compare groups between participants with and without a diagnosis of AD.

A non-parametric Kruskal–Wallis test was used to determine whether the data from the two groups differed from each other. The calculated *p*-value was compared to the significance level, usually set at 0.05. If the $p>0.05$, the null hypothesis can be retained, signifying that the Kruskal–Wallis test does not detect a significant difference between categories of independent variables with respect to the dependent variable. Otherwise, the null hypothesis can be rejected.

Upon completion of the preprocessing steps, a total of 5186 records with 84 variables (83 predictors and 1 outcome) were retained for the following analysis. The outcome denotes the anxiety diagnosis of each participant, determined using HADs. In the following experiments, an analysis of risk factors was conducted.

### 3.4. Risk factor analysis

While the exact reasons for mental health problems remain unclear, attempts have been made to identify potential risk factors.[32,33] Our previous research using self-reported ADs diagnoses[28] identified several potential risk factors for ADs in older adults. These include female sex, loneliness, separated/divorced status, low socioeconomic status, lifestyle-related factors such as smoking and alcohol intake, as well as medical conditions such as cardiovascular disease, lipidemia, diabetes, hypertension, some chronic inflammatory diseases, and family history of diseases such as stroke and presenile dementia. In this study, we used the HADs ADs diagnosis to analyse risk factors. Identifying and intervening in modifiable risk factors can contribute to the mitigation or prevention of ADs.[9]

Table 3 shows the results of the univariate analysis for participant characteristics related to ADs diagnosis based on the total HADs score. Since only the anxiety questions are included in the HADs score in the TUDA study, we used HADs to assess anxiety.

Notably, "Gender" emerges as a key feature, with 5.9% of females experiencing ADs, compared to 3.7% of males. These differences may be attributed to hormonal fluctuations and variations in brain chemistry throughout a woman's life, potentially linked to ADs. Marital status also appears influential, as participants in the separated/divorced and widow/widower categories showed a slight susceptibility to ADs. Interestingly, living with others seems to mitigate the prevalence of ADs in older people, potentially due to the emotional support from others in shared living arrangements. Participants residing in low socioeconomic areas demonstrated a higher likelihood of experiencing ADs, aligning with findings from previous studies.[26] Lifestyle factors such as smoking were identified as risk factors for ADs, consistent with existing research.[34] Our analysis revealed that approximately 5.0% of participants had ADs, and the presence of a family history of stroke, heart disease, or presenile dementia increased the risk of ADs (Table 3).

Figures 3 and 4 illustrate the percentages of frequency distribution of predictors related to the effect of food supplements, Vitamin B, and Vitamin D in terms of ADs diagnosis. Vitamin D helps our body better absorb calcium, a key building block for our bones and an essential nutrient for overall health. It has been reported that nearly one billion people worldwide have low levels

**Table 3. Univariate analysis of general characteristics of the Trinity-Ulster Department of Agriculture study participants**

| Variables | HADs (anxiety diagnosis) | Effect size estimates | Effect size estimates |
|---|---|---|---|
| | Yes (*n*=260) n (%) | Cramer's V | Odds ratio |
| Gender | | 0.044 | 1.591 |
| Male | 62 (3.7) | | |
| Female | 198 (5.9) | | |
| Marital status | | 0.017 | |
| Single | 25 (3.8) | | 1.377 |
| Married/common law | 134 (5.0) | | 1.031 |
| Separated/divorced | 13 (5.2) | | 0.958 |
| Widow/widower | 88 (5.6) | | 0.843 |
| Area deprivation | | 0.041 | 1.737 |
| Normal | 159 (4.3) | | |
| SESlow | 95 (7.2) | | |
| Accommodation status | | 0.022 | |
| Alone | 82 (4.7) | | 1.114 |
| Spouse/partner | 135 (5.0) | | 0.995 |
| Other | 8 (3.4) | | 0.691 |
| Children | 35 (6.8) | | 1.511 |
| Smoking | | 0.014 | 1.204 |
| No | 112 (4.6) | | |
| Yes | 148 (5.4) | | |
| Drinking alcohol | | 0.013 | |
| Never | 65 (5.1) | | 0.986 |
| Past | 54 (5.8) | | 0.820 |
| Current | 141 (4.7) | | 1.145 |
| Vitamin D | | 0.024 | 1.356 |
| >50 nmol/L | 125 (4.4) | | |
| ≤50 nmol/L | 135 (5.9) | | |
| Fortified food and supplements | | 0.0057 | 0.948 |
| No/low | 106 (5.2) | | |
| Medium/high | 154 (4.9) | | |
| Diagnosis of | | | |
| Hypertension | | | |
| Yes | 184 (5.0) | 0.016 | 0.930 |
| No | 76 (5.3) | | |
| Hyperlipidemia | | | |
| Yes | 150 (5.5) | 0.016 | 1.200 |
| No | 106 (4.6) | | |
| Diabetes | | | |
| Yes | 30 (4.5) | 0.011 | 0.884 |
| No | 230 (5.1) | | |
| Other serious diseases | | | |
| Yes | 62 (4.3) | 0.019 | 0.816 |
| No | 198 (5.3) | | |
| Self-memory concern | | | |
| Yes | 113 (7.5) | 0.048 | 1.875 |
| No | 138 (4.1) | | |
| Family-memory concern | | | |
| Yes | 67 (10.2) | 0.061 | 2.461 |
| No | 184 (4.4) | | |
| Family history | | | |
| Cancer | | | |
| Yes | 70 (4.9) | 0.0028 | 0.960 |
| No | 190 (5.1) | | |
| Stroke | | | |
| Yes | 27 (5.8) | 0.0073 | 1.168 |
| No | 233 (5.0) | | |
| Heart disease | | | |
| Yes | 79 (5.2) | 0.0033 | 1.048 |
| No | 181 (5.0) | | |
| Presenile dementia | | | |
| Yes | 7 (9.6) | 0.017 | 2.014 |
| No | 253 (5.0) | | |
| Senile dementia | | | |
| Yes | 34 (4.9) | 0.0023 | 0.958 |
| No | 225 (5.1) | | |

Abbreviations: HADs: Hospital anxiety and depression scale; SESlow: Low standard socioeconomic status.

of Vitamin D, and approximately 20% of the population in the UK suffers from a Vitamin D deficiency.[35] Numerous studies have highlighted the link between Vitamin D deficiency and increased anxiety.[36,37] Figure 3A shows the percentages of participants who have an anxiety diagnosis and their Vitamin D levels. 5.85% of participants with a Vitamin D deficiency experienced AD, compared to 4.38% with normal vitamin D levels. Participants with low Vitamin D levels appeared to have a higher chance of experiencing ADs. Research suggests that Vitamin D, influencing brain receptors and mental health, may play a role in cell growth promotion. In theory, a Vitamin D deficiency could potentially limit this behaviour, impeding overall brain function.[38,39] Vitamin D supplements could serve as preventive measures or potential treatments for

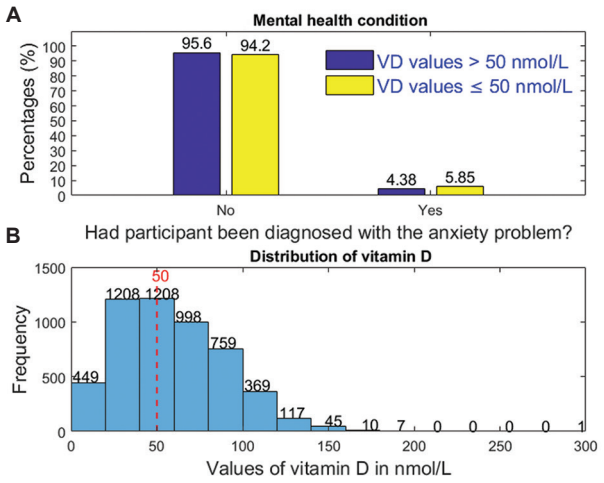



**Figure 3.** Bar plot showing percentages of participants with anxiety diagnosis and their Vitamin D levels (A) and distribution of Vitamin D values (B)

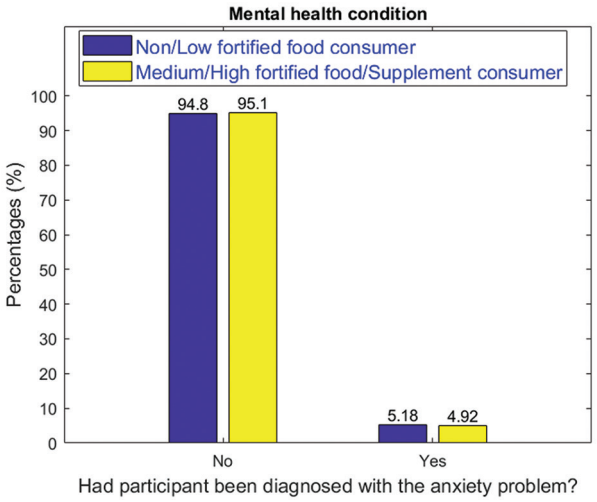



**Figure 4.** Bar plot showing percentages for non/low fortified food consumers and medium/high fortified food/supplement consumers classified based on their ADs diagnosis

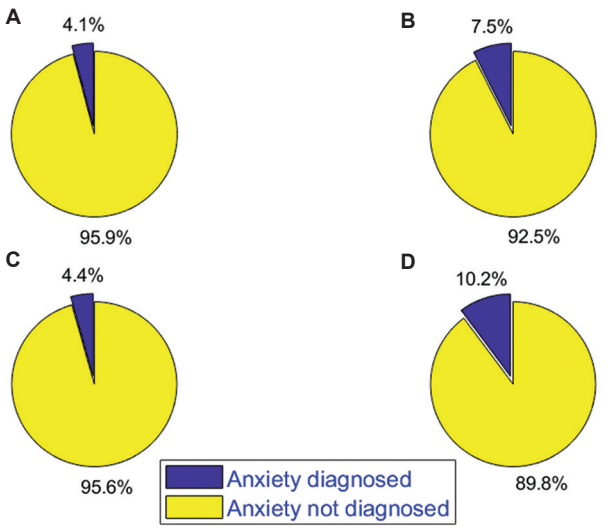



**Figure 5.** Pie chart illustrating percentages of anxiety diagnosis for participants with (A) non-self-memory concern, (B) self-memory concern, (C) non-family-memory concern and (D) family-memory concern

anxiety and depression due to their strong associations. Another study found that individuals with anxiety tend to have lower levels of calcidiol, the most active form of Vitamin D. Calcidiol helps the body use more calcium from foods or supplements and regulates the production of parathyroid hormones in the body.[40]

Figure 4 shows the percentage distribution for predicting ADs diagnosis among consumers of non- or low-fortified foods and consumers of medium- or high-fortified foods or supplements. It should be noted that this feature was engineered by combining variables related to fortified food and supplements in the original dataset. Consumers of non- or low-fortified food included participants who did not consume fortified food and supplements or consumed 1–4 servings of fortified foods per week. Consumers of medium- or high-fortified foods or supplements included participants who consumed 5–7 servings of fortified foods per week (medium consumers), or more than 8 servings of fortified foods per week (high consumers), or users of supplements. Participants with non/low fortified food intake appeared to have a higher chance of experiencing ADs compared to those with medium/high fortified food/supplements intake. This suggests that fortified foods may offer mental health benefits.

The participant was asked if their family had any concerns with regard to their memory (family-memory concern), and if they themselves had any concerns about their memory (self-memory concern). Figure 5 shows percentages of anxiety diagnosis for participants with (a) self-memory concern and (b) family-memory concern. From the results, we can see that participants who had a concern about their memory, and/or whose family had concerns about their memory, were more prone to ADs than those with no memory concerns reported.

Figures 6 and 7 present the results of multivariate analysis for lifestyle characteristics in terms of smoking status and alcohol intake of participants, and in relation to ADs diagnosis. Figure 6 shows that, regardless of gender, smoking poses a higher risk of ADs in older people. The diagnosis of ADs increases from 8.92% for non-smoking females to 11.2% for females who smoke. The difference is less pronounced for males at 2.25% for non-smokers

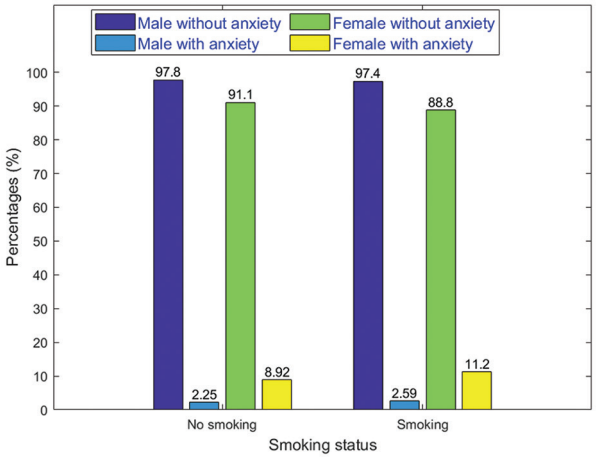



**Figure 6.** Bar plot showing percentages of participants who smoke versus those who do not, for males and females, respectively, in terms of their anxiety disorder diagnosis

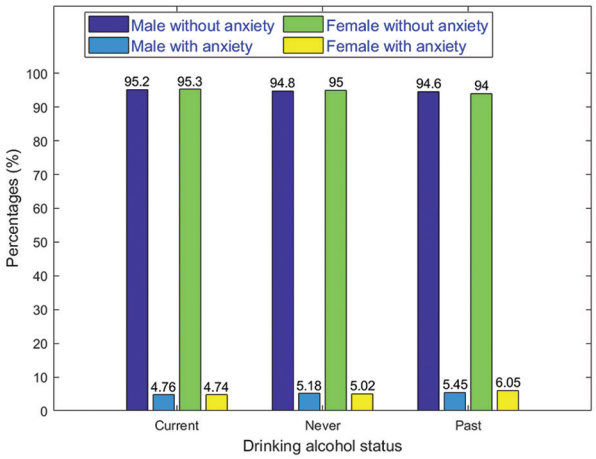



**Figure 7.** Bar plot showing percentages of participants based on their alcohol consumption status, for males and females, respectively, and in terms of anxiety disorder diagnosis. "Current:" They currently consume alcohol; "Past:" They consumed alcohol in the past but not currently; "Never:" They never consumed alcohol.

and 2.59% for smokers. Smoking appears to have a more pronounced impact on women.

Figure 7 shows how participants' alcohol intake status affects the diagnosis of ADs for males and females, respectively. The participants with a history of past alcohol consumption are more prone to ADs, especially among females.

The analysis highlights several factors associated with a higher risk of ADs, including being female, experiencing separation or divorce, being widowed, having self- or family-reported memory concerns, having a family history of chronic diseases (such as stroke or presenile dementia), smoking, and having a history of alcohol consumption. Furthermore, participants from areas with higher socioeconomic deprivation were found to have a higher prevalence of ADs, with a prevalence of 7.2% compared to 4.3% in areas with normal socioeconomic status.

Feature selection and engineering, which involves identifying and selecting the most relevant features in a high-dimensional dataset, is crucial due to the challenges posed by high dimensionality, noise reduction, and model interpretability requirements. This study used a variety of methods to reduce features: (i) Domain experts were involved in grouping the features and selecting the most relevant features; (ii) The percentage of missing values for a specific feature was compared to a threshold to determine whether to keep or remove the feature; (iii) Filter-based feature selection techniques such as statistical correlation analysis were used in the TUDA dataset to eliminate irrelevant or redundant features, and generate new features from existing similar features, for example, by removing a predictor that had a strong association with another predictor, retaining a predictor that had high correlation to the outcome feature, and by merging variables "Bone fracture" and "Hip fracture" into one predictor. These methods make feature selection and feature engineering an integral part of building effective models in this field.

## 4. Results

### 4.1. Preparation of training and test sets

The preprocessed dataset was randomly divided into two sets: A training set containing 70% of the data (3631 records) and a test set consisting of 30% of the data (1555 records). Figure 8 illustrates the process of preparing the training and test sets. To ensure consistency in model comparisons, each model followed the same procedure and was trained and tested on the same datasets. The response variable, anxiety diagnosis, relates to the diagnosis of ADs in participants using the HADs method. Among the 5186 participants, 4918 participants (94.83%) were not diagnosed with an AD, whereas 260 participants (5.01%) were diagnosed with an AD, resulting in an imbalanced class distribution of approximately 95:5. Due to this extreme imbalance, standard SMOTE oversampling could introduce bias, as a large proportion of records would require oversampling. To address this issue, we developed an adapted bagging ensemble diagnosis and prediction system, which involves reconstructing the training data.

The m value determines how many submodels are ensembled. We let *Nrn* represent all the records with "No anxiety" diagnosis, and *Nra* represent all the records with "Anxiety" diagnosis in the training set. *m* is calculated based on Equation II:

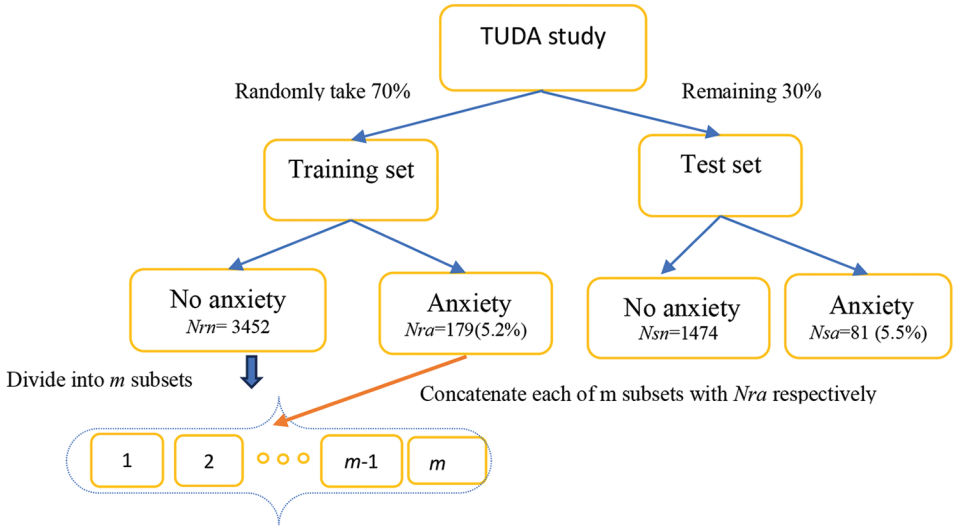



**Figure 8.** Preparation of the training and test sets, and construction of subsets in the training set

$$m = \text{round}(\frac{Nrn}{Nra}) \quad \text{(II)}$$

Then, the records with "No anxiety" diagnosis in the training set were split into m chunks with no repetition. In this study, *m* equals 20.

### 4.2. Structure of the system

Various techniques can be employed to address the challenge of imbalanced datasets. Two common methods are under-sampling and over-sampling, which aim to create a balanced dataset from an imbalanced one. Under-sampling is used when the volume of data is sufficient; it involves reducing the size of the over-represented class to balance the dataset. Conversely, over-sampling is used when the volume of data is insufficient, and it involves increasing the size of the under-represented class to balance the dataset, without removing abundant samples. Over-sampling techniques may generate new synthetic samples for the rare class. In previous work,[28] approximately 22% of participants self-reported as having ADs, and the class imbalance was addressed using two standard oversampling techniques: SMOTE[29,30] and ADASYN.[31] These methods synthetically generated additional records of the minority categories in the training dataset. Importantly, the test set was unchanged to preserve representativeness of the original population, ensure a fair comparison with other methods, and provide reliable predictions on the unseen test set. In this paper, we developed a procedure to overcome the challenge imposed by imbalanced data.

The bagging method involves generating *m* bootstrapped samples to construct models in parallel. Then, the individual predictions from the *m* models are aggregated through voting or averaging to obtain the final prediction from the ensemble of models. The main purpose of the bagging method is to minimise diversity and mitigate the risk of overfitting across the various models created. In this process, each bootstrapped sample is randomly generated from the given data records with replacement, allowing some individual records to be chosen more than once. In this paper, we use every record in the training set without repetition. We adapted the traditional bagging method and allocated the records with "No anxiety" to *m* groups; the number of records in each group roughly equals *Nra*. As 20 groups were allocated, the number of records in the first 19 groups is 173, and the number of records in the last group is 165. The flowchart illustrating this adaptation is presented in Figure 9.

First, we trained multiple small/simpler models instead of training one complex/large model for our training dataset. Normal records in the training set were divided into *m* small trunks, with the size of each trunk ensured to be the same or a size similar to that of the "Anxiety" records in the training set. Each divided subset, containing records with only "No anxiety" diagnosis, was then concatenated with records labeled with "Anxiety" diagnosis (*Nra*) in the training set. During the training phase, the generated *m* subsets were used to build *m* submodels using the same weaker learner, respectively. Second, after *m* submodels were trained, the prediction (1 = "Anxiety" diagnosis; 0 = "No anxiety" diagnosis) for each participant in the test set was obtained. During the testing phase, the held-out test set, which contains 1555 participants, was sent through to each of the generated *m* submodels, and a prediction

was generated by each submodel for a participant. These predictions of *m* submodels for each participant in the test set were summed (*Sum*), which is in the range of [0, *m*]. Then, a threshold-moving strategy by the trial-and-error method could be employed for prediction making. After a threshold (*Th*) was set, the final prediction could be obtained. If the *Sum* exceeds *Th*, the participant can be said to have an "Anxiety" diagnosis; otherwise, the participant is classified as having "No anxiety" diagnosis. The algorithm logic is based on concatenating predictions of each submodel and decision rule, ensuring the opinions of at least 80% submodels are respected. For example, if *Th* is taken as 16, the participant is diagnosed as having AD if the predictions of at least 80% of the submodels were 1. This is represented in Figure 10, showing the flowchart of the proposed bagging ensemble system for the analysis and prediction of ADs in the TUDA dataset using one specific weak learner across all submodels to predict the output. Figure 11 shows the prediction of the test set.

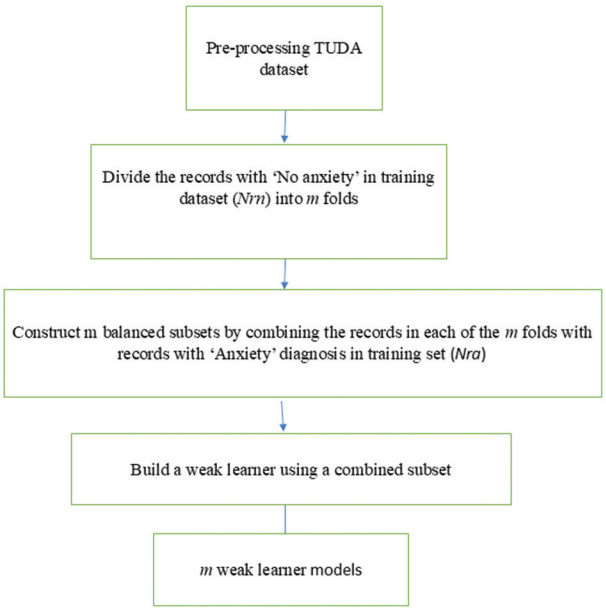



**Figure 9.** Flowchart for reconstructing *m* subsets and building *m* weak learners in parallel

### 4.3. Evaluation

Evaluation metrics such as specificity (true negative rate, TNR), sensitivity (true positive rate, TPR), accuracy, F1 score, the area under the receiver operating characteristic curve (AUC-ROC), and MCC can be used to evaluate the performance of an approach. The AUC-ROC represents the relationship between TPR and false positive rate. F1 score, AUC, and accuracy are three of the most widely adopted metrics in binary classification tasks. However, for imbalanced datasets, these statistical measures can produce misleadingly over-optimistic information. In such cases, MCC can more accurately reflect general diagnosis and prediction problems.[25]

MCC is a correlation coefficient between predicted and observed binary classifications. It is a more reliable

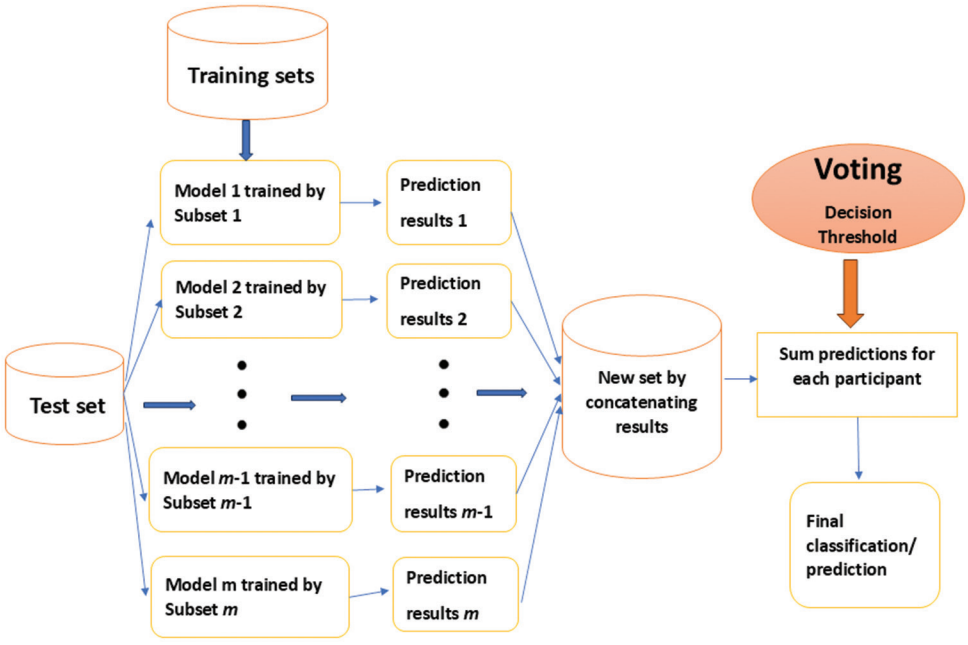



**Figure 10.** A pipeline for analysis and prediction of a high-dimensional Trinity-Ulster-Department of Agriculture dataset using an ensemble machine learning method ($m = 20$), *m* submodels with changing data records for the normal class in the training set

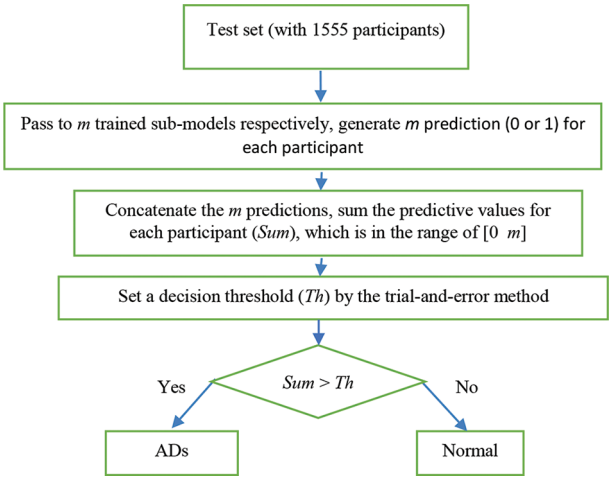



**Figure 11.** Flowchart of prediction on the test set
Abbreviation: ADs: Anxiety disorders.

statistical measure. MCC only provides a high value if good results are achieved in predictions across all four confusion matrix categories, *i.e.*, true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP). It is proportional to the size of positive and negative elements in the dataset. A perfect classification is indicated by an MCC value of 1, whereas values close to 0 represent predictions made at random, and −1 represents an opposite prediction where all positive samples were predicted as negative and *vice versa*. The MCC is often normalized to the range of [0, 1], referred to as the normalized coefficient (normMCC), to match the value range and meaning of the other statistical rates. In evaluating binary classifications, MCC can produce a more informative and truthful score compared to accuracy and F1 score. In previous work,[41] the mathematical properties and use cases of MCC were explained and presented, establishing its preference over accuracy and F1 score as the standard metric for evaluating binary classification tasks. There has even been a suggestion that MCC should replace the AUC-ROC as the standard metric for assessing binary classification.[42] In this paper, we used specificity, sensitivity, accuracy, and normMCC as evaluation metrics. The formulas of MCC and normMCC are listed in Equations III and IV.

$$C = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \quad \text{(III)}$$

$$\text{normMCC} = \frac{MCC+1}{2} \quad \text{(IV)}$$

In Equation IV, 0 represents the worst and minimum value, whereas 1 represents the best and maximum value.

### 4.4. Classical ML methods

Common ML techniques such as SVM, RF, GB, MLP, and LR were used in this work as weak learners to build submodels. Embedding these ML techniques into the pipeline of the proposed system can help with decision-making. Doctors can use it to recognize patterns of ADs, and distinguish an AD patient from a healthy patient. The accuracy of the results can be optimized, and adequate treatment can be provided. We briefly explore these ML methods below.

RF takes the mean of the results of a number of distinct decision trees, which work collectively as a group. A majority voting method is used to make the model's final prediction. Voting or averaging mechanisms may deal with the problem of overfitting. It performs well with missing values.

SVM is a supervised ML algorithm, which is good at finding the optimal decision boundary to best separate the hyperplane via linear separation. SVM transforms the input space into a high-dimensional feature space so that the non-linear problem can be solved. SVM is more accurate than other ML methods and is less likely to suffer from overfitting issues, suitable for modeling complex non-linear decision domains. However, SVMs are not suitable for larger datasets and are more sensitive to missing data.

LR predicts a dependent data variable by analyzing the relationship between one or more existing independent variables and helps model a binary dependent variable. LR is easier to interpret, implement, and very quick to train, with no assumptions needed to be made about distributions of classes in the entire feature space. Since linear boundaries are constructed, it is necessary to assume that there is a linear relationship between the independent variable and the dependent variables.

MLP is a fully connected multilayer neural network. It has three layers, including one hidden layer. It is an integral part of a deep neural network. When the number of hidden layers is more than one, they are called deep neural networks. MLP can model complex non-linear relationships and handle various types of data. However, MLP is sensitive to feature scaling; several hyperparameters need to be tuned, such as the number of hidden neurons, layers, and iterations.

GB is a high-performance algorithm that is mainly used for ML sorting or classification. It is generally more accurate compared to other models, but it may require more resources and time compared to simpler ML methods. In the case of high learning rates and complex models, GB can be prone to the overfitting issue, often considered a black box model that is less interpretable.

**Table 4. Predictive performance of the proposed system with various weak learners for anxiety disorder using 83 variables of the Trinity-Ulster-department of Agriculture dataset**

| Models | Metrics | | | | Voting |
|---|---|---|---|---|---|
| | TPR (%) | TNR (%) | Accuracy (%) | normMCC | *Th* |
| RF | 63.0 | 87.8 | 86.5 | 0.6585 | >17 |
| SVM | 58.0 | 89.1 | 87.5 | 0.6543 | >17 |
| GB | 65.4 | 90.5 | 89.2 | 0.6885 | >17 |
| MLP | 60.5 | 89.8 | 88.2 | 0.6668 | >17 |
| LR | 59.3 | 87.7 | 86.2 | 0.6473 | >17 |

Abbreviations: GB: Gradient Boosting; LR: Linear regression; MLP: Multilayer Perceptron; normMCC: Normalized Matthew's correlation coefficient; RF: Random Forest; SVM: Support vector machine; *Th*: Threshold; TNR: True negative rate; TPR: True positive rate.

### 4.5. Performance of the system

In this paper, we used the HADs score as the outcome variable. The ratio of the two classes is approximately 95:5, indicating an extremely imbalanced dataset. The system proposed in section 4.2. is applied to address the class imbalance. To ensure consistency in the comparisons using various weak learners, every weak learner follows the same procedure and is given the same subsets of data for the training and test sets.

Table 4 lists the predictive performance of the proposed system with weak learners of RF, SVM, GB, MLP, and LR, respectively, with a decision threshold taken as 17. The linear kernel function was used in SVM. In MLP, 200 hidden neurons were used with the adaptive moment estimation (Adam) optimizer; for both MLP and LR, 0.001 was the initial learning rate, and the mini-batch size was taken as 25. The number of decision trees was 300 in RF. Other parameters were employed in the standard procedures. Figures 12-16 illustrate the predictive performance in terms of accuracy, TPR (sensitivity), TNR (specificity), and normMCC against various decision thresholds with various weak learners.

The major difference between oversampling techniques and SMOTE is that SMOTE produces synthetic samples by interpolating samples among existing minority samples. However, oversampling techniques replicate existing minority samples to make the dataset balanced. Due to the extremely imbalanced data, SMOTE can cause problems when generating a large number of minority samples. Therefore, we opted to use an oversampling technique by replicating the existing minority samples in the training set 20 times. This approach resulted in a roughly balanced training set for model training.

Table 5 compares the performance in terms of sensitivity, specificity and normMCC between the oversampling

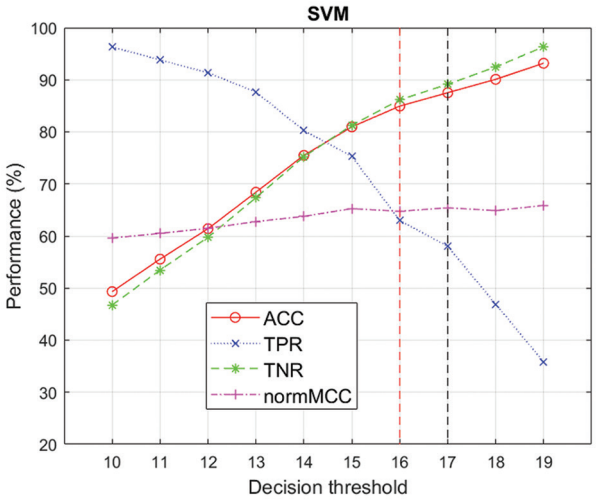



**Figure 12.** Predictive performance in terms of accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and normalized Matthew's correlation coefficient (normMCC) against decision threshold with support vector machine (SVM) weak learner

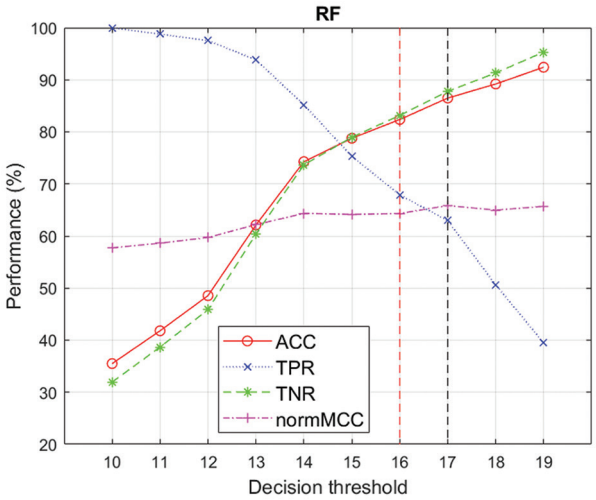



**Figure 13.** Predictive performance in terms of accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and normalized Matthew's correlation coefficient (normMCC) with random forest (RF) weak learner against decision threshold

method with repetition and the proposed system when the decision threshold is taken as 17, and RF, SVM, GB, MLP, and LR ML methods are used as weak learners, respectively, for ADs diagnosis. For fair comparison, each result used the same 83 predictor variables of the TUDA dataset.

From Table 5 , we can see that the proposed ensemble system achieves better performance than the base approach using the embedded oversampling technique. The proposed homogeneous ensemble system that used a single-base ML model across all submodels indeed increased reliability and stability in predictions compared

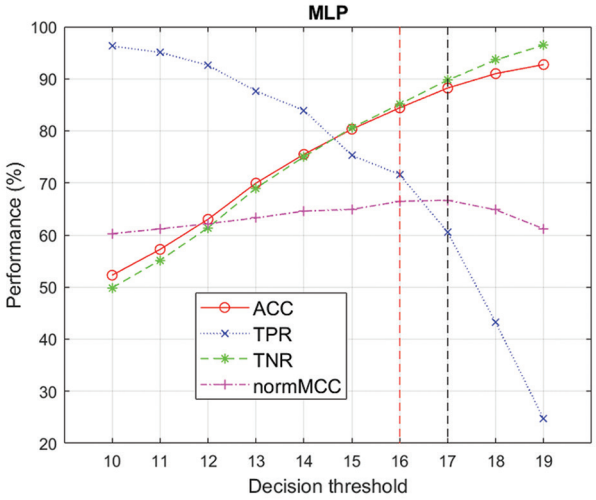



**Figure 14.** Predictive performance in terms of accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and normalized Matthew's correlation coefficient (normMCC) with a multilayer perceptron (MLP) weak learner against a decision threshold.

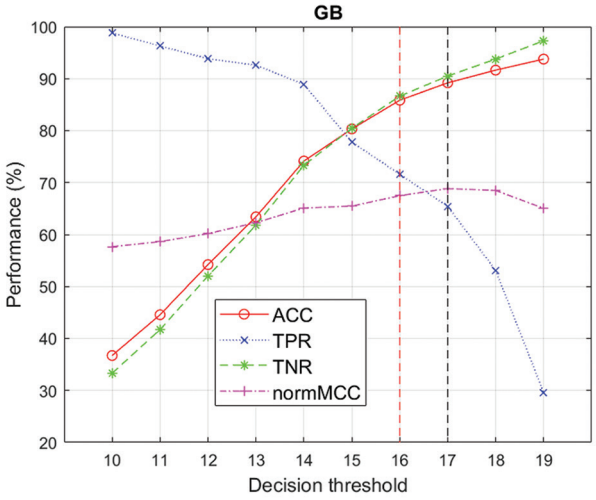



**Figure 15.** Predictive performance in terms of accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and normalized Matthew's correlation coefficient (normMCC) with gradient boosting (GB) weak learner against decision threshold.

to using the oversampling technique. The numMCC value of the proposed system with GB as a weak learner is 0.6885, which is the highest among these five approaches.

In Figures 17-19, we use an interpretability tool, such as the Shapley function in Matlab, to interpret the prediction of a testing instance using a trained RF, GB, and SVM submodel, respectively. Shapley values are calculated to show how much each predictor variable contributes to that prediction.

Figure 20 shows the 30 most important predictor variables when a trained RF submodel was used to predict

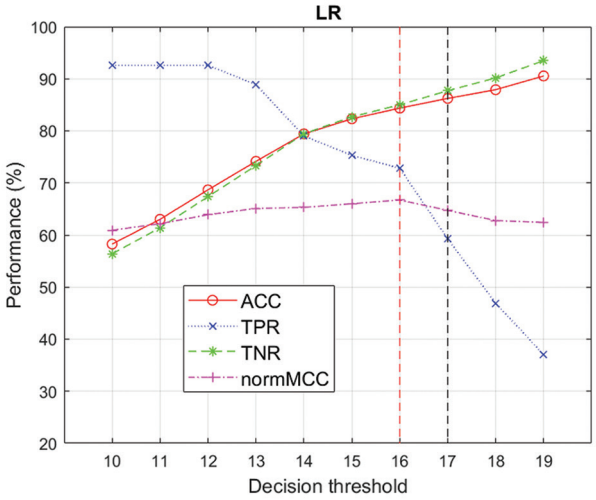



**Figure 16.** Predictive performance in terms of accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and Matthew's correlation coefficient (normMCC) with linear regression (LR) weak learner against decision threshold

the test set. Note that variables relevant to anxiety diagnosis, depression diagnosis, total score of depression scale, and age were among the top 10 most important variables.

## 5. Discussion

In this paper, the imbalanced HADS variable was the outcome variable in data from the TUDA study. The results show that the proposed adapted bagging ensemble system, using a variety of sociodemographic, lifestyle, clinical, and biochemical factors, may effectively predict ADs in older adults with a high degree of accuracy. An accuracy of 89.2% (sensitivity: 65.4%, specificity: 90.5%) was achieved with the GB weak learner method when the decision threshold was set to more than 17. In this context, the sensitivity for the ADs class was 65.4% when the decision threshold was more than 17, meaning that 65.4% of true ADs diagnoses are correctly identified as ADs. Therefore, 34.6% (1–65.4%) of true AD diagnoses are unfortunately missed. For fairness, the oversampling method was investigated on the same dataset, and the results show that the proposed bagging ensemble system achieved significant improvements over the oversampling method. The threshold-moving strategy was adopted to add the predictions of multiple submodels for the final prediction, which reduced the sensitivity of each submodel to outliers or noise and avoided overfitting. Compared with using a single model, better generalization performance and higher accuracy can be achieved.

To enhance the sensitivity of ADs (the TPR) and reduce AD losses (minimize FNR), a threshold-moving strategy can be employed. By lowering the decision threshold to >16,

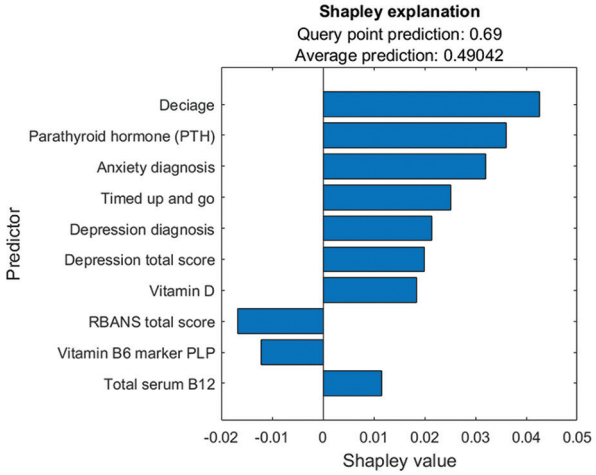



**Figure 17.** Shapley's explanation of the top 10 predictors predicted for a testing instance using a random forest (RF) submodel
Abbreviations: PLP: Pyridoxal phosphate; RBANS: Repeatable battery for the assessment of neuropsychological status.

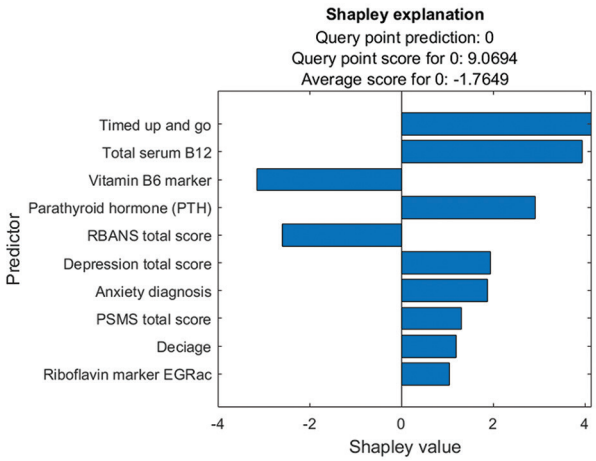



**Figure 18.** Shapley's explanation of the top 10 predictors predicted for a testing instance using a gradient boosting (GB) submodel
Abbreviations: EGRac: Erythrocyte glutathione reductase activation coefficient; PSMS: Physical self-maintenance scale; RBANS: Repeatable battery for the assessment of neuropsychological status.

we can identify instances as ADs when the summarized predictive value surpasses the decision threshold. Table 3 presents the results for the decision threshold values taken as more than 17. The results indicate that the proposed system achieved an accuracy of 86.5% (sensitivity: 63.0% and specificity: 87.8%) with the decision threshold taken at >17, using RF as the weak learner. The results indicate that models incorporating a combination of features, including nutrition, health, clinical, biochemical, and lifestyle factors, should be encouraged. For GB as a weak learner, we have reduced the false negative rate from 34.6%

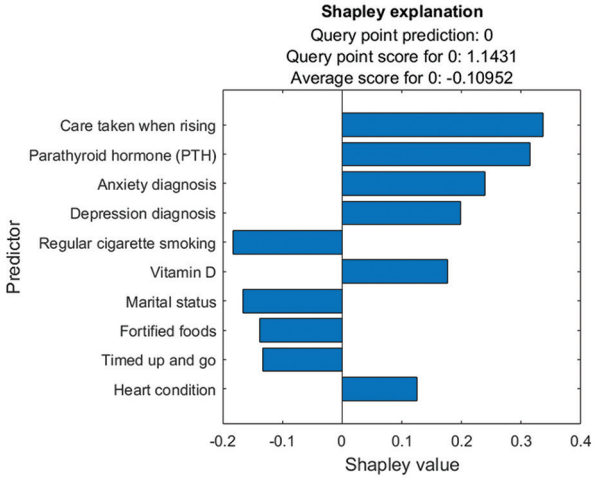



**Figure 19.** Shapley's explanation of the top 10 predictors predicted for a testing instance using a support vector machine submodel

down to 28.4% whereas the decision threshold changed from more than 17 to more than 16. The trade-off of misclassifying approximately 3.9% of the normal class may be deemed worthwhile to correctly classify more than 6.2% of the ADs.

As a trade-off, the number of FP will inevitably increase as we adjust to decrease the decision threshold that we apply to the model's prediction. To elucidate this trade-off and assist in threshold selection, Figures 12-16 illustrate the predictive performance in terms of the TPR and TNR against various decision thresholds. This approach is a variant of the ROC curve, with a focus on stressing the decision threshold.

Several potential risk factors for a diagnosis of ADs were identified. Understanding risk factors for ADs in people with specific chronic diseases can aid health-care professionals in immediately identifying at-risk patients, allowing improved screening activities for psychological assessment and the introduction of personalized treatments within the care settings for specific illnesses. In the long term, it will be crucial to assess the impact of real-time feedback and identify specific triggers that lead to inappropriate and high levels of anxiety.

In this study, potential risk factors were identified. Older people who have a marital status of separated/divorced or widow/widower seem to be more prone to anxiety problems. Other risk factors for ADs were identified, including areas of higher deprivation, being female, smoking, and alcohol drinking in the past. Vitamin D and fortified food/Vitamin B supplements intakes may be beneficial to ADs. Furthermore, consistent with evidence from other large cohort studies, the variables including lack of formal education,[7] functional and cognitive

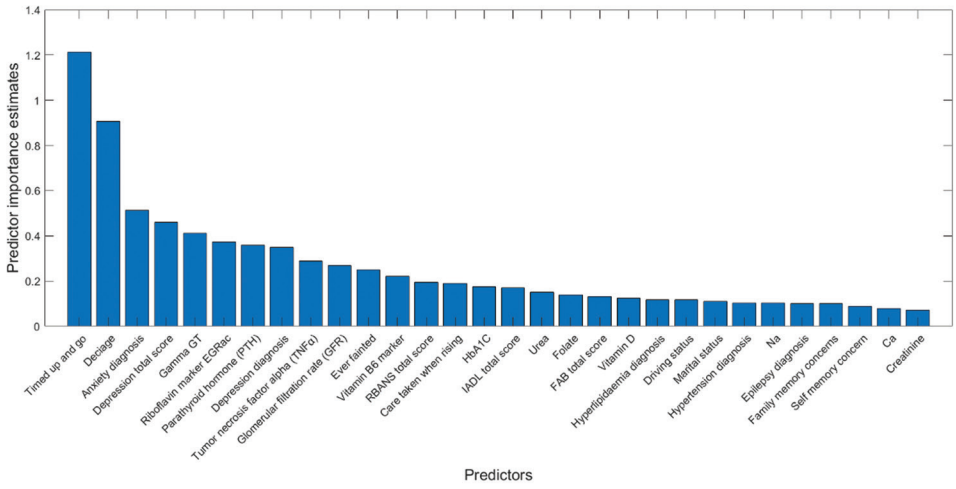



**Figure 20.** The 30 most important predictor variables using a random forest (RF) submodel
Abbreviations: EGRac: Erythrocyte glutathione reductase activation coefficient; FAB: Frontal assessment battery; Gamma GT: Gamma-glutamyl transferase; IADL: Instrumental activities of daily living; RBANS: Repeatable battery for the assessment of neuropsychological status.

**Table 5. Performance comparison between the proposed system and oversampling method**

| Models | Metrics | | | | | | pa |
|---|---|---|---|---|---|---|---|
| | Oversampling | | | Proposed | | | |
| | TPR (%) | TNR (%) | normMCC | TPR (%) | TNR (%) | normMCC | |
| RF | 4.94 | 99.9 | 0.5861 | 63.0 | 87.8 | 0.6585 | <2.5e-16**** |
| SVM | 79.0 | 76.2 | 0.6387 | 58.0 | 89.1 | 0.6543 | <2.2e-16**** |
| GB | 18.5 | 98.6 | 0.6285 | 65.4 | 90.5 | 0.6885 | <2.2e-16**** |
| MLP | 28.4 | 96.5 | 0.6301 | 60.5 | 89.8 | 0.6668 | <2.2e-16**** |
| LR | 81.5 | 75.8 | 0.6430 | 59.3 | 87.7 | 0.6473 | <2.2e-16**** |

Note: **** *p*-value indicating that Kruskal–Wallis's test rejects the null hypothesis at the 0.01% significance level.
Abbreviations: GB: Gradient boosting; LR: Linear regression; MLP: Multilayer perceptron; normMCC: Normalized Matthew's correlation coefficient; RF: Random forest; SVM: Support vector machine; TNR: True negative rate; TPR: True positive rate.

impairment,[7] poor quality of life,[19] low status of folate and metabolically related B vitamins, namely Vitamin B6, Vitamin B12, and riboflavin deficiencies have been identified as risk factors for ADs among older people.[27] It has also been considered that fortified foods could play a role in optimizing B vitamin status and potentially reduce the risk of these mental health disorders.[27] An intriguing finding is the association between Vitamin D and fortified food/Vitamin B supplements and ADs, indicating their potential value as contributing factors.

This developed system trains multiple models from generated multiple subsets of the training data, which can reduce variance, mitigate overfitting, and improve the ability of generalization of the system. The training process is scalable because each submodel is trained independently, and the developed system has the ability of expansion to handle increasing amounts of data efficiently. However, compared to training a single model, training multiple submodels may increase computational overhead for large datasets; the aggregated prediction output can reduce interpretability, too, and the reasoning behind a specific prediction can be hard to understand. In addition, there is a risk of the diluted insights if a rare but important pattern is captured by a single submodel; the impact might be diluted by aggregating it with other submodels. The choice of base model affects the performance of the system. The limitations also include that the observational studies are subject to bias due to their inherent flaws; there were fewer positive cases in the training set, resulting in a decrease in the ability of the system to predict the occurrence of positive cases. The findings of this study, focusing on a specific cohort, are not generalizable to other racial groups. The developed system is able to learn from different subsets of data and reduce variance, making it suitable for various external

applications, such as those in clinical settings. Identified risk factors, such as the link between certain lifestyle factors and ADs, may have clinical implications and help health-care professionals invest more effort in educating people about these factors. The system could help doctors identify and treat patients earlier and even improve treatment outcomes.

In Figures 17-19, we show Shapley's explanation of the top 10 predictors predicted for the same testing instance using RF, GB, and SVM submodels trained with the same subdataset, respectively. The interpretability of each trained RF, GB, and SVM submodel varied, with predictors such as anxiety diagnosis, depression diagnosis, total score of depression scale, age, and timed up and go, ranking among the top ten factors contributing most to the prediction.

ML algorithms such as RF, MLP, SVM, GB, and LR are widely used and typically achieve good performance.[18] The success of these algorithms depends on the quality and quantity of features used, as well as the characteristics of the available dataset. However, when dealing with extremely imbalanced datasets, using these methods directly for specificity in the disadvantaged class can lead to poor performance. In this study, we demonstrated how to adapt an ensemble bagging ML method to deal with imbalanced classes. Through this adaptation, satisfactory performance was achieved.

## 6. Conclusion

ML technologies have become increasingly popular for disease diagnosis in the field of mental health. ADs are one of the major health burdens facing older adults worldwide. Despite evidence[43-46] that the impact of ADs can be reduced through prevention and intervention, the prevalence remains high worldwide, highlighting that people with such disorders need intervention.[47-49] The widespread prevalence of anxiety-related disorders presents many challenges for mental health-care providers, who find it difficult to provide face-to-face treatments to those who need it in a timely manner. As more complex health data is becoming available, ML can deal with data that can be from multiple sources, predict the risk of developing ADs by identifying individual characteristics and risk factors, and perform mental health diagnoses. ML can help personalize treatment plans and ensure that individuals receive the most appropriate care. In this study, an adapted bagging ensemble approach was proposed to identify ADs, in which traditional ML methods act as weak learners. In future work, the goal is to identify ADs in its earlier or prodromal stage, when interventions may be more effective, and treatments can be personalized based on individuals' unique characteristics. This could pave the way for developing a "mental health status indicator" to monitor an individual's mental health, establish different alert levels, and efficiently address emerging issues.

In this study, key predictors were identified that could effectively predict ADs using the proposed learning system, achieving a satisfactory level of accuracy. Some variables were determined to be closely associated with an increased risk of ADs, such as gender, marital status, accommodation status, lifestyle-related factors, quality of life, fortified food/supplements intake level, and family history of certain diseases and chronic diseases. Efforts were made to assess risk factors of anxiety in older adults. More studies are needed to fully understand the characteristics of anxiety in this population.

Future works involve conducting a longitudinal study by studying the individuals in the TUDA dataset over the years to observe how they develop over time, to understand physical and cognitive developmental processes, and to predict future development of ADs. The long-term consequences of interventions can be investigated.

ADs and depression are commonly found to coexist. Participants with one condition were generally at higher risk for the other condition. Future studies should explore the relationship between ADs and depression, especially given that older adults often suffer from comorbidities. Maintaining mental health is crucial, not only to improve daily functioning, strengthen relationships, and enhance self-image, but also to address physical health issues related to mental health conditions. It could help reduce the prevalence of ADs by supporting older adults' participation in physical activities and reducing social isolation.[49] Encouraging proper intake of fortified food and supplements to prevent deficiencies of necessary vitamins is also essential. Given the burden on health-care resources, these efforts may promote inclusivity in policymaking, especially in identifying strategies in the public health field to promote reduced inequalities and improved mental health.

## Acknowledgments

The authors would like to acknowledge the support of all who are involved in the AIM4HEALTH programme.

## Funding

The TUDA study was supported by government funding from the Irish Department of Agriculture, Food and the Marine, and Health Research Board (under the Food Institutional Research Measure), as well as from the Northern Ireland Department for Employment and Learning (under its Strengthening the All-Island Research

Base Initiative). The AIM4HEALTH project gratefully acknowledges the support of the higher education authority, Department of Further and Higher Education, Research, Innovation and Science, and the Shared Island Fund, and the SFI grant 21/RC/10295_P2.

## Conflict of interest

The authors declare they have no competing interests.

## Author contributions

*Conceptualization:* Jinling Wang, Michaela Black, Debbie Rankin
*Formal analysis:* Jinling Wang
*Investigation:* Catherine F. Hughes, Leane Hoey, Geraldine Horigan, Helene McNulty, Anne M. Molloy
*Methodology:* Jinling Wang, Michaela Black, Debbie Rankin, Catherine F. Hughes, Leane Hoey, Anne M. Molloy, Helene McNulty, Mimi Zhang
*Writing–original draft:* Jinling Wang, Debbie Rankin
*Writing–review & editing:* All authors

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

The data will not be shared due to some concerns. For more data information, please refer to the relevant research works.[7,26-28]

## References

1. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *Lancet*. 2021;398:1700-1712.

   doi: 10.1016/S0140-6736(21)02143-7

2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: A systematic analysis for the global burden of disease study 2019. *Lancet Psychiatry*. 2022;9:137-150.

   doi: 10.1016/S2215-0366(21)00395-3

3. Lauderdale SA, Sheikh JL. Anxiety disorders in older adults. *Clin Geriatr Med*. 2003;19(4):721-741.

   doi: 10.1016/s0749-0690(03)00047-8

4. Sheikh JI. Investigations of anxiety in older adults: Recent advances and future directions. *J Geriatr Psychiatry Neurol*. 2005;18(2):59-60.

   doi: 10.1177/0891988705276253

5. Andreescu C, Varon D. New research on anxiety disorders in the elderly and an update on evidence-based treatments. *Curr Psychiatry Rep*. 2015;17(7):53.

   doi: 10.1007/s11920-015-0595-8

6. Ishikawa RZ, Vyas C, Okereke O. Anxiety disorders among older adults: Empirically supported treatments and special considerations. In: Bui E, Charney ME, Baker AW, editors. *Clinical Handbook of Anxiety Disorders: From Theory to Practice*. United States: Humana Press/Springer Nature; 2020. p. 175-189.

   doi: 10.1007/978-3-030-30687-8_9

7. Rankin D, Black M, Flanagan B, *et al*. Identifying key predictors of cognitive dysfunction in older people using supervised machine learning techniques: Observational study. *JMIR Med Inform*. 2020;8(9):e20995.

   doi: 10.2196/20995

8. Javaid SF, Hashim IJ, Hashim MJ, Stip E, Samad MA, Ahbabi AA. Epidemiology of anxiety disorders: Global burden and sociodemographic associations. *Middle East Curr Psychiatry*. 2023;30:44.

   doi: 10.1186/s43045-023-00315-3

9. Fusar-Poli P, Correll CU, Arango C, Berk M, Patel V, Ioannidis JP. Preventive psychiatry: A blueprint for improving the mental health of young people. *World Psychiatry*. 2021;20:200-21.

   doi: 10.1002/wps.20869

10. Jorm AF, Patten SB, Brugha TS, Mojtabai R. Has increased provision of treatment reduced the prevalence of common mental disorders? Review of the evidence from four countries. *World Psychiatry*. 2017;16:90-99.

    doi: 10.1002/wps.20388

11. Jain PR, Quadri SMK. Emerging role of intelligent techniques for effective detection and prediction of mental disorders. In: Hemanth J, Bestak R, Chen JIZ, editors. *Intelligent Data Communication Technologies and Internet of Things. Lecture Notes on Data Engineering and Communications Technologies*. Vol. 57. Singapore: Springer; 2021.

    doi: 10.1007/978-981-15-9509-7_16

12. Meehan AJ, Lewis SJ, Fazel S, *et al*. Clinical prediction models in psychiatry: A systematic review of two decades of progress and challenges. *Mol Psychiatry*. 2022;27:2700-2708.

    doi: 10.1038/s41380-022-01528-4

13. Graham S, Depp C, Lee EE, *et al*. Artificial intelligence for mental health and mental illnesses: An overview. *Curr Psychiatry Rep*. 2019;21:116.

    doi: 10.1007/s11920-019-1094-0

14. Cearns M, Hahn T, Baune BT. Recommendations and future

directions for supervised machine learning in psychiatry. *Transl Psychiatry*. 2019;9:271.

doi: 10.1038/s41398-019-0607-2

15. Thieme A, Belgrave D, Doherty G. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ml systems. *ACM Trans Comput Hum Interact*. 2020;27(5):1-53.

doi: 10.1145/3398069

16. Ancillon I, Elgendi M, Menon C. Machine learning for anxiety detection using biosignals: A review. *Diagnostics (Basel)*. 2022;12(8):1794.

doi: 10.3390/diagnostics12081794

17. Khan A, Husain MH, Khan A. Analysis of mental state of users using social media to predict depression: A survey. *Int J Adv Res Comput Sci*. 2018;9:100-106.

doi: 10.26483/ijarcs.v9i0.6146

18. Agarwal D, Singh V, Singh AK, Madan P. Stacked ensemble model for analyzing mental health disorder from social media data. *Multimed Tools Appl*. 2023;83:53923-53948.

doi: 10.1007/s11042-023-17395-2

19. Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Sci Rep*. 2021;11(1):1980.

doi: 10.1038/s41598-021-81368-4

20. Shen ZX, Cui LJ, Mou SQ, *et al*. Combining S100B and cytokines as neuro-inflammatory biomarkers for diagnosing generalized anxiety disorder: A proof-of-concept study based on machine learning. *Front Psychiatry*. 2022;13:881241.

doi: 10.3389/fpsyt.2022.881241

21. Byeon H. Exploring factors for predicting anxiety disorders of the elderly living alone in South Korea using interpretable machine learning: A population-based study. *Int J Environ Res Public Health*. 2021;18(14):7625.

doi: 10.3390/ijerph18147625

22. Henry M, Isa SM. Mental health treatment prediction for Tech Employee with the implementation of ensemble methods. *J Theor Appl Inf Technol*. 2022;100(8):2675-2685.

23. Rocca J. *Ensemble Methods: Bagging, Boosting and Stacking--Understanding the Key Concepts of Ensemble Learning*; 2019. Available from: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205. [Last accessed on 2024 Jan 27].

24. Patel A. *Ensemble Learning- the Heart of Machine Learning*. Available from: https://medium.com/ml-research-lab/ensemble-learning-the-heart-of-machine-learning-b4f59a5f9777 [Last accessed on 2020 Jan 03].

25. Chicco D, Tötsch N, Jurman G. The matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min*. 2021;14(1):13.

doi: 10.1186/s13040-021-00244-z

26. McCann A, McNulty H, Rigby J, *et al*. Effect of area-level socioeconomic deprivation on risk of cognitive dysfunction in older adults. *J Am Geriatr Soc*. 2018;66(7):1269-1275.

doi: 10.1111/jgs.15258

27. Moore K, Hughes CF, Hoey L, *et al*. B-vitamins in relation to depression in older adults over 60 years of age: The trinity ulster department of agriculture (TUDA) cohort study. *J Am Med Dir Assoc*. 2019;20(5):551-557.e1.

doi: 10.1016/j.jamda.2018.11.031

28. Wang J, Black M, Rankin D, *et al*. Analysis of Risk Factors and Diagnosis for Anxiety Disorder in Older People with the Aid of Artificial Intelligence: Observational Study. In: *2023 the 31st Irish Conference on Artificial Intelligence and Cognitive Science,* Letterkenny, Ireland, IEEE. p. 1-8.

doi: 10.1109/aics60730.2023.10470782

29. Larsen BS. *Synthetic Minority Over-Sampling Technique (SMOTE)*. Available from: https://github.com/dkbsl/matlab_smote/releases/tag/1.0),github [Last accessed on 2023 May 31].

30. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.

doi: 10.1613/jair.953

31. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In: *2008 IEEE International Joint Conference on Neural Networks*; 2008. p. 1322-1328.

doi: 10.1109/ijcnn.2008.4633969

32. Edwards R. *Causes and Risk Factors of Anxiety*; 2021. Available from: https://www.verywellhealth.com/anxiety-causes-and-risk-factors-5191778 [Last accessed on 2023 Dec 03].

33. Narmandakh A, Roest AM, De Jonge P, *et al*. Psychosocial and biological risk factors of anxiety disorders in adolescents: A TRAILS report. *Eur Child Adolesc Psychiatry*. 2021;30:1969-1982.

doi: 10.1007/s00787-020-01669-3

34. *UK Statistics on Vitamin and Mineral Deficiency*; 2023. Available from: https://vitall.co.uk/health-tests-blog/statistics-vitamin-mineral-deficiency-uk [Last accessed on 2023 Aug 23].

35. *Vitamin D: The Connection to Depression and Anxiety*. Available from: https://montarebehavioralhealth.com/vitamin-d-the-connection-to-depression-and-anxiety [Last accessed on 2023 Aug 23].

36. Chang S, Lee H. Vitamin D and health - the missing vitamin in humans. *Pediatr Neonatol*. 2019;60(3):237-244.

doi: 10.1016/j.pedneo.2019.04.007

37. Menon V, Kar SK, Suthar N, Nebhinani N. Vitamin D and depression: A critical appraisal of the evidence and future directions. *Indian J Psychol Med*. 2020;42(1):11-21.

doi: 10.4103/ijpsym.ijpsym_160_19

38. Kowalówka M, Gówka AK, Karaniewiczada M, Kosewski G. Clinical significance of analysis of vitamin D status in various diseases. *Nutrients*. 2020;12(9):2788.

doi: 10.3390/nu12092788

39. Chiang JJ, Park H, Almeida DM, *et al*. Psychosocial stress and C-reactive protein from mid-adolescence to young adulthood. *Health Psychol*. 2019;38(3):259-267.

doi: 10.1037/hea0000701

40. Anjum I, Jaffery SS, Fayyaz M, Samoo Z, Anjum S. The role of vitamin D in brain health: A mini literature review. *Cureus*. 2018;10(7):e2960.

doi: 10.7759/cureus.2960

41. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(6):6.

doi: 10.1186/s12864-019-6413-7

42. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min*. 2023;16(1):4.

doi: 10.1186/s13040-023-00322-4

43. Adeniji OD, Adeyemi SO, Ajagbe SA. An improved bagging ensemble in predicting mental disorder using hybridized random forest - artificial neural network model. *Informatica*. 2022;46(4):543-550.

doi: 10.31449/inf.v46i4.3916

44. Ogunseye EO, Adenusi CA, Nwanakwaugwu AC, Ajagbe SA, Akinola SO. Predictive analysis of mental health conditions using adaboost algorithm. *Paradigmplus*. 2022;3:11-26.

doi: 10.55969/paradigmplus.v3n2a2

45. Alabi EO, Adeniji OD, Awoyelu TM, Fasae EO. Hybridization of machine learning techniques in predicting mental disorder. *Int J Hum Comput Stud*. 2021;3(6):22-30.

doi: 10.31149/ijhcs.v3i6.2083

46. Obiedat R, Toubasi SA. A combined approach for prediction employee's productivity based on ensemble machine learning methods. *Int J Comput Inform*. 2022;46:49-58.

doi: 10.31449/inf.v46i5.3839

47. Moreno C, Wykes T, Galderisi S, *et al*. How mental health care should change as a consequence of the COVID-19 pandemic. *Lancet Psychiatry*. 2020;7:813-824.

doi: 10.1016/S2215-0366(20)30307-2

48. Campion J, Javed A, Sartorius N, Marmot M. Addressing the public mental health challenge of COVID-19. *Lancet Psychiatry*. 2020;7(8):657-659.

doi: 10.1016/S2215-0366(20)30240-6

49. Baxter AJ, Scott KM, Vos T, Whiteford HA. Global prevalence of anxiety disorders: A systematic review and meta-regression. *Psychol Med.* 2013;43(5):897-910.

doi: 10.1017/S003329171200147X

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

# Large language models-in-the-loop: Leveraging expert small artificial intelligence models for multilingual anonymization and de-identification of protected health information

**Murat Gunay[1]*, Bunyamin Keles[2†], and Raife Hizlan[1†]**

[1]Department of Research and Development, AI Handed LLC, Lewes, Delaware, United States of America

[2]Department of Health Management, Hacettepe University Institute of Social Sciences, Ankara, Turkey

## Abstract

The rise of chronic diseases and pandemics, such as COVID-19 has emphasized the need for effective patient data processing while ensuring privacy through anonymization and de-identification of protected health information. Anonymized data facilitates research without compromising patient confidentiality. This paper introduces expert small artificial intelligence (AI) models developed using the large language model (LLM)-in-the-loop methodology to meet the demand for domain-specific de-identification of named entity recognition (NER) models. These models overcome the privacy risks associated with LLMs used through application programming interfaces by eliminating the need to transmit or store sensitive data. More importantly, they consistently outperform LLMs in de-identification tasks, offering superior performance and reliability. Our de-identification NER models, developed in eight languages—English, German, Italian, French, Romanian, Turkish, Spanish, and Arabic—achieved F1-macro score averages of 0.931, 0.960, 0.955, 0.937, 0.930, 0.963, 0.957, and 0.922, respectively. These results establish our de-identification NER models as the most accurate healthcare anonymization solutions, surpassing existing small models and even general-purpose LLMs, such as GPT-4o. While Part I of this series introduced the LLM-in-the-loop methodology for biomedical document translation, this second paper showcases its success in developing cost-effective expert small NER models in de-identification tasks. Our findings lay the groundwork for future healthcare AI innovations, including biomedical entity and relation extraction, demonstrating the value of specialized models for domain-specific challenges.

***Keywords:*** De-identification; Health Insurance Portability and Accountability Act; Protected health information; Patient safety; Large language models-in-the-loop; Anonymization

†*These authors contributed equally to this work.*

***Corresponding author:**
Murat Gunay
(murat.esra.gunay@gmail.com)

**Citation:** Gunay M, Keles B, Hizlan R. Large language models-in-the-loop: Leveraging expert small artificial intelligence models for multilingual anonymization and de-identification of protected health information. *Artif Intell Health*. 2026;3(1):138-151. doi: 10.36922/AIH025120021

**Received:** March 19, 2025

**1st revised:** June 19, 2025

**2nd revised:** July 2, 2025

**3rd revised:** July 22, 2025

**4th revised:** August 21, 2025

**Accepted:** August 26, 2025

**Published online:** September 19, 2025



## 1. Introduction

Patient data are essential for improving public health, expanding preventive health services, preventing diseases, and formulating necessary health policies. Recent studies show that almost all (99%) hospitals in the United States[1] use electronic health records

(EHR). Similarly, Wales, Scotland, Denmark, and Sweden have implemented EHRs in the past few years. In contrast, the United Kingdom still lacks a nationally accessible health data system, despite the COVID-19 pandemic underscoring the critical importance of EHR data.[2] EHRs enable the examination of disease trends, support predictive modeling, and inform the development of health policies.

Technology, which has become more complex and has developed with medical practices, necessitates the development of methods that will protect patient privacy.[3] With information security and information leakage recently gaining more importance, patient safety may have significant consequences beyond ethical violations in fundamental and health law.[4] Personal data are sensitive information associated with an individual and protected by various laws.[5] Personal privacy data in healthcare, known as protected health information (PHI), includes private information, such as a patient's health history, treatments received, and more.[6]

EHRs contain both valuable clinical information and PHI. While EHRs are a rich data source for research, their usability is restricted due to the confidentiality of PHI.[7-10] For example, the Health Insurance Portability and Accountability Act (HIPAA) regulates the use of 18 types of PHI, such as name, phone number, dates, and more (Table 1).[11,12] Therefore, PHI must be extracted from the text before EHR data can be used. Automating de-identification systems is needed since manually extracting PHI is time-consuming and costly. In addition, coordination between annotators is also an important consideration.[13,14] While early approaches to de-identification relied on complex rules to detect PHI, recent developments use machine learning (ML) methods and train on expert-annotated records. Hybrid systems integrate practices as features into statistical models, such as conditional random fields (CRFs).[15]

The de-identification method makes it possible to use EHRs in research by removing confidential information.[15] Basic de-identification rules include removing direct identifying statements, such as name, date, and more. Advanced statistical methods anonymize the data, reducing de-identification risk.[16] However, new techniques may also introduce unknown privacy risks. Therefore, continuous evaluation and improvement efforts are necessary.[17] Advanced methods can enable extensive collections of EHRs to be used efficiently and securely in research.

According to HIPAA, there are two possible methods of identity masking. The "Expert Determination" method, which requires employing an expert in the field, involves a small risk in identifying the individual whose information is used and is performed using different statistical methods. In this method, the expert must have sufficient experience and knowledge. The other method is the "Safe Harbor" method, which involves de-identifying 18 pre-determined relevant identifiers that must be removed and/or modified from the corpus.[11,18] In studies using deep learning (DL) models, the Safe Harbor method is used, and the relevant PHI is de-identified. The de-identification process was primarily guided by the HIPAA framework for protecting patient health information. For non-English languages, entity definitions were carefully adapted to reflect HIPAA requirements and global privacy standards, such as the General Data Protection Regulation for European languages. Where relevant, local legal and regulatory frameworks were considered to ensure our approach aligns with both United States and international best practices for data privacy.

**Table 1. Types of protected health information**

| No | Type of protected health information |
|---|---|
| 1 | Names |
| 2 | All geographic subdivisions smaller than a state |
| 3 | Dates |
| 4 | Telephone numbers |
| 5 | Vehicle identifiers |
| 6 | Fax numbers |
| 7 | Device identifiers and serial numbers |
| 8 | Emails |
| 9 | Uniform resource locators |
| 10 | Social security numbers |
| 11 | Medical record numbers |
| 12 | Internet protocol addresses |
| 13 | Biometric identifiers |
| 14 | Health plan beneficiary numbers |
| 15 | Full-face photographic images and any comparable images |
| 16 | Account numbers |
| 17 | Certificate/license numbers |
| 18 | Any other unique identifying number, characteristic, or code |

The lack of comprehensive data privacy frameworks can lead to vulnerabilities, leaving sensitive patient information susceptible to breaches and misuse. Despite efforts to anonymize this data, reidentification is still feasible through just a few spatiotemporal data points.[19] Recent advancements in privacy-preserving technologies have increased adoption,[20] particularly in artificial intelligence (AI) and big data analytics. These technologies are vital in addressing major global health challenges by enhancing access to healthcare, promoting health, preventing diseases, and improving the overall experience for healthcare professionals and patients. AI, coupled with

big data analytics, is the backbone for many innovations in digital health, driving improvements in care delivery and decision-making processes. These domains are supported by additional technologies, such as the Internet of Things, next-generation networks (e.g., 5G), and privacy-preserving platforms, such as blockchain.[21,22]

However, questions remain regarding the accountability for AI and large language models (LLMs). Given that AI lacks autonomy and sentience, it cannot hold moral responsibility, leaving uncertainty about who should be accountable for its decisions and actions.[23]

LLMs, particularly GPT-4o, have demonstrated remarkable success in zero-shot de-identification tasks. However, significant challenges remain in utilizing proprietary paid application programming interfaces (APIs) and open-source small LLMs. Paid APIs raise privacy concerns, as hospitals are often unwilling to transmit sensitive patient data to cloud-based services. Conversely, locally deployed small LLMs present difficulties in production due to their limited accuracy, resource-heavy requirements, and complex deployment processes. In response to these challenges, we propose expert small AI models—lightweight, plug-and-play solutions that offer superior performance and accuracy compared to LLMs, while being more practical and efficient for deployment in secure, on-premises environments. Hence, we did not directly implement LLMs to address these challenges; instead, we developed expert small named entity recognition (NER) models using the LLM-in-the-loop methodology.

## 2. What is the LLM-in-the-loop?

From our perspective, the "LLM-in-the-loop" methodology is an integral part of the development process for expert small models, without relying on LLMs as the final solution. Instead of directly using LLMs for tasks, we utilize them selectively at various stages, such as synthetic data generation, rigorous evaluation, and agent orchestration, to improve the performance of smaller, domain-specific models. This approach allows us to benefit from the capabilities of LLMs while keeping the models efficient, focused, and specialized for specific tasks.

Recently, there has been a growing emphasis on the work done within the scope of LLM-in-loops. Studies have shown that LLMs perform better on tasks traditionally completed by humans,[24-26] and the potential for effective utilization of LLMs is emphasized. It is seen that the innovative approach of "LLM-in-the-loop" is used in different fields today. In a study conducted to analyze social media content and reveal hidden themes,[27] the advanced capabilities of LLMs were leveraged to gain a deeper understanding of social media content by analyzing social media messages, discovering thematic structures and nuances in texts, and effectively matching texts to themes. Another study using the LLM-in-loops technique to improve the performance of LLMs was aimed at continuously improving the model outputs through iterative feedback loops, and this was applied in a study in the medical field. The aim was to increase the accuracy and reliability of the model and reduce hallucinations. The LLM-in-loops study, which integrated human expert evaluation of model outputs, feedback provision, and subsequent retraining, focused on reducing model errors and obtaining more reliable results in medical question-answering and summarization tasks.[28]

Another study, which examined the potential of LLMs to recognize and examine intertextual relationships in biblical and Koine Greek texts, highlighted how LLMs evaluate different intertextual scenarios and how these models can detect direct quotations, allusions, and echoes between texts. The study also mentioned the ability of LLMs to generate intertextual observations and connections and the potential of these models to reveal new insights. However, it is noted that the model had difficulties with long query texts and can create incorrect intertextual connections, which reveals the importance of expert evaluation.[29]

We first used the "LLMs-in-the-loop" method in the context of biomedical document translation.[30] In the previous work, we demonstrated its success in developing cost-effective expert small NER models for de-identification tasks. Our findings laid the groundwork for future healthcare AI innovations, including biomedical entity and relation extraction, and demonstrated the value of specialized models for domain-specific challenges.

As we navigate the evolving landscape of AI in healthcare, the LLM-in-the-loop methodology stands out as a transformative approach. Recent studies have highlighted its capacity to enhance the performance of the models by leveraging human expertise to refine outputs continuously. This innovative strategy addresses the traditional challenges faced in biomedical text processing, such as accuracy and reliability, and mitigates issues, such as hallucinations that commonly occur in AI-generated content. By fostering a symbiotic relationship between human input and ML, we pave the way for advanced applications, including more effective biomedical entity extraction and improved medical summarization techniques. Ultimately, this research underscores the significance of integrating human intelligence with AI capabilities, setting the stage for more robust and trustworthy healthcare solutions.

## 3. Background

The de-identification model, called the NER classification model, can be considered under four headings:[31] (i) Rule-based models, (ii) ML models, (iii) hybrid models, and (iv) DL models.

Techniques, such as rule-based models and dictionaries can be easily implemented without labels but are vulnerable to input errors.[31-34] ML methods, such as support vector machines and CRFs can recognize complex patterns but require large amounts of labeled data and feature engineering, and are poor at generalization.[35-37] Hybrid systems combine rule-based and ML models, providing high accuracy but requiring intensive feature engineering.[38,39]

Considering the disadvantages of the past three approaches in creating the de-identification systems, the latest state-of-the-art systems employ DL techniques to achieve better results than the best hybrid systems without requiring a time-consuming feature engineering process. DL is an ML subset using multilayered artificial neural networks and is very successful in most natural language processing (NLP) tasks. Recent advances in DL and NLP (especially in the field of NER) enable the systems to outperform the winning hybrid system proposed by Yang and Garibaldi[39] on the 2014 i2b2 de-identification challenge dataset.[31,35]

De-identifying unstructured data is a widely recognized challenge[40] in NLP, involving two key tasks: Identifying PHI and replacing it through masking or obfuscation. Research has primarily focused on PHI identification. Early de-identification approaches,[41,42] especially in healthcare, were rule-based, using regular expressions, syntactic rules, and specialized dictionaries to detect PHI, such as phone numbers and emails. However, they struggled with identifying more complex entities, such as names and professions and required significant adjustments to function in different datasets, limiting their flexibility. The 2014 i2b2 project[34] introduced automatic de-identification, fueling the advancement of ML and DL models for more accurate PHI detection. Early ML methods, such as CRF,[43] used hand-crafted features and lexical rules,[44] signaling a shift to more adaptive and scalable approaches.

Work in the de-identification context has achieved human-level accuracy in de-identifying clinical notes from research datasets. Still, challenges remain in scaling this success to large, real-world environments. The hybrid context-based model outperformed traditional NER models by 10% in the i2b2-2014 benchmark. It also has significantly fewer errors (93% accuracy) compared to ChatGPT (60% accuracy).[45]

Large language-based methods have been used in the development of de-identification models. However, these are still in the early stages, and further development is still needed to protect the privacy and security of health data.[46] The continued need to use APIs in LLM models and the challenge of storing patient data reveal that expert models are still needed.

## 4. Methodology

This section details the purpose of the research, the datasets employed, the methods for training and testing, the data preparation process, and the modeling and evaluation phases. The protection of personal data, compliance with legal regulations, and mitigation of risks associated with processing sensitive patient information are central to this study.

Our LLM-in-the-loop methodology leveraged LLMs at key stages, such as synthetic data generation, labeling, and evaluation, focusing on developing high-performance, expert small models. To this end, we used a combination of proprietary closed-source data, open-source datasets, and synthetic data, all annotated by our labeling team in accordance with i2b2 labeling logic. Incorporating synthetic data and LLM-assisted labeling further enhanced the scope and quality of our training datasets.

For English-language de-identification NER models, we utilized the entire dataset for training. The i2b2 test dataset served as the exclusive test set for evaluation purposes, allowing us to benchmark performance with high precision. For non-English languages, we applied an 80–20 split for training and testing. In addition, our medical translation models[30] were used to translate the English datasets into non-English languages, generating high-quality parallel datasets across multiple languages.

In the data pre-processing phase, we employed language-specific tools to ensure accurate de-identification across different languages. The "Stanza" library was utilized for Romanian-language tasks, while the Natural Language Toolkit library was used for other languages. Word tokenization for all datasets was performed using the "word-punct tokenizer" from the Natural Language Toolkit library.

For evaluation, we adopted the strict evaluation method, where both the chunk and the label had to match to be considered a correct prediction. This rigorous approach ensured the accuracy and reliability of our models, particularly in handling PHI.

By integrating proprietary, open-source, and LLM-synthesized datasets and utilizing real and translated data, this methodology demonstrates the capability of

expert small models to provide accurate, domain-specific de-identification solutions. Our approach minimizes reliance on large LLMs while ensuring privacy and top-tier performance in medical data anonymization.

The results in Tables 2 and 3 were achieved using a structured and detailed prompt to extract PHI from clinical notes. The prompt provided a comprehensive list of entity definitions, such as "AGE," "CITY," "DEVICE," and "ORGANIZATION," along with examples for clarity. It instructed GPT-4o to identify and mark entities using a consistent tagging format (e.g., BEGINER LABEL CHUNK ENDNER) while preserving the original text. Specific guidelines were included for nuanced cases, such as excluding titles (e.g., "Dr.") from names and marking only actual dates for the "DATE" label. This rigorous approach ensured precision in high-performing categories and highlighted areas for improvement in more challenging entities. The prompt used in the study is presented in Appendix A.

### 4.1. Datasets

"i2b2-2014" is a research project (https://portal.dbmi.hms.harvard.edu) on de-identification and heart disease in clinical texts, and its labeling logic was used in our study. For English-language de-identification NER models, we used a training dataset composed of approximately 78% synthetic, AI-generated data and 22% proprietary, closed-source data. The i2b2 training dataset was not used at any stage. The synthetic data were created by generating artificial EHR records and annotating PHI entities using our automated

**Table 2. Categories and comparison of the de-identification model in English i2b2-protected health information**

| Protected health information/model owners | Our scores | Khin *et al.*[31] | Yang and Garibaldi[39] | Kocaman *et al.*[45] | GPT-4o |
|---|---|---|---|---|---|
| AGE | 0.981 | 0.973 | 0.948 | 0.964 | 0.781 |
| CITY | 0.944 | 0.909 | 0.776 | 0.949 | 0.917 |
| COUNTRY | 0.881 | 0.805 | 0.303 | 0.920 | 0.802 |
| DATE | 0.978 | 0.987 | 0.976 | 0.996 | 0.494 |
| DEVICE | 0.762 | - | - | 0.286 | 0.217 |
| DOCTOR | 0.966 | 0.962 | 0.945 | 0.980 | 0.743 |
| HOSPITAL | 0.920 | 0.928 | 0.864 | 0.972 | 0.575 |
| IDNUM | 0.867 | 0.756 | 0.838 | 0.909 | 0.288 |
| LOCATION-OTHER | 1 | - | - | 0.722 | - |
| MEDICALRECORD | 0.942 | 0.979 | 0.971 | 0.980 | 0.716 |
| ORGANIZATION | 0.876 | 0.719 | 0.427 | 0.874 | 0.400 |
| PATIENT | 0.967 | 0.961 | 0.933 | 0.967 | 0.535 |
| PHONE | 0.868 | 0.970 | 0.952 | 0.978 | 0.456 |
| PROFESSION | 0.900 | 0.899 | 0.688 | 0.925 | 0.583 |
| STATE | 0.961 | 0.932 | 0.863 | 0.969 | 0.932 |
| STREET | 0.985 | 0.989 | 0.978 | 0.996 | 0.900 |
| USERNAME | 0.962 | 0.957 | 0.978 | 0.954 | 0.635 |
| ZIP | 0.989 | 0.982 | 0.986 | 0.982 | 0.975 |
| Macro score average | 0.931 | 0.919 | 0.840 | 0.863 | 0.548 |

**Table 3. i2b2 test set scores using GPT-4o**

| Entity | Precision | Recall | F1-score |
|---|---|---|---|
| B-AGE | 0.688 | 0.937 | 0.791 |
| B-CITY | 0.948 | 0.904 | 0.925 |
| B-COUNTRY | 0.908 | 0.718 | 0.832 |
| B-DATE | 0.808 | 0.834 | 0.821 |
| B-DEVICE | 0.132 | 0.625 | 0.217 |
| B-DOCTOR | 0.956 | 0.810 | 0.877 |
| B-HOSPITAL | 0.916 | 0.675 | 0.775 |
| B-IDNUM | 0.340 | 0.672 | 0.531 |
| B-MEDICALRECORD | 0.960 | 0.794 | 0.869 |
| B-ORGANIZATION | 0.303 | 0.695 | 0.422 |
| B-PATIENT | 0.852 | 0.779 | 0.814 |
| B-PHONE | 0.757 | 0.726 | 0.741 |
| B-PROFESSION | 0.695 | 0.637 | 0.665 |
| B-STATE | 0.902 | 0.974 | 0.937 |
| B-STREET | 0.933 | 0.927 | 0.930 |
| B-USERNAME | 0.563 | 0.728 | 0.635 |
| B-ZIP | 1.000 | 0.993 | 0.997 |
| I-AGE | 0.175 | 0.453 | 0.253 |
| I-CITY | 0.872 | 0.852 | 0.862 |
| I-COUNTRY | 0.800 | 0.615 | 0.696 |
| I-DATE | 0.755 | 0.755 | 0.755 |
| I-DEVICE | 0.133 | 1.000 | 0.235 |
| I-DOCTOR | 0.490 | 0.767 | 0.605 |
| I-HOSPITAL | 0.891 | 0.715 | 0.793 |
| I-IDNUM | 0.392 | 0.550 | 0.458 |
| I-LOCATION | 0.114 | 0.121 | 0.118 |
| I-MEDICALRECORD | 0.763 | 0.457 | 0.571 |
| I-ORGANIZATION | 0.246 | 0.750 | 0.370 |
| I-PATIENT | 0.535 | 0.652 | 0.587 |
| I-PHONE | 0.749 | 0.755 | 0.752 |
| I-PROFESSION | 0.628 | 0.693 | 0.659 |
| I-STATE | 0.917 | 0.688 | 0.786 |
| I-STREET | 0.839 | 0.964 | 0.897 |
| I-ZIP | 0.714 | 0.625 | 0.667 |
| O | 0.986 | 0.984 | 0.985 |
| Macro average | 0.5819 | 0.6247 | 0.5775 |
| Weighted average | 0.970 | 0.967 | 0.968 |

Note: Data using the "beginning, inside, outside" (BIO) format.

labeling pipeline. The i2b2 test dataset served as the exclusive evaluation benchmark set, per standard practices. For non-English models, the entire training data were derived by translating[30] the English dataset into the target language, followed by an 80:20 split for training and testing. No open-source test sets were used for non-English languages; the i2b2 test set was used exclusively for English evaluation.

In addition, we utilized several NLP techniques and open-source third-party tools (LangTest by John Snow Labs: https://langtest.org/) to enhance and augment the training datasets. Although the i2b2 2014 dataset was not utilized for training purposes, we provide relevant information and statistics to offer a more comprehensive understanding of its role in our evaluation process. i2b2/UTHealth is a dataset focused on identifying medical risk factors for coronary artery disease in the medical records of diabetic patients, where risk factors include hypertension, hyperlipidemia, obesity, smoking status, and family history, as well as diabetes, coronary artery disease, and indicators suggestive of the presence of these diseases.[47] The i2b2 dataset consists of 1,304 progress notes of 296 diabetic patients. All PHIs in the i2b2-2014 dataset were already de-identified by the dataset's creators before our study, using automated replacement of real entities with synthetic, randomly generated identifiers (e.g., fictitious names or dates). We did not apply any further de-identification to this dataset. Our study used the i2b2-2014 test set exclusively for evaluation and benchmarking. The term "randomly" refers to the replacement strategy used by the original dataset providers to ensure that PHI tokens were substituted with realistic but non-identifying values. The PHIs in this dataset were first categorized into HIPAA categories and then into i2b2-PHI categories, as shown in Table 4. Overall, the i2b2 dataset contains 56,348 sentences with 984,723 individual tokens, of which 41,355 are individual PHI tokens representing 28,867 particular PHI instances.[31]

**Table 4. Protected health information categories of the HIPAA and our study's entities**

| HIPAA | i2b2 dataset | Our dataset |
|---|---|---|
| Name | Patient, doctor, username | Patient, doctor profession |
| Profession | Profession | Profession |
| Location | Street, city, state, country, zip, hospital, organization | Street, city, country, zip, hospital, location, organization |
| Age | Age | Age |
| Date | Date | Date |
| Contact | Phone, fax, email, URL, IP address | Phone, fax, email |
| ID | Medical record, ID no, SSN, license no | Medical record, ID, ID no, SSN, sex, family |

Abbreviations: HIPAA: Health Insurance Portability and Accountability Act; ID: Identification; IP: Internet protocol; SSN: Social security number; URL: Uniform resource locator.

In the literature review, it is seen that there are relative limitations in terms of data sets in de-identification model studies other than English. For this reason, it can be stated that only a few de-identification models have been developed for different languages. In this respect, the de-identification models in different languages developed in this study will contribute to the literature and data scientists working on these models and the health institutions that will use them.

### 4.2. Experimental setup and metrics

#### *4.2.1. Clinical English de-identification model*

The corpus of clinical admission discharge and private clinical reports from private hospitals and healthcare organizations was used to develop the English de-identification model. Labeling was done according to the i2b2-2014 data principles as described previously. The labels employed in the model, which uses a fine-tuned version of the "microsoft/deberta-v3-small" model as an embedding, are shown in Table 5.

In the study, 10 labels were used for the rule-based method, and 18 labels were used for the DL methods. The training dataset was augmented for these labels since "ORGANIZATION," "PROFESSION," and "LOCATION-OTHER" entities gave low results due to the first training process with the DL method. The augmentation stages of the model were performed as follows. First, a fake chunk data frame was created for each label in various formats. Sentences with the labels "ORGANIZATION," "PROFESSION," and "LOCATION-OTHER" in the training dataset and CoNLL file were extracted. Each labeled chunk was removed and replaced with label abbreviations. The sentences were translated from English to the working language. For the translation, our medical translation models used the work by Keles *et al.*[30] The label abbreviations in the new sentences were replaced with new chunks of those labels from the fake data frame.

This new dataset was converted to "beginning, inside, outside" (BIO) format and added to the training dataset. The BIO tagging scheme is a widely adopted convention for NER tasks, where each token in a sentence is labeled as either the beginning (B) of an entity, inside (I) an entity, or outside (O) any entity. This format allows the model to accurately learn both the boundaries and the types of entities in text, and is particularly effective for training DL models for sequence labeling.

The model's performance, implemented with the DL method used in this study, was tested with the i2b2-2014

**Table 5. English de-identification model labels**

| Method | Rule-based | Deep learning |
|---|---|---|
| Labels | ACCOUNT, DLN, EMAIL, FAX, IP, LICENSE, PLATE, SSN, URL, VIN | AGE, CITY, COUNTRY, DATE, DEVICE, DOCTOR, HOSPITAL, IDNUM, LOCATION-OTHER, MEDICAL RECORD, ORGANIZATION, PATIENT, PHONE, PROFESSION, STATE, STREET, USERNAME, ZIP |

Abbreviations: ID: Identification; IP: Internet protocol; URL: Uniform resource locator; VIN: Vehicle identification number.

test set. It was observed that the retrained dataset with augmented labels showed better classification results when evaluated using the i2b2 2014 test set.[33] In the de-identification study conducted in English and with the DL method, a learning rate of $2\times10^5$, a max sentence length of 512, a batch size of two, and ten epochs of training were used. For the rule-based method, regexes suitable for each format were created for the selected labels.

#### *4.2.2. Non-English de-identification models*

To understand which labels could be used in de-identification models and which labels would be appropriate for which aggregates, and to determine the principles, the labeling team organized meetings with relevant hospital staff to develop the models in German, French, Italian, Romanian, Spanish, and Turkish. Data collected from clinical admission reports, discharge reports, and special clinic reports obtained from hospitals and health institutions were labeled according to these principles.

The training was conducted with the obtained data set. In the study, the 0.20 parts of the dataset determined during the division process were used as the test dataset. The dataset was pre-processed and converted into BIO format. For German: bert-base-german-cased, for Italian: bert-base-italian-cased, for French: camembert-bio-base, for Romanian: bert-base-ro-cased, for Turkish: bert-base-turkish-cased, and for Spanish: roberta-base-biomedical-clinical-es were used as embeddings.

The augmentation stages of the other language models were performed as follows. In the dataset used for the English de-identification model, a fake chunk data frame was created for each label in various formats. Each labeled chunk was removed and replaced with label abbreviations. The sentences were translated from English to the working language. For the translation, our medical translation models used the work by Keles *et al.*[30] The label abbreviations in the new sentences were replaced with new chunks of those labels from the fake data frame. This new data set was converted to BIO format and added to the train data set.

First, for each entity label (e.g., ORGANIZATION, PROFESSION, LOCATION-OTHER), we created a data frame of "fake chunks": Short text snippets (typically phrases or entity-level fragments) that could be substituted in place of real PHI. For example, for the label "ORGANIZATION," the fake chunk data frame included items, such as "Springfield General Hospital" or "Central Medical Associates." During augmentation, original EHR sentences with labeled PHI were modified by replacing real entities with randomly sampled fake chunks. For instance, the original EHR sentence reads, "The patient was transferred to Mercy Hospital under the care of Dr. Smith." After augmentation: "The patient was transferred to Springfield General Hospital under the care of Dr. Williams." This process was repeated for each target entity in the dataset. The modified sentences were then converted to the BIO format, where each token is labeled as "B-ENTITY," "I-ENTITY," or "O" (non-entity), as required for training sequence labeling models.

This approach aims to increase training diversity, prevent model memorization of real PHI, and ensure compliance with privacy standards. Detailed examples and code will be available in the project repository upon publication.

The de-identification research was performed with the DL method in seven languages other than English, a learning rate of $2 \times 10^5$, a max sentence length of 512, a batch size of 16 (batch size = 2 in Romanian), and ten epochs were trained.

## 5. Results

The results obtained from the de-identification NER models are shown in Table 2. In addition, the results obtained using GPT-4o and the comparison results of other studies utilizing the same dataset with the results obtained in this study are also included in the same table.

As seen in Table 2, the model realized in this study includes PHIs not used in other studies, and satisfactory results were obtained. When the performance results of the studies are compared with the results of this study, it is determined that new state-of-the-art values were obtained with this study. Although the train was performed with 18 PHI labels (DEVICE and LOCATION-OTHER labels were not used in other studies), and high scores of some labels were not obtained, the F1 macro score (0.931) obtained in this study was higher than the other models, and a new state-of-the-art value was achieved.

GPT-4o performs well in classes, such as "CITY," "COUNTRY," "ZIP," and "STATE," achieving high precision, recall, and F1-scores. However, it struggles significantly with "IDNUM," "LOCATION-OTHER," "ORGANIZATION," "EMAIL," "FAX," and "DEVICE," where the scores are notably low. The macro average (0.5757) indicates that the model's performance varies significantly across classes, with weaker performance in certain categories. In contrast, the micro average (0.5907) is slightly higher, reflecting the model's stronger performance in more frequent classes, but overall, the scores are low.

The results obtained for 13 labels in German, Italian, and French are shown in Table 6, while the results obtained for Turkish (13 labels), Spanish (14 labels), and Romanian (14 labels) are shown in Table 7.

The table presents F1-scores for de-identification tasks across German, Italian, and French datasets. Overall, the German model achieved the highest macro-average (0.960), followed by Italian (0.955) and French (0.937). "DATE" and "PHONE" categories exhibited consistently strong performance across all languages, achieving nearly perfect scores (0.995). In contrast, the "ORGANIZATION" category showed notable variability, with the French model scoring significantly lower (0.699). These results highlight the robustness of the models in categories, such as "AGE," "IDNUM," and "ZIP" while identifying areas for improvement in language-specific challenges, particularly for underperforming categories, such as "ORGANIZATION" in French (Table 6). However, since it was impossible to find any benchmark tests for these languages, comparing the scores obtained in this study was impossible.

Table 7 highlights strong performances for Turkish (0.963) and Spanish (0.957) models, followed by Romanian (0.930) and Arabic (0.922). Categories, such as "DATE," "PHONE," and "MEDICAL RECORD" achieved near-perfect scores across languages, demonstrating model robustness. Lower scores were observed for "CITY" and "ORGANIZATION" in Romanian and Arabic, indicating room for improvement. Missing or language-specific labels (e.g., EMAIL, SSN) show variability in evaluation, reflecting dataset differences. Turkish and Spanish excel in most categories, with consistent performance across diverse labels.

Table 3 presents the evaluation results for the B- and I- tags separately. The model achieved high overall accuracy (0.9672). Classes, such as "B-STATE," "I-CITY," and "I-COUNTRY" performed very well, while "B-EMAIL," "B-FAX," and "I-LOCATION" had lower precision and recall values, indicating challenges in identifying these entities. The macro average (0.5775) was lower than the weighted average (0.968), suggesting that less frequent or more difficult classes were pulling down the macro scores,

**Table 6. German, Italian, and French de-identification model outputs**

| Language/labels | German | Italian | French |
|---|---|---|---|
| AGE | 0.985 | 0.983 | 0.981 |
| CITY | 0.963 | 0.922 | 0.939 |
| COUNTRY | 0.954 | 0.906 | 0.926 |
| DATE | 0.997 | 0.998 | 0.998 |
| DOCTOR | 0.944 | 0.955 | 0.952 |
| HOSPITAL | 0.981 | 0.975 | 0.915 |
| IDNUM | 0.987 | 0.998 | 0.997 |
| ORGANIZATION | 0.865 | 0.916 | 0.699 |
| PATIENT | 0.903 | 0.920 | 0.918 |
| PHONE | 0.995 | 0.995 | 0.995 |
| PROFESSION | 0.980 | 0.917 | 0.941 |
| STREET | 0.945 | 0.952 | 0.949 |
| ZIP | 0.975 | 0.982 | 0.975 |
| Macro score average | 0.960 | 0.955 | 0.937 |

**Table 7. Turkish, Spanish, Romanian, and Arabic de-identification model outputs**

| Language/labels | Turkish | Spanish | Romanian | Arabic |
|---|---|---|---|---|
| AGE | 0.988 | 0.980 | 0.984 | 0.980 |
| CITY | 0.979 | 0.958 | 0.889 | 0.867 |
| COUNTRY | 0.917 | 0.969 | 0.899 | 0.881 |
| DATE | 0.997 | 0.997 | 0.973 | 0.987 |
| DOCTOR | 0.953 | 0.969 | 0.966 | 0.908 |
| EMAIL | - | 0.994 | 0.857 | - |
| HOSPITAL | 0.942 | 0.976 | 0.935 | 0.988 |
| ID | - | 0.995 | - | - |
| IDNUM | 0.979 | - | 0.997 | 0.962 |
| LOCATION | 1 | - | 0.846 | - |
| MEDICAL RECORD | 1 | 0.991 | 0.999 | - |
| ORGANIZATION | 0.975 | 0.734 | 0.768 | 0.978 |
| PATIENT | 0.946 | 0.967 | 0.944 | 0.856 |
| PHONE | 0.982 | 0.981 | 1 | 0.984 |
| PROFESSION | 0.924 | 0.912 | 0.888 | 0.877 |
| SEX | - | 0.971 | - | - |
| SSN | - | 0.937 | - | - |
| STREET | 0.913 | 0.959 | 0.953 | 0.768 |
| ZIP | 0.913 | 0.980 | 0.992 | 0.950 |
| FAX | - | - | 0.923 | - |
| FAMILY | 1 | - | - | - |
| Macro score average | 0.963 | 0.957 | 0.930 | 0.922 |

whereas the model was quite successful in predicting the more common entities.

The low scores can be attributed to several factors. The model struggled to recognize patient and doctor names embedded in the middle of the text, despite successfully identifying those at the beginning and end. Some hospital names were partially labeled, affecting overall precision and recall. Occasionally, the model included extra tokens within labels, leading to incorrect annotations. Furthermore, although the prompt explicitly specified which labels to use, the model occasionally introduced unintended labels (e.g., time). Confusion between labels or failure to identify them also contributed to the lower performance.

## 6. Discussion

This work showed that expert small NER models, built with an LLM-in-the-loop development process, can deliver strong multilingual PHI de-identification while keeping inference on-premises. Across eight languages, macro-F1 scores ranged from 0.922 (Arabic) to 0.963 (Turkish), with consistently high scores for common identifiers, such as "DATE" and "PHONE." Deploying lightweight, on-premises models avoids routine transmission of clinical text to external services, which better aligns with privacy regimes, such as HIPAA and General Data Protection Regulation when paired with organizational controls (access management, audit logging, and defined data-retention policies).

Performance was not uniform across entity types or languages. "ORGANIZATION" was the most fragile category in French (0.699), and "CITY" was comparatively weaker in Romanian (0.889) and Arabic (0.867). We attribute these gaps to linguistic factors (orthography, compounding, and rich morphology), domain naming conventions (hospital and clinic aliases), tokenizer mismatch, and limited language-specific coverage in gazetteers and training corpora. Targeted remedies include (i) language-tailored augmentation that preserves morphology and diacritics, (ii) curated medical-organization gazetteers and alias tables, (iii) character-aware and subword-robust encoders; and (iv) post-processing with constrained decoding and span-consistency checks to reduce boundary errors.

The LLM-in-the-loop paradigm was most valuable during data synthesis, labeling quality assurance, and error triage, while excluding LLMs from deployment helped mitigate API-related privacy risk. Reliance on synthetic and translated corpora for non-English training limits real-world generalization. Future work must prioritize evaluation on native clinical notes where feasible, out-of-distribution stress tests (novel facilities, regional toponyms), and ablations that quantify the incremental value of each loop component versus purely supervised baselines. To strengthen the privacy posture, future work must also explore federated fine-tuning across institutions and differentially private optimization to bound memorization risks, and report operational characteristics (latency, memory footprint, and minimal hardware) to support adoption.

Finally, we identified several immediate paths to broader utility, such as extending models to related biomedical entities and relations, uncertainty-aware inference to flag low-confidence spans for human review, and releasing prompts, evaluation scripts, and error taxonomies to enable reproducibility. Addressing the noted weaknesses—especially expanding native, annotated non-English resources—will make these expert small models more inclusive, robust, and clinically practical while preserving patient privacy.

## 7. Conclusion

This study underscores the importance of de-identification as a key method for safeguarding patient/personal health information and ensuring its ethical use in scientific research. By removing identifiable details through techniques, such as anonymization, generalization, and differential privacy, de-identification allows data to be used for diverse scientific applications, including epidemiological studies, disease modeling, and AI development, while maintaining patient privacy.

Recent advancements have demonstrated the potential of LLMs in de-identification tasks. However, challenges remain, particularly around issues of patient data security, API dependencies, and the need for domain-specific expertise in handling EHRs. Our "LLMs-in-the-loop" approach addresses these concerns by integrating small, specialized models tailored to the medical field. This method enhances both privacy and reliability, enabling the secure use of data without relying on external APIs or compromising sensitive patient information.

The multilingual nature of this research, spanning several languages, shows the adaptability and robustness of our models across diverse healthcare environments. While there are inherent risks associated with data anonymization, this study demonstrates that when properly applied, de-identification models can strike a delicate balance between protecting individual privacy and maximizing the utility of health data.

Furthermore, as the field progresses, it is crucial to establish globally recognized standards, raise awareness

of best practices, and ensure that ethical principles guide the deployment of de-identification technologies. Transparency, accountability, and a rigorous risk-benefit analysis must remain at the forefront of these efforts.

Ultimately, the findings of this study highlight the potential of expert small models developed through the LLMs-in-the-loop methodology to meet the evolving demands of healthcare research. The models presented here offer a reliable and scalable solution for future de-identification applications, advancing the capabilities of AI in healthcare while safeguarding patient privacy.

Future research should focus on refining and expanding de-identification models to cover a wider range of languages and healthcare contexts. One of the primary challenges is the scarcity of high-quality, annotated datasets in languages other than English, which limits the development of robust models for non-English speaking regions. Addressing this gap will require collaborative efforts to create and share multilingual datasets, ensuring more comprehensive language coverage. In addition, future studies could explore more advanced augmentation techniques and develop models capable of handling increasingly complex medical data types, such as clinical narratives and imaging reports. Continuous innovation in privacy-preserving methods, such as federated learning, may also prove valuable in safeguarding sensitive patient information while advancing the performance and applicability of de-identification technologies across diverse healthcare systems.

## Acknowledgments

None.

## Funding

None.

## Conflict of interest

The authors declare that they have no competing interests.

## Author contributions

*Conceptualization:* Murat Gunay
*Methodology*: Murat Gunay
*Software:* Bunyamin Keles, Raife Hizlan
*Validation:* Murat Gunay, Bunyamin Keles
*Writing – original draft:* Murat Gunay, Raife Hizlan
*Writing – review & editing:* Bunyamin Keles, Raife Hizlan

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

Data will be made available upon request to the corresponding author after the evaluation process.

## References

1. Ahmed T, Al Aziz MM, Mohammed N. De-identification of electronic health record using neural network. *Sci Rep*. 2020;10(1):18600.

   doi: 10.1038/s41598-020-75544-1

2. Wood A, Denholm R, Hollings S, *et al*. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: Data resource. *BMJ*. 2021;373:n826.

   doi: 10.1136/bmj.n826

3. Gungoren M, Orhan F, Kurutkan N. *Mikro Rekabetcilikte Yeni Yaklasımlar: Hastanelerde Olusan Etik Iklimin Kalite ve Akreditasyon Acısından Degerlendirilmesi [New Approaches in Micro-Competitiveness: Evaluating the Ethical Climate in Hospitals in Terms of Quality and Accreditation*]. Vol. 18. Suleyman Demirel Universitesi Iktisadi ve Idari Bilimler Fakultesi Dergisi; 2013. p. 221-241. Available from: https://dergipark.org.tr/tr/pub/sduiibfd/issue/20819/222797 [Last accessed on 2025 Sep 17].

4. Varol S, Orhan F, Tuncer S, Akyuz S. Saglık kurumlarında bilgi guvenligi baglamında biyometrik sistemler [Biometric systems in the context of information security in healthcare institutions]. *Saglık Akadem Derg*. 2016;3(4):155-162.

   doi: 10.5455/sad.13-1483706096

5. Yilmaz D, Ozkoc EE, Ogutcu Ulas G. Elektronik saglik kayitlarinda farkındalık [Awareness of electronic health records]. *Hacettepe Sağlık İdaresi Derg*. 2021;24(4):777-792.

6. HealthITSecurity. *De-Identification of PHI According to the HIPAA Privacy Rule*. Available from: https://healthitsecurity.com/features/de/identification/of/phi/according/to/the/hipaa/privacy/rule [Last accessed on 2023 Apr 13].

7. Act A. *Health Insurance Portability and Accountability Act of 1996*. Vol. 104. Public Law; 1996. p. 191. Available from: https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf [Last accessed on 2025 Sep 17].

8. Fernández-Alemán JL, Señor IC, Lozoya PÁ, Toval A. Security and privacy in electronic health records: A systematic literature review. *J Biomed Inform*. 2013;46(3):541-562.

   doi: 10.1016/j.jbi.2012.12.003

9. Office for Civil Rights HH. Standards for privacy of individually identifiable health information. Final rule. *Fed Regist*. 2002;67(157):53181-53273.

10. Toscano F, O'Donnell E, Unruh MA, *et al*. Electronic health records implementation: Can the European union learn from the United States? *Eur J Public Health*. 2018;28 Suppl 4:pcky213.401.

doi: 10.1093/eurpub/cky213.401

11. *Guidance on De-Identification of Protected Health Information hhs Deid Guidance.pdf*; 2012. Available from: https://www.hhs.gov/sites/default/files/ocr/priv/identification/hhs/deid/guidance.pdf [Last accessed on 2023 Jul 17].

12. *Standards for Privacy of Individually Identifiable Health Information HHS.gov*,; 2013. Available from: https://www.hhs.gov/hipaa/for/professionals/privacy/guidance/standards/privacy/individually/identifiable/health/information/index.html [Last accessed on 2023 Jul 17].

13. Neamatullah I, Douglass MM, Lehman LW, *et al*. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak*. 2008;8:32.

doi: 10.1186/1472-6947-8-32

14. Paul T, Rana MKZ, Tautam PA, *et al*. Investigation of the utility of features in a clinical de-identification model: A demonstration using EHR pathology reports for advanced NSCLC patients. *Front Digit Health*. 2022;4:728922.

doi: 10.3389/fdgth.2022.728922

15. Garfinkel S. *De-Identification of Personal Information, 2015: US Department of Commerce, National Institute of Standards and Technology*. Available from: https://nvlpubs.nist.gov/nistpubs/ir/2015/nist.ir.8053.pdf [Last accessed on 2025 Sep 17].

16. Wu H, Toti G, Morley KI, *et al*. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J Am Med Inform Assoc*. 2018;25(5):530-537.

doi: 10.1093/jamia/ocx160

17. Stubbs A, Uzuner O. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *J Biomed Inform*. 2015;58 Suppl: S78-S91.

doi: 10.1016/j.jbi.2015.05.009

18. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M. A novel COVID-19 data set and an effective deep learning approach for the de-identification of italian medical records. *IEEE Access*. 2021;9:19097-19110.

doi: 10.1109/ACCESS.2021.3054479

19. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc*. 2020;27(3):491-497.

doi: 10.1093/jamia/ocz192

20. Ong JCL, Seng BJ, Law JZ, *et al*. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: Current state and future directions. *Cell Rep Med*. 2024;5(1):101356.

doi: 10.1016/j.xcrm.2023.101356

21. Gunasekeran DV, Tham YC, Ting DS, Tan GS, Wong TY. Digital health during COVID-19: Lessons from operationalising new models of care in ophthalmology. *Lancet Digit Health*. 2021;3(2):e124-e134.

doi: 10.1016/S2589-7500(20)30287-9

22. Ting DS, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. *Nat Med*. 2020;26(4):459-461.

doi: 10.1038/s41591-020-0824-5

23. Verdicchio M, Perin A. When doctors and AI interact: On human responsibility for artificial risks. *Philos Technol*. 2022;35(1):11.

doi: 10.1007/s13347-022-00506-6

24. Dai SC, Xiong A, Ku LW. *LLM-in-the-Loop: Leveraging Large Language Model for Thematic Analysis*. [arXiv Preprint ]; 2023.

doi: 10.48550/arXiv.2310.15100

25. De Paoli S. *Can Large Language Models Emulate an Inductive Thematic Analysis of Semi-Structured Interviews? An Exploration and Provocation on the Limits of the Approach and the Model*. [arXiv Preprint]; 2023.

doi: 10.48550/arXiv.2305.13014

26. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text- annotation tasks. *Proc Natl Acad Sci U S A*. 2023;120(30):e2305016120.

doi: 10.1073/pnas.2305016120

27. Islam T, Goldwasser D. *Discovering Latent Themes in Social Media Messaging: A Machine-in-the-Loop Approach Integrating Llms*. [arXiv Preprint]; 2024.

doi: 10.48550/arXiv.2403.10707

28. Pham DK, Vo BQ. *Towards Reliable Medical Question Answering: Tech- niques and Challenges in Mitigating Hallucinations in Language Models*. [arXiv Preprint]; 2024.

doi: 10.48550/arXiv.2408.13808

29. Umphrey R, Roberts J, Roberts L. *Investigating Expert-in-the-Loop LLM Discourse Patterns for Ancient Intertextual Analysis*. [arXiv Preprint ]; 2024.

doi: 10.48550/arXiv.2409.01882

30. Keles B, Gunay M, Caglar SI. *LLMs-in-the-loop Part-1: Expert Small AI Models for Bio-Medical Text Translation*. [arXiv Preprint]; 2024.

doi: 10.48550/arXiv.2407.12126

31. Khin K, Burckhardt P, Padman R. *A Deep Learning Architecture for De- identification of Patient Notes: Implementation and Evaluation*. [arXiv Pre-print ]; 2018.

doi: 10.48550/arXiv.1810.01570

32. Morrison FP, Sengupta S, Hripcsak G. Using a pipeline to improve de-identification performance. *AMIA Annu Symp Proc*. 2009;2009:447-451.

33. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UThealth shared task track 1. *J Biomed Inform*. 2015;58 Suppl: S11-S19.

    doi: 10.1016/j-bi.2015.06.007

34. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*. 2007;14(5):550-563.

    doi: 10.1197/jamia.M2444

35. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc*. 2017;24(3):596-606.

    doi: 10.48550/arXiv.1606.03475

36. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. *BMC Med Res Methodol*. 2012;12:109.

    doi: 10.1186/1471-2288-12-109

37. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Med Res Methodol*. 2010;10:70.

    doi: 10.1186/1471-2288-10-70

38. Liu Z, Chen Y, Tang B, *et al*. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *J Biomed Inform*. 2015;58 Suppl: S47-S52.

    doi: 10.1016/j.jbi.2015.06.009

39. Yang H, Garibaldi JM. Automatic detection of protected health information from clinic narratives. *J Biomed Inform*. 2015;58 Suppl: S30-S38.

    doi: 10.1016/j.jbi.2015.06.015

40. Nadkarni PM, Ohno-Machado L, Chapman WW. *Natural language processing: An introduction*. *J Am Med Inform Assoc*. 2011;18(5):544-551.

    doi: 10.1136/amiajnl-2011-000464

41. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. *Proc AMIA Annu Fall Symp*. 1996:333-337.

42. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. *Am J Clin Pathol*. 2004;121(2):176-186.

    doi: 10.1309/E6K3-3GBP-E5C2-7FYU

43. He B, Guan Y, Cheng J, Cen K, Hua W. CRFs based de-identification of medical records. *J Biomed Inform*. 2015;58 Suppl: S39-S46.

    doi: 10.1016/j.jbi.2015.08.012

44. Lafferty J, McCallum A, Pereira F. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. Williamstown, MA: ACM; 2001.

    doi: 10.1145/3696410.3714901

45. Kocaman V, Talby D, Hak HU. Beyond accuracy: Automated de-identification of large real-world clinical text datasets. *Value in Health*. 2023;26(12):S532.

    doi: 10.48550/arXiv.2312.08495

46. Liu Z, Huang Y, Cao C, *et al*. *Deid-Gpt: Zero-Shot Medical Text de-Identification by Gpt-4*. [arXiv Preprint]; 2023.

    doi: 10.48550/arXiv.2303.11032

47. Stubbs A, Kotfila C, Xu H, Uzuner O. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task track 2. *J Biomed Inform*. 2015;58 Suppl: S67-S77.

    doi: 10.1016/j.jbi.2015.07.001

## Appendix

### A. The prompt used to obtain benchmarks with GPT-4o

prompt = f""" You are tasked with extracting Protected Health Information (PHI) from clinical notes. Your job is to identify and mark specific entities within the text. Here are the entities you need to look for:

<entities>

AGE (Identifies the age number or age-related information. Example: In "88 years old," 88 would be marked as AGE. In "in his 50's," 50's would be marked as AGE.)

CITY (Identifies the name of a city.)

COUNTRY Identifies the name of a country.)

DATE (Identifies specific dates or years. Example: In "He was admitted on 03/29/2089," 03/29/2089 would be marked as DATE. In "His surgery was in the 1980s," 1980s would be marked as DATE. In "His record was marked on 2089-08-24," 2089-08-24 would be marked at DATE.)

DEVICE (Identifies serial numbers, item code, or product code of a medical device mentioned. Example: In " The AA 737 pacemaker was implanted," AA 737 would be marked as DEVICE.)

DOCTOR (Identifies the name of a doctor or healthcare professional. Only the name should be marked, not the title, such as " Dr.," " M.D.")

HOSPITAL (Identifies the name of a hospital or nursing home.)

IDNUM (Identifies identification numbers, such as medical record or patient numbers.)

LOCATION (Identifies specific locations related to healthcare, excluding city or country.)

MEDICALRECORD (Identifies medical record numbers or similar identifiers.)

ORGANIZATION (Identifies names of organizations or institutions.)

PATIENT (Identifies the patient's name. Only the name should be marked, not titles like "Mr." or " Mrs.")

PHONE (Identifies phone numbers, including fax numbers.)

PROFESSION (Identifies professions or job title.)

STATE (Identifies the name of a state or region.)

STREET (Identifies street addresses.)

USERNAME (Identifies usernames or account IDs.)

ZIP (Identifies postal or zip codes.)

</entities >

I will provide you with a clinical note. Your task is to process this note and mark all instances of the PHI entities listed above.

Here is the clinical note:

{clinical_note}

Instructions for marking PHI entities:

* Carefully read through the entire clinical note.

* Identify any text that matches one of the PHI entity types listed above.

* For each identified PHI entity, mark the beginning and end of the relevant text chunk using the following format:

BEGINNER_ LABEL CHUNK ENDNER where ENTITY LABEL is one of the entity types from the list, and CHUNK is the actual text containing the PHI.

* While marking, DO NOT EDIT OR CHANGE the original clinical text, only put marks described above.

Here are a few examples of correct markup: Original text:

Mrs. Linda Martinez, a 45-year-old architect, having MR \#:2775283 for an evaluation on 2023-05-10. Her insulin pump model ZX900 was assessed by Dr. Michael Brown, M. D. The patient's condition has improved since the 1990s, but she mentioned feeling unwell for the past 6 months. MF381/1183 was referenced during her visit, which lasted approximately 5 hours and concluded at 10:05:03. She was discharged on 20/10/2023.

Marked text:

Mrs. BEGINER_PATIENT Linda Martinez ENDNER, a BEGINER_AGE 45 ENDNER year- old BEGINER_ PROFESSION architect ENDNER, having MR\#: BEGINER_MEDICALRECORD 2775283 ENDNER for an evaluation on BEGINER_DATE 2023-05-10 ENDNER. Her insulin pump model BEGINER_DEVICE ZX900 ENDNER was assessed by Dr. BEGINER_DOCTOR Michael Brown ENDNER, M. D. The patient's condition has improved since the BEGINNER_DATE 1990s ENDNER, but she mentioned feeling unwell for the past 6 months. BEGINER_IDNUMMF381/1183 ENDNER was referenced during her visit, which lasted approximately 5 hours and concluded at 10:05:03. She was discharged on BEGINER_DATE 20/10/2023 ENDNER.

Important notes:

* Be sure to process the entire clinical note and mark all instances of PHI entities.
* If a chunk of text could belong to multiple entity types, choose the most specific or appropriate one.
* Do not mark information that is not part of the specified PHI entity types.
* Preserve the original text exactly as it appears, including any spelling errors or formatting.
* Label the data, ensuring that professional titles or suffixes, such as "M. D.," "Ph. D.," or similar, are not removed. These titles must be preserved exactly as they appear in the text, without alteration or omission, and should NEVER be inside the label.
* Apostrophe "s" ('s) should not be included within the label when associated with Names. Only the person's name should be inside the label, and the apostrophes should remain outside the marked text. However, apostrophe s' is allowed within the DATE label when referring to a decade (e.g., 80's).
* Mark only specific calendar dates as DATE. Do not mark relative time expressions like "6 months," "1 year ago," "5 weeks," "5 wks," "yesterday," "today," "days," or similar units of time (months, years, weeks), as they do not represent actual dates.
* Mark only actual dates as DATE. Do not mark time-related expressions, such as "10:05:03," "10 am," or durations like "5 hours" as DATE, since they refer to times or durations rather than specific calendar dates.
* Fax numbers should be treated as PHONE entities and marked the same way as phone numbers.

Please process the provided clinical note and return it with all PHI entities appropriately marked.

"""

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

# Forecasting world health expenditures: A hybrid artificial intelligence framework

**Taegeon Yu[1], Daipayan Bera[1], Abbas Maazallahi[2], Roschlynn Dsouza[1], Francina Pali[1], Wen-Shan Liu[1], Payam Norouzzadeh[3], Eli Snir[4], and Bahareh Rahmani[1]***

[1]Department of Health and Clinical Outcomes Research, School of Medicine, Saint Louis University, Saint Louis, Missouri, United States of America

[2]Department of Computer Science, School of Mathematics, Statistics, and Computer Science, College of Science, University of Tehran, Tehran, Iran

[3]Department of Analytics, School for Professional Studies, Saint Louis University, Saint Louis, Missouri, United States of America

[4]Department of Business Analytics, Olin Business School, Washington University in Saint Louis, Saint Louis, Missouri, United States of America

## Abstract

Global healthcare expenditures continue to rise, posing substantial economic challenges, particularly for low- and middle-income countries (LMICs), where resource constraints intensify the impact. Accurate forecasting, efficient resource allocation, and equitable policy development are essential to address these growing pressures. This study presents a hybrid analytical framework that integrates generative artificial intelligence (AI) with traditional econometric and machine learning models to analyze and predict trends of healthcare expenditure. Utilizing data from the World Bank and World Health Organization, we applied generative adversarial networks, hierarchical clustering, support vector machines, and autoregressive integrated moving average models to uncover spending patterns, simulate policy scenarios, and tackle disparities in health investment. Generative AI played a pivotal role by augmenting sparse and incomplete datasets, particularly from underrepresented LMICs, identifying anomalies, and generating realistic synthetic data to support robust forecasting. This enabled the development of more inclusive, equity-focused health resource planning tools. The results demonstrate improved forecasting accuracy and offer deeper insights into regional and income-based differences in expenditure trends. By combining traditional machine learning with cutting-edge generative models, this study advances a scalable, data-driven approach to support global health decision-making. Ultimately, generative AI is highlighted as a transformative enabler of equitable, informed strategies in the management of global healthcare expenditures.

***Keywords:*** Healthcare expenditure; Generative artificial intelligence; Autoregressive integrated moving average; Health equity; Support vector machines; Low- and middle-income countries

***Corresponding author:**
Bahareh Rahmani
(bahareh.rahmani@health.slu.edu)

**Citation:** Yu T, Bera D, Maazallahi A, *et al*. Forecasting world health expenditures: A hybrid artificial intelligence framework. *Artif Intell Health*. 2026;3(1):152-163. doi: 10.36922/AIH025170033

**Received:** April 21, 2025

**1st revised:** June 22, 2025

**2nd revised:** July 4, 2025

**Accepted:** August 12, 2025

**Published online:** September 22, 2025



## 1. Introduction

Rising global healthcare expenditure presents a critical challenge for policymakers, especially in low- and middle-income countries (LMICs), where financial and

infrastructural limitations exacerbate health inequities. Traditional econometric models often fall short in capturing the complexity of health spending patterns across diverse socioeconomic and regional contexts. Moreover, the scarcity and inconsistency of data in LMICs hinder the development of robust forecasting tools and the formulation of health policies.

Following substantial increases of 7.4% in 2022 and 10.7% in 2023, it is estimated that the growth in global healthcare expenditure will remain high at 9.9% in 2024. While there is debate over whether population aging is the primary cause of rising healthcare costs, many analysts and representatives consider it one of the major contributors.[1] Healthcare expenditure has become increasingly burdensome globally, driven by various factors, including gross domestic product (GDP) growth, aging populations, and healthcare strategies.[2,3] The United States (US) exemplifies the surge in healthcare expenditure across both private and public sectors, with per capita costs and the share of GDP rising sharply due to rapid advancements in medical treatments and technology. Technological advancements, especially in developed nations, such as the US, have significantly expanded treatment options and increased per capita spending, often exceeding expenditures in many countries with universal healthcare structures.[2] Estimates of changes in global health research and development expenditures are crucial for improving and setting boundaries in health research planning.[4,5] However, comparing this information across countries can be a challenging task, as there is a gap in measuring data per area.[4]

Over the past couple of decades, healthcare spending has nearly doubled in many high-income countries, with projections suggesting that global spending could reach $15 trillion within 25 years.[5] In 2021, US healthcare spending was nearly twice that of other high-income countries, with significantly higher costs for inpatient, outpatient, and administrative facilities.[6] In addition, it is estimated that the growth of US per capita spending on health, at an annual average rate of 5.0%, will continue to outpace the GDP growth from 2023 to 2027, stressing the need for systemic reform. At this rate, it is estimated that health spending will reach 17.9% of GDP in 2025, growing slightly faster than the economy. In addition, health spending growth is expected to exceed overall economic growth, reaching 19.7% of GDP within 7 years.[7]

These figures underscore the significance of government regulations and policies in mitigating the upward trend in healthcare costs. However, fluctuations in government revenues, driven by events, such as the COVID-19 pandemic, the Russo-Ukrainian War, inflation, and recessionary pressures, have widened spending disparities between high- and low-income countries. In addition, the weakening global economy has also constrained health funding. The global economic slowdown in 2021 squeezed health budgets, reducing health aid contributions from high-income countries, such as the United Kingdom (UK) and Sweden, thereby diminishing support for low-income nations.[8]

Several interconnected factors contribute to the excessive rise in healthcare expenditures, including expanded insurance coverage, supplier-induced demand, defensive medicine, factor productivity, and technological advances.[9] While these technological developments have improved treatment options and quality of life, they have also contributed to sustained increases in healthcare expenses in the US. Conversely, lifestyle factors, particularly modifiable risk behaviors, such as regular physical activity and healthy dietary choices, can reduce healthcare expenses. Moreover, environmental factors, including carbon dioxide emissions and fossil fuel consumption, negatively impact population health.[10] For example, one study indicates that air pollution in China significantly increases healthcare costs, with the economic burden extending well beyond classic respiratory illnesses.[11]

Innovations in technology have introduced automation and cost-effectiveness into healthcare, transforming research, diagnostics, and treatment delivery.[12] Artificial intelligence (AI) has become a transformative instrument across several fields, particularly in health economics and outcomes research. The application of AI, particularly machine learning, offers novel approaches for enhancing prediction models, economic analysis, and healthcare decision-making trials[13] For example, to evaluate the possibility of patient hospitalization, several machine learning strategies have been developed using prevalent methodologies with insurance claim datasets to improve prediction accuracy.[14,15] Other studies have validated how machine learning is revolutionizing medical commerce.[16,17] It is a crucial task for humanity to enhance the quality of healthcare through machine learning techniques, as it was shockingly found that a high proportion of healthcare expenditure failed to protect countries from COVID-19.[18] Nonetheless, machine learning should be viewed as a complement to, rather than a replacement for, human judgment, allowing human oversight to remain as a core standard.[19]

The integration of AI into healthcare has transformed the landscape of diagnostics, treatment, public health, and health systems management. Over the past decade, advancements in machine learning and deep learning have significantly driven the development of intelligent, data-

driven health technologies. For example, Rajkomar *et al.*[20] provided a foundational overview of how machine learning is revolutionizing diagnostics and clinical decision-making, enabling earlier and more accurate treatment strategies. Complementing this, Topol[21] emphasized the synergistic potential of AI and human intelligence, coining the term "high-performance medicine" to describe the future of personalized, precise care. Moreover, Beam and Kohane[22] further highlighted the importance of integrating big data and machine learning tools into healthcare workflows, noting that the scalability of these technologies can improve both outcomes and efficiency.

The rapid pace of innovation is evident, with AI applications in diagnostics, drug development, and remote patient monitoring advancing swiftly, as noted by Heaven.[23] These advances are particularly crucial for enhancing the delivery of healthcare, as outlined by Reddy *et al.*,[24] who discussed both the opportunities and limitations of AI in real-world systems. Similarly, Yu *et al.*[25] provided a more comprehensive review of the challenges in implementing AI technologies, including regulatory hurdles, ethical concerns, and data privacy considerations.

Deep learning, a subset of AI, has become particularly important in medical applications. Esteva *et al.*[26] proposed a practical guide to deep learning tools in healthcare, demonstrating how models, such as convolutional neural networks are now used to interpret medical images and analyze unstructured clinical data. In low-resource settings, AI also holds great promise. Wahl *et al.*,[27] explored how it can be used to reduce health disparities and enhance access to care in underrepresented regions.

The COVID-19 pandemic has accelerated digital adoption in healthcare, including the deployment of AI for surveillance, diagnostics, and modeling. Keesara *et al.*,[28] described this shift as a digital revolution catalyzed by the pandemic. AI's clinical utility is already evident in specialties, such as cardiology, where Dilsizian and Siegel[29] demonstrated its efficacy in cardiac imaging diagnostics.

However, ethical concerns remain. Obermeyer and Emanuel[30] critically examined racial bias in healthcare algorithms, showing how even data-driven tools can perpetuate inequities. This is echoed by Holmes *et al.*,[31] who advocated for a more equity-focused approach to AI deployment, especially in global health contexts.

The role of big data is central to these transformations. Chen and Chen[32] explored how AI-powered analytics can support public health interventions, while Miotto *et al.*[33] proposed "deep patient," an unsupervised model that can accurately predict health outcomes from electronic health records. The global health crisis sparked by COVID-19 further underlined the importance of rapid data modeling and response strategies, as reviewed by Wang *et al.*[34] In addition, LeCun *et al.*[35] laid the theoretical foundation of deep learning technologies, offering insights into neural networks that now underpin most AI systems in healthcare. Building on this, Wang *et al.*[36] emphasized the organizational advantages of big data analytics in hospital and system-level planning.

Furthermore, the use of AI in specific clinical applications is highlighted by Krittanawong *et al.*,[37] who detailed how deep learning enhances cardiovascular risk prediction and diagnostics. From a practical perspective, Davenport and Kalakota[38] and Hinton[39] underscored the growing integration of AI into clinical routines, making a case for workflow optimization and clinical support tools. Finally, the future of AI in healthcare depends on its alignment with human oversight and decision-making. Shortliffe and Sepúlveda[40] highlighted the importance of clinical decision support systems that enhance, rather than replace, physician judgment, ensuring that ethical and contextual considerations remain central to care.

Together, these works establish a comprehensive framework for understanding the potential, progress, and pitfalls of AI in global healthcare systems. The literature collectively supports the premise that, when implemented thoughtfully and equitably, AI holds transformative power to reshape medicine, improve patient outcomes, and reduce disparities worldwide.

This study proposed a novel, AI-driven forecasting framework that combines generative AI (e.g., generative adversarial networks [GANs]) with conventional machine learning methods, such as support vector machines (SVMs) and autoregressive integrated moving average (ARIMA) models. By leveraging generative AI, we aimed to augment incomplete or imbalanced datasets—particularly those from underrepresented regions—thereby enabling more inclusive and equitable global health expenditure forecasting. The proposed methodology not only improves prediction accuracy but also uncovers latent patterns and clusters of countries based on spending behavior. This hybrid approach enhances the interpretability and relevance of predictions for LMICs, empowering policymakers with reliable, data-driven insights to allocate resources more effectively and equitably.

## 2. Methods

The data utilized in this study were retrieved from the database of the World Bank Group (https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS?most_recent_year_desc=true&locations=1W), which retrieved its data from the World Health Organization's Global

Health Expenditure Database (GHED). The GHED is an integrated database comprising a total of 192 countries. A subset of 168 countries was selected in this study. The data were represented in a wide format, comprising country-level healthcare expenditures as a percentage of GDP, collected over the years from 2000 to 2021. In the selected data, the highest health expenditure as a percentage of GDP occurred in Nauru in 2007 (24.23%), whereas the lowest ever was recorded in Qatar in 2011 (1.60%). The US consistently allocated the highest share of GDP spent on healthcare throughout the study period, with a mean of 15.66%, except in 2001, 2007, and 2008 (when Nauru exceeded the US), and in 2014 and 2015 (when Sierra Leone exceeded the US). In contrast, despite maintaining high standards of medical care, Qatar's expenditure on healthcare is one of the lowest among all other countries in the world, averaging 2.61%. This dataset was subsequently used for machine learning classification.

Data from 24 countries were selected from the above dataset and transposed. The new dataset comprised primarily countries from North America and Europe, along with a few Asian countries, including India, China, and Iran. The resulting data are a time series indicating the health expenditure of 24 countries from 2000 to 2021. This dataset was used to predict the health expenditure of selected countries in 2025.

The healthcare expenditure data used in this study were expressed as a percentage of GDP, which inherently normalizes for inflation at the national level by relating spending to the overall economic output. As such, an explicit inflation adjustment was not required for this analysis. Nonetheless, we acknowledge that inflation can still indirectly influence healthcare costs, and future work should consider integrating inflation-adjusted absolute spending values. The following methods were employed to analyze and predict future healthcare expenditures.

### 2.1. Hierarchical clustering

A hierarchical clustering approach was employed in this study. The healthcare expenditure dataset was organized by nations nested within broader geographic or economic regions, using the city block distance metric and average linkage methods. This structure allowed the understanding of both country-specific variations and regional or global trends over time. In this hierarchical framework, the assumptions were:

(i) Level 1 represents the annual healthcare expenditure for each country, varying by year
(ii) Level 2 captures regional-level or income-based groupings.

The model is represented by Equation I:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij} \quad \text{(I)}$$

Where $Y_{ij}$ represents the healthcare expenditure for country $j$ in year $i$, $X_{ij}$ denotes time-specific covariates, and $\varepsilon_{ij}$ is the error term. The intercept $\beta_{0j}$ and slope $\beta_{1j}$ terms may vary across countries, allowing the capturing of both the overall trend and country-specific deviations.

### 2.2. SVMs

The high dimension of the data remained a consistent challenge in employing different algorithms. To reduce the complexity, the SVM model was used for predictive analysis due to its strong performance in classification and regression tasks. SVM aims to identify an optimal hyperplane in higher dimensions that maximally separates data points from different classes or accurately fits the data in regression tasks. For classification, the SVM model constructs a decision boundary that maximizes the margin between data points from distinct classes, thereby minimizing classification error and enhancing generalization to new data.

### 2.3. ARIMA

In this study, the ARIMA model was employed to analyze and forecast time-dependent trends in healthcare expenditures across countries. The ARIMA model is a widely used statistical method for analyzing time series data, especially when a series exhibits temporal autocorrelation. The model is particularly advantageous for handling non-stationary data by integrating differencing, autoregressive, and moving average components, making it suitable for our dataset, which spans multiple years. Based on the autocorrelation function and partial autocorrelation function diagnostics, initial ARIMA model configurations were identified and fitted to the data. Model performance was assessed using standard evaluation metrics, including the Akaike information criterion (AIC) and Bayesian information criterion (BIC), to select the most parsimonious model with adequate predictive accuracy.

### 2.4. Generative AI

Generative AI, a subset of AI that focuses on generating new data instances resembling existing data, offers innovative solutions for complex problems in health economics. Unlike traditional predictive models, generative AI models, such as GANs and variational autoencoders (VAEs), excel in simulating realistic scenarios and exploring data-driven solutions, making them a valuable tool for understanding and forecasting healthcare expenditure under various economic, demographic, or policy-driven conditions. By leveraging models, such as GANs or conditional VAEs,

researchers can generate plausible future trends based on historical data and specific input variables. For example, these models can simulate how an economic downturn, a new healthcare policy, or a pandemic may affect spending patterns across regions or income groups. Such simulations enable policymakers to explore potential outcomes before implementing changes, enabling proactive planning and risk mitigation.

## 3. Results

### 3.1. Machine learning

#### *3.1.1. Hierarchical clustering*

The results of the hierarchical clustering showed six clusters with similar healthcare expenditure from 2003 to 2021 (Figure 1). The information of each cluster, including size and countries, is depicted as follows:

(i) Cluster 0 (30 countries): The UK, Armenia, Canada, Austria, Switzerland, Netherlands, Sweden, Portugal, Belgium, Denmark, Japan, Spain, Malta, Australia, Finland, Lesotho, Norway, New Zealand, Maldives, Serbia, Brazil, Iceland, El Salvador, Argentina, Bosnia and Herzegovina, Slovenia, Namibia, Italy, Uruguay, and Greece.

(ii) Cluster 1 (46 countries or regions): Timor-Leste, Lebanon, Nicaragua, Panama, Czechia, Cyprus, Chile, Republic of Korea, Honduras, Mozambique, Latvia, Colombia, Bulgaria, North Macedonia, Georgia, Andorra, Ecuador, South Africa, Guinea-Bissau, Latin America and Caribbean (excluding those high-income countries), Bolivia, Croatia, Barbados, Paraguay, Tajikistan, Ukraine, San Marino, Israel, Lithuania, Slovak Republic, Costa Rica, Cambodia, Estonia, Malawi, Hungary, Rwanda, Jordan, Albania, Eswatini, Tunisia, Guatemala, Belarus, Poland, Botswana, Mexico, and Iran.

(iii) Cluster 2 (51 countries or regions): Uzbekistan, Russia, Jamaica, The Bahamas, Trinidad and Tobago, Mongolia, Cabo Verde, Samoa, Zambia, Dominica, Romania, Mauritius, Burkina Faso, Comoros, Caribbean small states, Tonga, Peru, Saudi Arabia, Philippines, Niger, Morocco, Grenada, Suriname, Luxembourg, Turkmenistan, Togo, Algeria, Nepal, China, Seychelles, Chad, Belize, Guyana, Dominican Republic, Uganda, Egypt, Vietnam, Türkiye, Kenya, Mali, Senegal, Bahrain, Ghana, Eritrea, Monaco, Madagascar, Haiti, Tanzania, Ethiopia, The Gambia, and Sudan.

(iv) Cluster 3 (1 country): The US.

(v) Cluster 4 (4 countries): Palau, Cuba, Germany, and France.

Cluster 5 (27 countries): Kuwait, Myanmar, Singapore, Fiji, United Arab Emirates, Iraq, Thailand, Azerbaijan, Malaysia, Vanuatu, Oman, Mauritania, Nigeria, Sri Lanka, Kazakhstan, Bhutan, Cameroon, Guinea, Indonesia, India, Angola, Pakistan, Qatar, Djibouti, Gabon, Benin, and Bangladesh.

It is noticeable that Clusters 0, 3, and 4 comprised developed economies, including the US, Japan, the UK, Nordic countries, Germany, and France, indicating that developed countries allocate a higher percentage of GDP than developing countries and emerging economies. Other

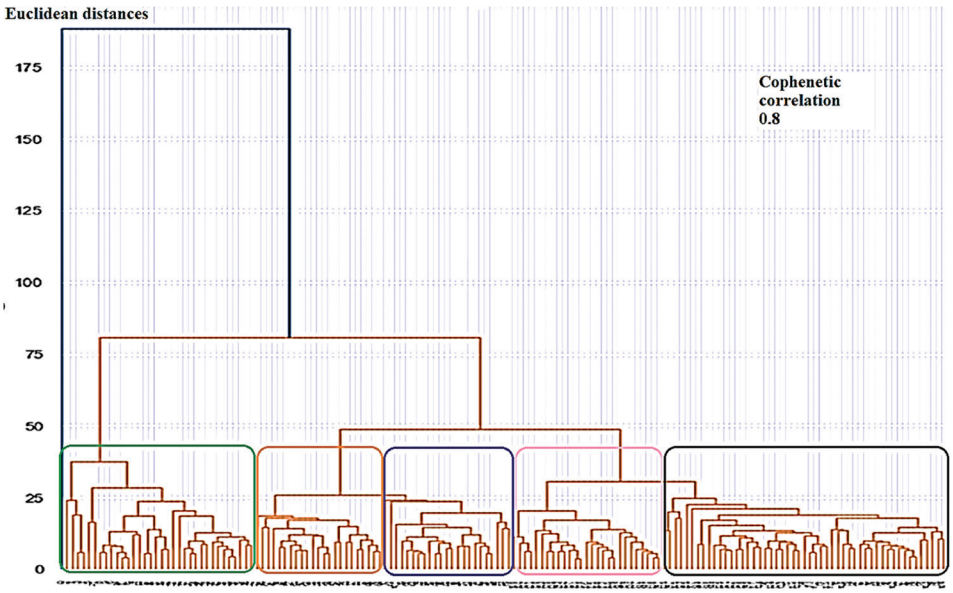



**Figure 1.** Hierarchical clustering graph of 158 countries
Notes: Green: Cluster 0; Orange: Cluster 1; Blue: Cluster 2; Pink: Cluster 3; Black: Cluster 4.

clusters consisted of countries from Eastern Europe, Latin America, Africa, the Caribbean, Asia, and the Middle East, implying that most countries in these regions tend to allocate a lower percentage of GDP to healthcare than other developed countries.

#### *3.1.2. SVMs*

The linear SVM, polynomial SVM, and Gaussian radial basis function (RBF) SVM models demonstrated accuracies of approximately 77, 71, and 81%, respectively. The graphs of 158 countries are shown in Figure 2. The US is in one cluster with a substantial difference from other countries. The linear SVM results showed straight lines across classes, indicating clear linear separations. However, it might also imply an oversimplification of the relationship across classes.

The polynomial SVM results demonstrated slightly curved boundaries compared to linear boundaries, suggesting the added complexity of the polynomial transformation was unnecessary. Moreover, the Gaussian model generated more complex and non-linear boundaries with curved and localized decision regions, implying potentially better capacity in capturing the actual association across classes. Its superior accuracy to other methods suggests that the data contain patterns, such as complicated overlapping classes and non-linearly separable data.

### 3.2. Predictive modeling

#### *3.2.1. ARIMA*

To analyze health expenditure trends across countries, a simple ARIMA model was initially fitted to each country. The model was specified with an order of (1, 1, 1), where the parameters represent the following: One lag for the autoregressive term, first-degree differencing to address non-stationarity, and one lag for the moving average term. This configuration utilized the values from one step prior for both autoregressive and moving average components while assuming stationarity of the time series as the primary requirement for ARIMA models.

However, visual inspection indicated that the data exhibited non-stationary behavior. To address this, first-degree differencing (d = 1) was applied, ensuring the data met the stationarity assumption. Figure 3 visualizes the predicted health expenditure of 25 selected countries in a map.

#### *3.2.2. Predicted values*

Table 1 presents the predicted health expenditures for the year 2025 across all selected countries, using both the

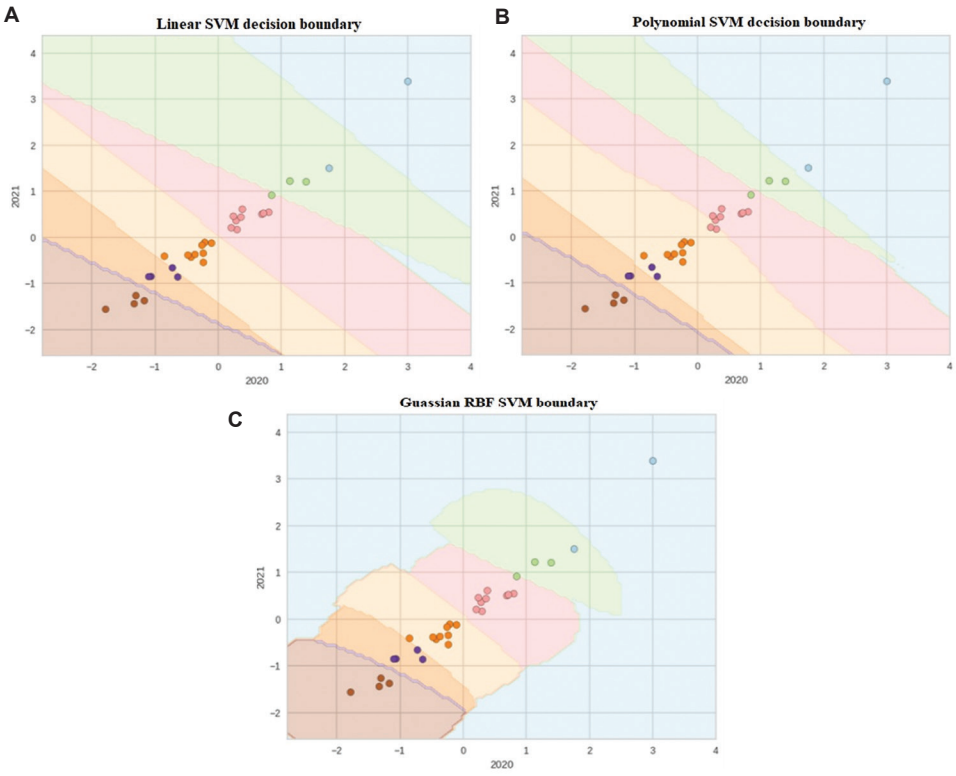



**Figure 2.** Decision boundaries of (A) linear, (B) polynomial, and (C) Gaussian radial basis function support vector machine models for 2020–2021

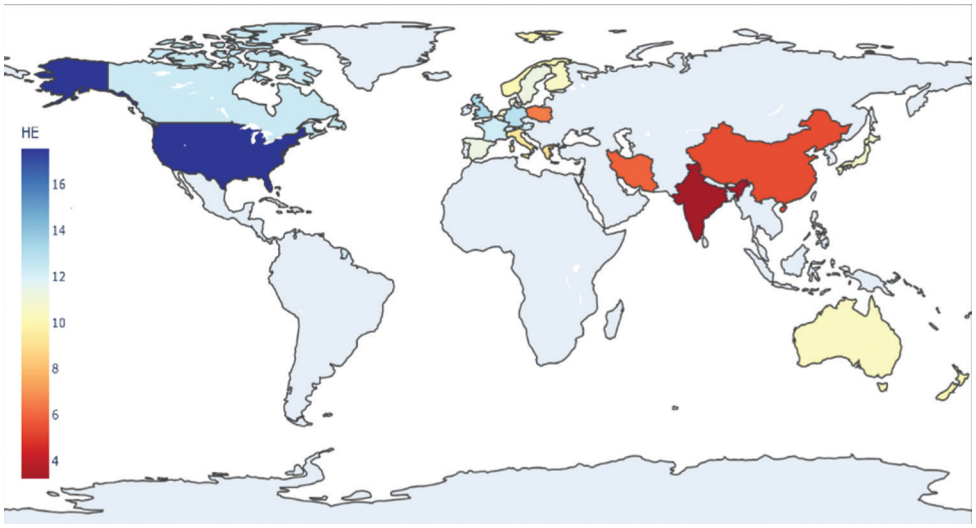



**Figure 3.** Predicted health expenditure of selected countries for 2025 using a simple autoregressive integrated moving average model (p = 1, d = 1, q = 1)

**Table 1. Predicted health expenditure as percentages of GDP for 2025 using a simple ARIMA model (p=1, d=1, q=1) and a multi-model ARIMA**

| Country | Simple ARIMA (%) | Multi-model ARIMA (%) |
|---|---|---|
| United States | 17.57 | 17.53 |
| United Kingdom | 13.13 | 12.36 |
| Switzerland | 11.77 | 11.80 |
| Sweden | 11.27 | 11.25 |
| Spain | 11.33 | 10.74 |
| Poland | 6.46 | 6.24 |
| Norway | 10.15 | 10.08 |
| New Zealand | 10.02 | 9.83 |
| Netherlands | 10.95 | 11.29 |
| Japan | 10.78 | 10.82 |
| Italy | 9.36 | 9.09 |
| Israel | 8.06 | 7.90 |
| Iran | 5.88 | 5.77 |
| India | 3.25 | 3.28 |
| Greece | 9.16 | 9.17 |
| Germany | 12.89 | 12.93 |
| France | 12.25 | 12.31 |
| Finland | 10.59 | 10.49 |
| Denmark | 11.03 | 10.82 |
| China | 5.36 | 5.38 |
| Canada | 12.38 | 12.42 |
| Belgium | 10.99 | 11.04 |
| Austria | 12.12 | 12.10 |
| Australia | 10.46 | 10.54 |

Abbreviations: ARIMA: Autoregressive integrated moving average; GDP: Gross domestic product.

simple ARIMA model and the enhanced multi-model approach. While the simple ARIMA model demonstrated favorable AIC scores, the multi-model approach achieved further reductions in AIC by fitting ARIMA models with all possible parameter combinations. This optimization led to losing their autoregressive component for most countries, except for Canada, New Zealand, Italy, and Poland.

The differences in predicted expenditures between the two approaches ranged from as little as 0.01 (for Greece) to 0.76 (for the UK). These results highlight the utility of the multi-model approach in refining parameter selection to enhance predictive accuracy. Subsequently, the fitted models were used to predict health expenditure for the next 4 years, with the results summarized in Figure 4, which presents the comprehensive prediction graphs for each country.

### 3.3. Generative AI

GANs play a transformative role in augmenting sparse and underrepresented datasets by generating synthetic yet realistic data. In the context of healthcare expenditure prediction, GANs can create additional data points for countries or regions where reliable data are scarce.

Figure 5 illustrates the progression of GAN training. In Figure 5A, epoch 0 represents the state before training, where the generated data (red points) were scattered randomly, showing no resemblance to the real data (blue points). This stage highlights the GAN's initial random generation, as the generator has not yet learned the underlying patterns of the real dataset.

The latent space dimension and batch size of trained data were 150 and 64, respectively. For epoch 0, the

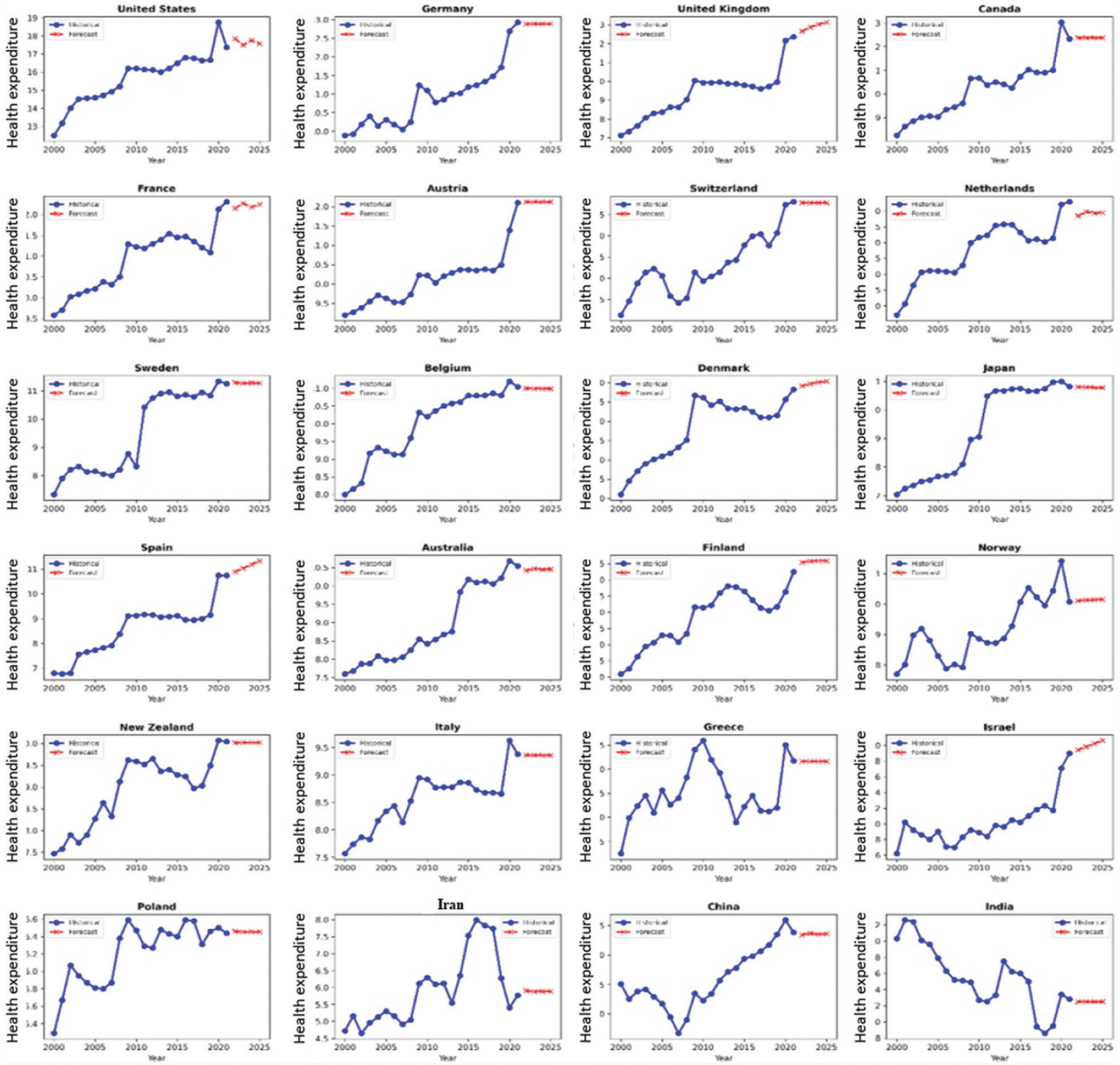



**Figure 4.** Predicted health expenditure of selected countries for 2022–2025 with a simple autoregressive integrated moving average model (p = 1, d = 1, q = 1)

discriminator loss was 1.396, implying that the generator makes many errors in classifying real and simulated samples. Meanwhile, the generator loss was 0.783, indicating the generation of fake data. In Figure 5B, epoch 2,000 presented the GAN's capability after extensive training. Here, the generated data aligned closely with the real data, demonstrating the generator's success in learning and replicating the distribution of healthcare expenditure patterns. This alignment indicates the potential of GANs to generate high-quality synthetic data, which can be used to address data sparsity issues in underrepresented regions or time periods. By filling gaps in datasets, GANs enable more accurate predictive modeling and equitable analysis of healthcare expenditures globally. Increasing the number of epochs to 2,000 reduced the discriminator and generator losses to 0.67 and 0.73, respectively.

Generative AI brings transformative advantages to healthcare expenditure analysis by addressing the limitations of traditional predictive models. One significant benefit is its ability to handle incomplete or imbalanced datasets through data augmentation. By generating synthetic yet realistic data, generative AI fills gaps in underrepresented

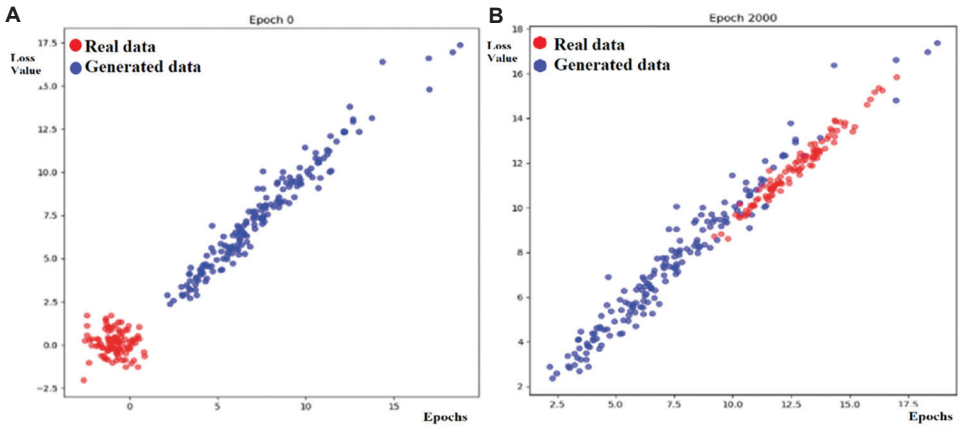



**Figure 5.** Outputs of generative adversarial networks at (A) epoch 0 and (B) epoch 2,000

regions, ensuring that models trained on such data are more inclusive and generalizable. This capability enhances the reliability of global healthcare expenditure predictions, particularly for low-income countries with limited historical data, thus fostering equitable decision-making.

## 4. Discussion

Despite its transformative potential, implementing generative AI in healthcare expenditure analysis faces significant challenges, particularly related to data quality and privacy. Generative AI models rely heavily on large, high-quality datasets to produce accurate and reliable outputs. However, healthcare expenditure data often suffer from inconsistencies, incomplete records, and regional disparities in reporting. For underrepresented countries or regions with scarce data, the generative AI model's ability to produce meaningful synthetic data might be compromised. In addition, concerns around data privacy and compliance with regulations, such as the General Data Protection Regulation and Health Insurance Portability and Accountability Act, pose obstacles to obtaining and utilizing sensitive healthcare-related information for training models.

Another major challenge is the computational complexity and resource requirements of generative AI models. Training sophisticated models, such as GANs or VAEs, requires substantial computational power, specialized infrastructure, and expertise, all of which may not be readily available in low-resource settings. Moreover, generative AI models often lack interpretability, making it difficult for policymakers and stakeholders to fully trust or understand the outputs. The potential for generating unrealistic or biased synthetic data further emphasizes the need for rigorous validation and oversight mechanisms. Overcoming these challenges requires collaboration across AI researchers, policymakers, and healthcare experts to ensure ethical, efficient, and equitable implementation of generative AI solutions. Validation processes include:

**Table 2. Validation methods and the accuracy of all models**

| Model | Data split | Validation method | Results |
|---|---|---|---|
| SVM (linear) | 70% train 15% validation 15% test | Cross-validation (k-fold) | Approximately 77% accuracy |
| SVM (polynomial) | 70% train 15% validation 15% test | Cross-validation (k-fold) | Approximately 71% accuracy |
| SVM (RBF) | 70% train 15% validation 15% test | Cross-validation (k-fold) | Approximately 81% accuracy (highest among SVM models) |
| ARIMA (simple) | Time-series split (train/test by year) | AIC, BIC, holdout year prediction | Accurate short-term forecast: AIC/BIC optimized |
| ARIMA (multi-model) | Time-series split (train/test by year) | Grid search for (p, d, q) combinations | Improved model fit for most countries |
| GAN | - (unsupervised) | Visual convergence and loss monitoring | Generator loss reduces to 0.73 and discriminator loss reduces to 0.67 by epoch 2000 |

Abbreviations: AIC: Akaike information criterion; ARIMA: Autoregressive integrated moving average; BIC: Bayesian information criterion; GAN: Generative adversarial network; RBF: radial basis function; SVM: Support vector machine.

(i) ARIMA models: AIC and BIC were used to select optimal parameters and evaluate forecast accuracy by comparing predicted values versus actual expenditures in a holdout year.
(ii) SVM models: Cross-validation was applied, and classification accuracy was reported across different kernels. The RBF kernel achieved the highest accuracy at 81%.
(iii) GAN models: Model fidelity was evaluated using

visual comparison of generated versus actual data distributions, and loss curves were monitored, with both generator and discriminator losses decreasing over training.

The validation and accuracy of the training, test, and validation sets are summarized in Table 2. A standard approach was used for data partitioning. The dataset was split into training (70%), validation (15%), and test (15%) sets based on a stratified sampling method to maintain the distribution of healthcare expenditure across different country groups. For the SVM models, cross-validation was applied during training to optimize model parameters and mitigate overfitting.

## 5. Conclusion

This study demonstrates the potential of integrating generative and traditional AI techniques to forecast healthcare expenditures globally, with a strong emphasis on addressing data sparsity and equity in LMICs. By applying a hybrid framework of GANs for synthetic data augmentation, SVMs for classification, and ARIMA for temporal forecasting, the robustness and accuracy of predictions were enhanced, particularly in data-scarce settings. The findings highlight that generative AI can mitigate the limitations of incomplete datasets, enabling a more representative modeling of global expenditure patterns. This methodological advancement is especially impactful for LMICs, where historical health data may be inconsistent or unavailable. By simulating realistic expenditure trends, the proposed framework equips policymakers with actionable forecasts tailored to their unique economic and demographic conditions. Ultimately, this work contributes to the development of a scalable and replicable AI-based model that supports equitable, evidence-based decision-making in global health finance, bridging a critical gap between advanced predictive technologies and real-world health policy needs.

## Acknowledgments

None.

## Funding

None.

## Conflict of interest

The authors declare they have no competing interests.

## Author contributions

*Conceptualization:* Taegeon Yu, Wen-Shan Liu, Payam Norouzzadeh, Eli Snir, Bahareh Rahmani

*Investigation:* Daipayan Bera, Abbas Maazallahi, Roschlynn Dsouza, Francina Pali, Eli Snir, Bahareh Rahmani

*Methodology:* Taegeon Yu, Daipayan Bera, Abbas Maazallahi, Roschlynn Dsouza, Francina Pali, Wen-Shan Liu, Bahareh Rahmani

*Writing – original draft:* Taegeon Yu, Daipayan Bera, Abbas Maazallahi, Roschlynn Dsouza, Francina Pali, Bahareh Rahmani

*Writing – review & editing:* Payam Norouzzadeh, Eli Snir, Bahareh Rahmani

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

You may download the data from https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS?most_recent_year_desc=true&locations=1W.

## References

1. *Global Medical Trends Survey*; 2024. Available from: https://www.wtwco.com/en-cz/insights/2023/11/2024-global-medical-trends-survey [Last accessed on 2025 Sep 18].

2. Meijer CD, Wouterse B, Polder J, Koopmanschap M. The effect of population aging on health expenditure growth: A critical review. *Eur J Ageing*. 2023;10(4):353-361.

   doi: 10.1007/s10433-013-0280-x

3. Global Burden of Disease Health Financing Collaborator. Past, present, and future of global health financing: A review of development assistance, government, out-of-pocket, and other private spending on health for 195 countries of 1995–2050. *Lancet*. 2019;393(10187):2233-2260.

   doi: 10.1016/S0140-6736(19)30841-4

4. Young AJ, Terry RF, Rottingen JA, Viergever RF. Global trends in health research and development expenditures-the challenge of making reliable estimates for international comparison. *Health Res Policy Syst*. 2015;13(1):7.

   doi: 10.1186/1478-4505-13-7

5. Wilensky G, Gordon H, Sun MD, *et al*. Global trends in biomedical R&D expenditures. *N Engl J Med*. 2014;370(1):3-6.

   doi: 10.1056/NEJMp1311068

6. Wager E, Rakshit S, Cox C. *What Drives Health Spending in the U.S. Compared to other Countries?* Washington, D.C: KFF Health Cost; 2024.

7. McGough M, Winger A, Kurani N, Cox C. *How Much is Health Spending Expected to Grow?* Washington, D.C: KFF Health Cost; 2024.

8. Glassman A, Keller JM, Smitham E. *The Future of Global Health Spending Amidst Multiple Crises*. United States: Center for Global Development; 2023.

9. Baltagi BH, Lagravinese R, Moscone F, Tosetti E. Health care expenditure and income: A global perspective. *Health Econ*. 2017;26(7):863-874.

doi: 10.1002/hec.3424

10. Gascon AE, Mico-Sanz JL, Ripolles AC. Global evidence of environmental and lifestyle effects on medical expenditures across 154 countries. *Prev Med Rep*. 2022;30:102036.

doi: 10.1016/j.pmedr.2022.102036

11. Chen F, Chen Z. Cost of economic growth: air pollution and health expenditure. *Sci Total Environ*. 2021;755(Pt 1):142543.

doi: 10.1016/j.scitotenv.2020.142543

12. Marley R. Top 10 current trends expected to transform healthcare in 2024. Healthcare Transformers. 2024. Available from: https://healthcaretransformers.com/digital-health/current-trends/current-trends-expected-to-transform-healthcare-in-2024/ [Last accessed on 2025 Sep 18].

13. Bertl M, Ross P, Draheim D. Systematic AI support for decision-making in the healthcare sector: obstacles and success factors. *Health Policy Technol*. 2023;12(3):100748.

14. Doupe P, Faghmous J, Basu S. Machine learning for health services researchers. *Value Health*. 2019;22(7):808-815.

doi: 10.1016/j.jval.2019.02.012

15. Langenberger B, Schulte T, Groene O. The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data. *PLoS One*. 2023;18(1):e0279540.

doi: 10.1371/journal.pone.0279540

16. Obermeyer Z, Emanuel EJ. Predicting the future, big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216-1219.

doi: 10.1056/NEJMp1606181

17. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: A practical introduction. *BMC Med Res Methodol*. 2019;19(1):64.

doi: 10.1186/s12874-019-0681-4

18. Khan JR, Awan N, Awan N, Islam M, Muurlink O. Healthcare capacity, health expenditure, and civil society as predictors of COVID-19 case fatalities: A global analysis. *Front Public Health*. 2020;8:347.

doi: 10.3389/fpubh.2020.00347

19. Mathauer I, Oranje M. Global medical trends survey report. *Bull World Health Organ*. 2024;102(3):216-224.

20. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-1358.

doi: 10.1056/NEJMra1814259

21. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med*. 2019;25:44-56.

doi: 10.1038/s41591-018-0300-7

22. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318.

doi: 10.1001/jama.2017.18391

23. Heaven WD. AI is changing medicine fast. *Nature*. 2023;614:234-237.

24. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. *J R Soc Med*. 2019;112(1):22-28.

doi: 10.1177/0141076818815510

25. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2:719-731.

doi: 10.1038/s41551-018-0305-z

26. Esteva A, Robicquet A, Ramsundar B, *et al*. A guide to deep learning in healthcare. *Nat Med*. 2019;25:24-29.

doi: 10.1038/s41591-018-0316-z

27. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health*. 2018;3:e000798.

doi: 10.1136/bmjgh-2018-000798

28. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. *N Engl J Med*. 2020;382:e82.

doi: 10.1056/NEJMp2005835

29. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep*. 2014;16(1):441.

doi: 10.1007/s11886-013-0441-8

30. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453.

doi: 10.1126/science.aax2342

31. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet*. 2020;395(10236):1579-1586.

doi: 10.1016/S0140-6736(20)30226-9

32. Chen M, Hao Y, Cai Y, Wang Y, Sun X. The role of big data and artificial intelligence in public health. *Biomed Res Int*. 2020;2020:1-10.

33. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep*. 2016;6:26094.

34. Wang L, Wang Y, Ye D, Liu Q. Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence. *Int J Antimicrob Agents*. 2020;55(6):105948.

doi: 10.1016/j.ijantimicag.2020.105948

35. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.

    doi: 10.1038/nature14539

36. Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. *Technol Forecast Soc Change*. 2018;126:3-13.

37. Krittanawong C, Johnson KW, Rosenson RS, *et al*. Deep learning for cardiovascular medicine: A practical primer. *Eur Heart J*. 2019;40(25):2058-2073.

    doi: 10.1093/eurheartj/ehz056

38. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94-98.

    doi: 10.7861/futurehosp.6-2-94

39. Hinton G. Deep learning technology with the potential to transform health care. *JAMA*. 2018;320(11):1101-1102.

    doi: 10.1001/jama.2018.11100

40. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA*. 2018;320(21):2199-2200.

    doi: 10.1001/jama.2018.17163

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

MINI-REVIEW

# Innovation management for artificial intelligence adoption in healthcare and biopharma: A mini-systematic review

**Thankgod Chimenem Kalagbor[1]*, Konstantin Koshechkin[2], Paul Ewa Oseshi[1], Samira Fatumata Sami[1], Josephine Ushang Adie[1], and Peter Ode Oto[1]**

[1]Department of Public Health and Healthcare, Faculty of Preventive Medicine, First Moscow State Medical University, Moscow, Russia

[2]Department of Information and Internet Technologies, Digital Health Institute, Faculty of Preventive Medicine, First Moscow State Medical University, Moscow, Russia

## Abstract

Recent advancements in artificial intelligence (AI) are reshaping core functions within healthcare and biopharmaceutical industries, particularly in diagnostics, personalized care, and drug development. However, the success of these innovations hinges on how well institutions manage their implementation. This systematic review investigates how innovation management influences AI adoption in healthcare and biopharma, highlighting both progress and persistent challenges. Following the preferred reporting items for systematic reviews and meta-analyses guidelines, this review was conducted using literature sourced from five major databases – PubMed, IEEE Xplore, Scopus, Web of Science, and Embase – focusing on peer-reviewed studies published between 2015 and 2024. A total of 82 studies were included, comprising 42 quantitative, 30 qualitative, and 10 mixed-methods studies. The population, intervention, comparison, and outcome framework guided study selection, while quality was assessed using the Joanna Briggs Institute checklist and Cochrane Risk of Bias 2.0 tool. Findings reveal that AI systems enable earlier disease detection, streamline patient triage, and improve operational workflows. In biopharma, companies, such as Moderna have shortened vaccine development timelines by integrating AI into molecular design. However, significant roadblocks remain, particularly regarding data privacy, infrastructure costs, and insufficient AI literacy among healthcare providers, especially in low- and middle-income countries. These barriers underscore the need for proactive innovation management approaches. To promote sustainable and ethical AI integration, this study recommends the development of governance frameworks, targeted workforce training, and increased interdisciplinary collaboration. As AI continues to evolve, managing its adoption thoughtfully will be essential to balancing technological potential with clinical realities and patient-centered care.

***Keywords:*** Artificial intelligence; Healthcare innovation; Biopharmaceutical industry; Innovation management; Artificial intelligence governance; Digital health; Ethical artificial intelligence; Healthcare administration

***Corresponding author:**
Thankgod Chimenem Kalagbor (kalagborthankgodc@gmail.com)

**Citation:** Kalagbor TC, Koshechkin K, Oseshi PE, Sami SF, Adie JU, Oto PO. Innovation management for artificial intelligence adoption in healthcare and biopharma: A mini-systematic review. *Artif Intell Health*. 2026;3(1):164-176. doi: 10.36922/AIH025180038

**Received:** April 29, 2025

**Revised:** June 18, 2025

**Accepted:** July 15, 2025

**Published online:** July 29, 2025

## 1. Introduction

The adoption of artificial intelligence (AI) in healthcare and biopharma is no longer a futuristic vision – it is actively reshaping how clinical and pharmaceutical services are delivered. From enhancing diagnostic precision to accelerating drug discovery timelines, AI has introduced tools that enable more informed and efficient decision-making at both the patient and organizational levels.[1] In an era where health systems are burdened by increasing populations, chronic disease prevalence, and rising demand for personalized care, AI offers scalable, data-driven solutions.[2,3]

Globally, countries are prioritizing investments in AI and digital health systems as part of broader strategies to strengthen health delivery. The Organisation for Economic Co-operation and Development (OECD) reported that nations with robust digital infrastructures were better equipped to manage the COVID-19 crisis, underscoring the importance of digital readiness and adaptability.[4] In low- and middle-income countries (LMICs), AI is being leveraged to close care gaps; for example, portable AI-powered ultrasound devices are improving breast cancer detection in remote African communities.[5]

Rather than remaining limited to rule-based systems, modern AI incorporates advanced machine learning and deep learning algorithms capable of processing vast, unstructured medical data. These technologies are now applied in radiology, early disease detection, predictive analytics, and patient monitoring, supporting clinicians with faster and more precise decisions.[6,7] In the pharmaceutical sector, AI models are reducing research bottlenecks by simulating molecular interactions, identifying viable drug targets, and optimizing clinical trial protocols, thereby compressing drug development timelines.[8]

However, successfully embedding AI into health systems requires more than technological capability. The OECD emphasizes that the main challenges to AI adoption are not purely technical but institutional, regulatory, and organizational.[9] Innovation management plays a pivotal role in ensuring AI integration is ethical, effective, and sustainable. This involves aligning AI development with policy frameworks, governance structures, and organizational readiness.

International organizations are also playing a more active role. The World Health Organization launched the global initiative on AI for Health to establish standards, promote ethical deployment, and guide countries in adopting AI responsibly.[10] The OECD similarly warns that without adequate oversight, AI systems may exacerbate inequality, compromise patient data security, or foster overdependence on opaque algorithms.[4]

Academic perspectives remain divided. While many researchers argue that AI can augment human judgment and enhance health outcomes, others express concern over biased algorithms, lack of transparency, and reduced clinician autonomy.[11,12] These contrasting views reinforce the need for robust governance, ethical foresight, and coordinated innovation management.

This article explores how innovation management can guide the responsible deployment of AI across healthcare and biopharma. By analyzing real-world applications, regulatory implications, and strategic approaches, this review offers practical insights into how stakeholders can align AI advancement with equitable, ethical, and patient-centered healthcare transformation.

## 2. Methodology

This systematic review was conducted in accordance with the preferred reporting items for systematic reviews and meta-analyses 2020 guidelines to ensure transparency, reproducibility, and methodological rigor in the identification, selection, and synthesis of studies.[13] The study followed a structured process, including the review question formulation, data collection, quality assessment, and findings analysis based on clearly defined criteria.

### 2.1. Study design

A systematic review approach was adopted to evaluate the role of innovation management in AI development within healthcare and biopharma. The review aimed to identify key strategies, applications, and challenges in managing AI-driven transformation, with a focus on ethical, operational, and regulatory implications.

### 2.2. Eligibility criteria

The inclusion and exclusion criteria were developed using the Population, Intervention, Comparison, and Outcome framework, which supports the formulation of focused and answerable research questions in systematic reviews.[14] The parameters included were:

(i) Population: Healthcare and biopharma sectors.
(ii) Intervention: Implementation of AI technologies.
(iii) Comparison: Traditional healthcare or research approaches.
(iv) Outcomes: Improvements in efficiency, accuracy, innovation management, and patient outcomes.

#### *2.2.1. Inclusion criteria*

The inclusion criteria were as follows:

(i) Peer-reviewed studies published between 2015 and 2024.
(ii) Articles discussing AI applications in healthcare and biopharma.

(iii) Studies addressing innovation management strategies for AI adoption.
(iv) Publications written in English.
(v) Full-text availability.

#### *2.2.2. Exclusion criteria*

The exclusion criteria were as follows:
(i) Studies unrelated to AI applications in healthcare or biopharma.
(ii) Articles focusing solely on technical AI algorithms without a managerial or strategic component.
(iii) Non-peer-reviewed literature (e.g., pre-prints, blog posts).
(iv) Conference abstracts without accessible full texts.

### 2.3. Data sources and search strategy

A comprehensive literature search was conducted across multiple databases, including PubMed, IEEE Xplore, Scopus, Web of Science, and Embase. The search terms included both MeSH terms and free-text keywords, such as "artificial intelligence in healthcare," "AI in drug discovery," "innovation management in AI," "digital health transformation," and "AI governance in biopharma." Boolean operators (AND, OR) were used to optimize and combine search terms. Reference lists of included articles were also manually screened for relevant additional studies.

### 2.4. Study selection and screening process

All identified articles were screened through a two-step process. First, two independent reviewers examined the titles and abstracts to exclude studies that did not meet the inclusion criteria. Second, the remaining full-text articles were assessed for eligibility. Discrepancies were resolved by a third reviewer. Cohen's Kappa coefficient was calculated to assess the inter-rater agreement during the selection process.

### 2.5. Data extraction and quality assessment

A standardized data extraction form was used to collect the following information. The extraction form included details, such as study title, author(s), and publication year, AI application domain (e.g., diagnostics, patient care, and drug development), innovation management frameworks or strategies discussed, and key findings, outcomes, and challenges reported.

To evaluate study quality, two established tools were applied: The Joanna Briggs Institute (JBI) Critical Appraisal Checklists and the Cochrane Risk of Bias 2.0 (RoB 2.0) tool. The JBI Critical Appraisal Checklists were used for both qualitative and quantitative studies to assess the methodological validity, relevance, and trustworthiness of the data. The JBI Manual for Evidence Synthesis offers structured guidance for applying these tools across different study designs.[15] The RoB 2.0 tool was used for assessing randomized studies, focusing on bias in randomization, deviations from intended interventions, missing outcome data, outcome measurement, and selective reporting.[15] Studies that were assessed as having a high risk of bias or scoring low on the quality assessment tools were excluded from the final synthesis.

### 2.6. Data synthesis and analysis

The data were analyzed using a narrative synthesis approach to identify recurring patterns and categorize findings into thematic areas, such as AI applications in healthcare and biopharma, innovation management frameworks for AI implementation, barriers and enablers of AI integration, and ethical and regulatory considerations. Where quantitative data were available (e.g., on adoption rates or AI effectiveness), descriptive statistics were used. Qualitative themes were identified using thematic analysis.

### 2.7. Ethical considerations

The ethical use of AI in healthcare requires strong regulatory frameworks to protect patient safety, privacy, and fairness. While regions, such as Europe have established regulations, such as the General Data Protection Regulation and the proposed AI Act, many African countries lack clear oversight, highlighting the need for locally relevant, culturally grounded AI ethics.

Bias in AI remains a major challenge, often stemming from unbalanced data or flawed models. Solutions include technical fixes, such as fairness-aware algorithms and organizational efforts, such as stakeholder involvement and ongoing monitoring.

Explainability is also crucial. Advanced AI models can be opaque, so tools, such as SHAP, LIME, and attention mechanisms help make decisions understandable to clinicians and regulators, building trust in AI systems.

Ultimately, ethical AI requires more than rules – it demands inclusive, transparent innovation management to ensure AI is both effective and equitable in real-world healthcare settings (Figure 1).

## 3. Results

In this review, key quality assessment criteria were applied across study types to only include high-quality and moderate-quality studies in the synthesis. The quantitative studies were assessed based on the following criteria: (i) Clearly defined inclusion criteria, (ii) detailed description of study setting and literature, (iii) valid and reliable measurement of variables, (iv) identification and appropriate management of confounding factors,

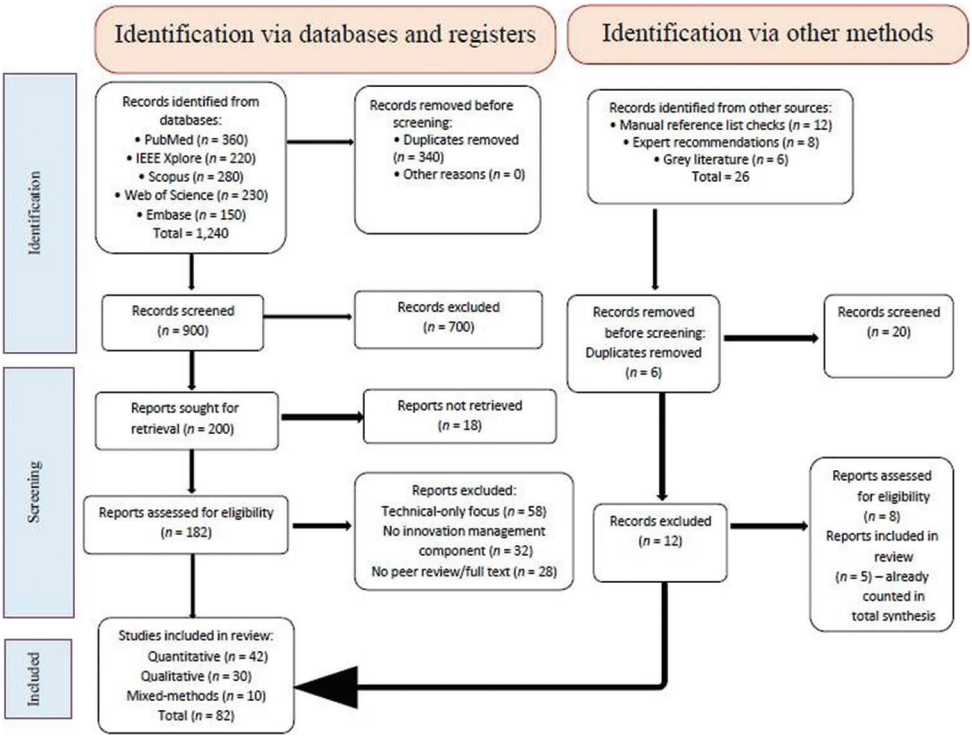



**Figure 1.** The preferred reporting items for systematic reviews and meta-analyses 2020 flowchart of the study selection process

(v) appropriate statistical analysis, and (vi) transparent reporting and follow-up (where applicable). For qualitative studies, the criteria were as follows: (i) Coherence between research aims and methodological approach, (ii) consistency between research design and data collection methods, (iii) clear representation of participants' perspectives, (iv) ethical considerations and approvals (if stated), (v) data-grounded interpretations, and (vi) researcher reflexivity and positionality are clearly stated.

Each item was assessed as "Yes," "No," "Unclear," or "Not Applicable". Studies with 80% or more "Yes: responses were classified as high quality, those with 60–79% were considered moderate quality, and studies scoring below 60% were deemed low quality and excluded from synthesis.

Subsequently, the findings of this review are categorized into key themes derived from the systematic literature review and case study analysis. These findings offer insights into AI adoption trends, challenges, and strategic approaches to effective innovation management.

## 3.1. AI adoption trends in healthcare and biopharma

### *3.1.1. AI-driven diagnostics and drug discovery*

Recent advancements in machine learning have significantly enhanced the accuracy of disease detection, particularly in radiology and pathology. AI systems now play a critical role in personalizing treatment plans by analyzing vast datasets to forecast patient responses. In the pharmaceutical sector, AI-driven approaches are streamlining drug discovery, predicting molecular interactions, optimizing clinical trial design, and reducing the time required for new drug development and market entry. In addition, funding models greatly influence access to AI-enabled personalized medicine. For instance, a study from Poland demonstrated that changing the funding model for genetic diagnostics between 2017 and 2019 substantially improved access to personalized oncology services, doubling genetic testing under hospital contracts and more than tripling of separately contracted services. This highlights the importance of adaptive funding strategies to support the scaling of advanced molecular and genetic testing for cancer diagnosis and treatment planning.[16]

### *3.1.2. Increased investment in AI innovation*

The past decade has witnessed a substantial increase in AI-focused research and development, driven by both private investment and public sector support. Governments are incentivizing innovation through targeted funding, while partnerships between academia and industry are fostering the co-creation of AI tools. At the same time, ethical and regulatory frameworks are gradually catching up to support safe and scalable AI integration. For example, an analysis of public payer

expenditures on drug programs in Poland between 2015 and 2018 revealed annual reimbursements ranging from approximately USD 635 million to USD 921 million, with oncology-related drug programs accounting for nearly half of this investment. Despite rising costs, such programs remain essential for improving patient access to innovative therapies and addressing diseases with limited treatment options, illustrating the significant financial commitment required to support AI-driven advances in drug discovery and development.[17]

### 3.2. Challenges in AI innovation management

#### *3.2.1. Ethical and regulatory barriers*

Despite its potential, AI adoption in healthcare remains constrained by complex ethical and legal considerations. Patient data privacy is a persistent concern, exacerbated by the lack of unified global regulations. In addition, the opacity of many AI algorithms, particularly in clinical decision-making, raises questions about bias, accountability, and explainability.

#### *3.2.2. Technological and operational challenges*

High infrastructure costs, limited digital literacy among healthcare professionals, and interoperability issues all present operational hurdles. AI integration often requires major changes in workflow and system architecture. These challenges can be particularly daunting for under-resourced institutions lacking the technical know-how and manpower for such tasks.

### 3.3. Strategic approaches to innovation management

#### *3.3.1. Development of AI governance frameworks*

Building trust in AI starts with transparent governance. Organizations are increasingly adopting governance models that prioritize ethical standards, patient safety, and regulatory compliance. Explainable AI models are gaining traction; offering clarity in clinical decisions and helping stakeholders make informed judgments. Meanwhile, continuous monitoring systems are being deployed to track AI performance and mitigate risks in real time.

#### *3.3.2. Workforce training and interdisciplinary collaboration*

Preparing the healthcare workforce for AI is as critical as AI itself. Institutions are conducting targeted training programs that equip clinicians with the skills needed to work alongside AI systems. Interdisciplinary teams comprising data scientists, healthcare providers, and regulatory experts are emerging as essential units for innovation. These collaborative efforts are proving vital for ensuring AI tools are not only functional but also aligned with clinical realities. These strategic elements are interdependent and form the foundation for successful AI integration. A conceptual framework summarizing these innovation management strategies is presented in Figure 2.

### 3.4. Case study findings

Real-world applications of AI in healthcare and biopharma demonstrate not only its technological promise but also the strategic maneuvers organizations have adopted to manage innovation effectively. Rather than isolated successes, these cases reflect broader patterns of collaboration, adaptation, and transformation that are reshaping the healthcare landscape.

CSL limited, a global biotechnology leader, offers a compelling example of how AI can be integrated into drug development pipelines to accelerate discovery and reduce research inefficiencies.[18] CSL has strategically invested in machine learning algorithms capable of predicting protein interactions and optimizing compound selection. By leveraging AI during early-stage research, CSL has significantly shortened the timeline for identifying promising drug candidates. More importantly, this integration is supported by a robust internal framework for data governance and cross-functional collaboration, underscoring the importance of structured innovation management in navigating both scientific and regulatory complexities.

Behold.ai, a United Kingdom-based medical technology company, has emerged as a key player in AI-driven diagnostics. Its red dot platform uses deep learning to interpret radiological scans with remarkable speed and accuracy, particularly in detecting lung cancer.[19] Behold.ai distinguishes itself not solely through the robust performance of its AI model but also through its strategic collaboration with the National Health Service to integrate the system into clinical workflows. This collaboration highlights the critical role of stakeholder engagement and regulatory alignment in scaling AI innovation. By ensuring that AI outputs are explainable and compliant with healthcare standards, Behold.ai demonstrates how ethical and operational concerns can be proactively managed within an innovation framework.

Moderna, widely recognized for its rapid COVID-19 vaccine development, exemplifies how AI can revolutionize vaccine research and development. In partnership with OpenAI and other data science entities, Moderna has incorporated large language models and predictive algorithms to refine antigen selection and optimize mRNA sequence designs.[20] This approach has not only accelerated pre-clinical testing but also improved responsiveness to

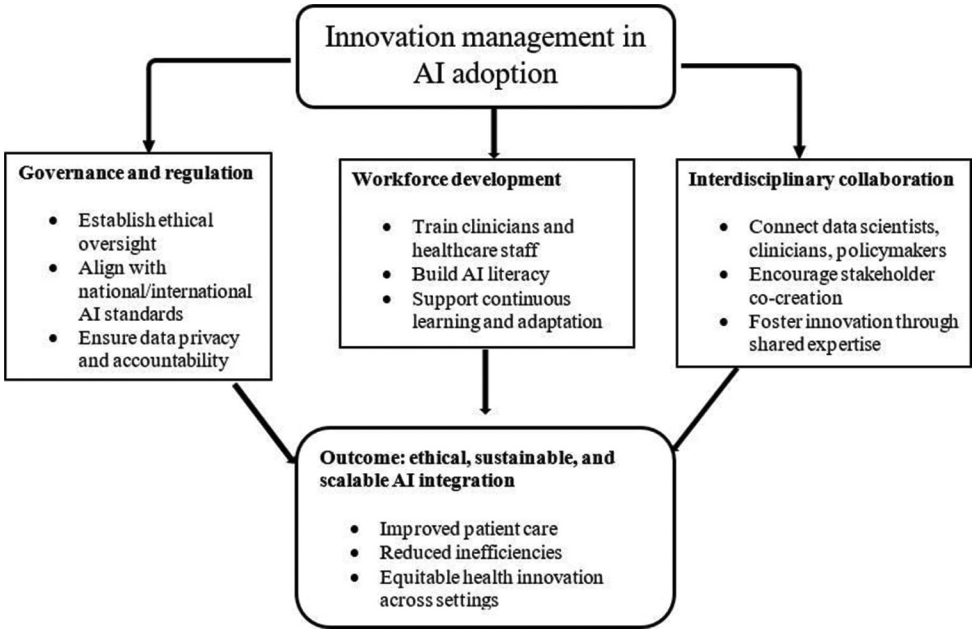



**Figure 2.** Conceptual framework illustrating key innovation management strategies for artificial intelligence (AI) adoption in healthcare and biopharma

emerging viral variants. Moderna's AI-driven strategies reflect a broader shift toward agile innovation ecosystems where data, modeling, and decision-making are well-integrated. Beyond the laboratory, this has reshaped public health response strategies, reinforcing the role of AI as a cornerstone of future-ready healthcare systems. These case studies illuminate the multidimensional nature of AI integration. They reveal that beyond technical capability, successful innovation depends on how organizations manage partnerships, align with regulations, and build internal capacity for continuous learning and ethical oversight (Table 1).

**Table 1. AI applications in healthcare**

| Organization | AI application | Key outcomes |
|---|---|---|
| CSL Limited | AI in drug development | Shortened drug discovery timelines by 40%; improved protein interaction predictions |
| Behold.ai | AI-driven diagnostics | Lung cancer detection with 94% accuracy; reporting time reduced by 22% |
| Moderna | AI in vaccine development | Accelerated mRNA development; enabled a rapid response to COVID-19 variants |

Abbreviation: AI: Artificial intelligence.

#### *3.4.1. AI applications in radiology*

AI has significantly enhanced radiological practices by improving diagnostic accuracy, streamlining workflows, and reducing turnaround times.[21] A previous study demonstrated that AI-assisted workflows reduced average radiology reporting times from 573 to 435 s without compromising diagnostic accuracy, indicating a substantial improvement in efficiency.[22] The United States Food and Drug Administration has approved over 340 AI tools for radiology, primarily for detection tasks, such as identifying brain tumors, strokes, and breast cancer. These tools assist in prioritizing urgent cases and improving diagnostic speed and accuracy.[23] Philips has developed AI-powered magnetic resonance imaging and computed tomography scanners that enhance diagnostic speed and accuracy. These innovations aim to alleviate pressure on health systems and improve patient care.[24] AI models have been developed to predict lung cancer risk from low-dose computed tomography scans, stratifying patients into different risk categories and guiding further diagnostic and treatment decisions.[25]

#### *3.4.2. AI during COVID-19*

During the COVID-19 pandemic, AI played a pivotal role in diagnostics, patient monitoring, and public health management. A scoping review identified 66 AI applications used in the clinical response to COVID-19, including tools for analyzing lung images, evaluating symptoms, monitoring vital signs, predicting infections, and aiding in breathing tube placement.[26]

In Singapore, a collaboration between Tan Tock Seng Hospital and research institutes developed "RadiLogic," a deep learning model that interprets chest radiographs quickly, prioritizing abnormal radiographs for early

review. The deployment of this solution resulted in a 22% reduction in turnaround times.[27] Researchers from Charles Darwin University and collaborators developed an AI model to diagnose pneumonia, COVID-19, and other lung diseases from lung ultrasound videos with an accuracy of 96.57%. This model also employs explainable AI techniques to assist radiologists in understanding and trusting the model's decisions.[28]

#### *3.4.3. AI in clinical trial design*

AI has transformed clinical trial design by enhancing patient recruitment, optimizing protocols, and improving data analysis.[29] Deloitte Insights reported that AI can reduce clinical trial cycle times while improving productivity and outcomes. AI algorithms assist in patient selection, site selection, and monitoring, thereby enhancing the efficiency of clinical trials.[29]

AstraZeneca has utilized AI to optimize trial protocols by analyzing historical trial data and real-world evidence. This approach has enabled the company to design more efficient protocols, resulting in faster trial execution and better resource allocation.[30]

#### *3.4.4. AI applications in LMICs*

While most of the literature on AI adoption originates from high-income countries, several LMICs are demonstrating innovative use cases of AI tailored to their specific healthcare challenges. In Nigeria, AI-powered portable ultrasound devices have improved breast cancer detection in rural settings where mammography services are scarce, enabling earlier diagnosis in women under 50.[31] In Brazil, during the COVID-19 pandemic, AI algorithms were deployed in public hospitals to triage patients, using imaging and vital sign data to prioritize Intensive Care Unit admissions and optimize limited critical care resources.[32] Similarly, India has piloted AI-based retinal imaging tools in rural clinics to screen for diabetic retinopathy, achieving a 30% increase in detection rates and reducing referral delays.[33] In the Philippines, AI-enhanced chatbots were integrated into public telemedicine platforms, allowing underserved populations to access remote symptom assessment and pandemic-related health information despite severe physician shortages.[34] These examples highlight how AI, when coupled with local adaptation and supportive innovation management, can bridge critical care gaps in resource-constrained environments.

### 3.5. Impact of AI on healthcare

AI is redefining innovation management in healthcare and biopharmaceuticals, driving groundbreaking advancements that enhance efficiency, accuracy, and patient outcomes. In 2024, the global AI healthcare market was valued at USD 10.94 billion, with projections estimating it will reach USD 15.48 billion by 2025 and USD 256.53 billion by 2033. This rapid growth is fueled by an annual expansion rate of 41.5% and is largely attributed to breakthroughs in deep learning, natural language processing, and AI-driven diagnostics that are transforming clinical decision-making and treatment strategies. In the United States, AI is accelerating medical innovation, with healthcare providers integrating machine learning models to streamline operations and enhance patient care. AI-powered imaging technologies have improved anomaly detection rates by 85%, significantly enhancing diagnostic precision. Meanwhile, AI-driven hospital management solutions have boosted operational efficiency by 45%, reducing administrative burdens and optimizing resource allocation. The biopharmaceutical industry is also undergoing a transformative shift, as AI-driven drug discovery is shortening research timelines by up to 60%. Advanced AI models facilitate precise molecular simulations, accelerating the identification of viable compounds and reducing the usual time required for drug development. Furthermore, AI-powered predictive analytics in hospitals has enhanced patient monitoring, lowering preventable medical errors by 50% and significantly improving healthcare outcomes. The post-pandemic era has further propelled AI-driven transformation, particularly in telemedicine, which has witnessed a 65% surge in adoption. AI-enhanced virtual consultations now offer real-time decision support, expanding healthcare accessibility and increasing patient engagement. In addition, wearable health devices integrated with AI provide continuous monitoring, with nearly half of users reporting improved health tracking and early intervention benefits. Globally, AI adoption in healthcare continues to accelerate, with over 80% of hospitals in developed regions implementing some form of AI-powered solutions, while emerging markets have experienced a 30% rise in AI-driven healthcare initiatives. Strategic collaborations between AI developers, healthcare institutions, and regulatory bodies are shaping a future where AI-driven healthcare is both scalable and ethically managed, ensuring its sustainable integration into diverse medical ecosystems.[35,36] (Figures 3 and 4).

Overall, based on the synthesis of the literature, several key findings can be highlighted. First, AI-driven diagnostics and drug discovery are experiencing rapid advancements. Moreover, ethical and regulatory concerns require structured governance mechanisms, while workforce training and interdisciplinary collaboration play a crucial role in AI adoption. In addition, case studies demonstrate the measurable benefits of AI in improving healthcare efficiency and drug development. These

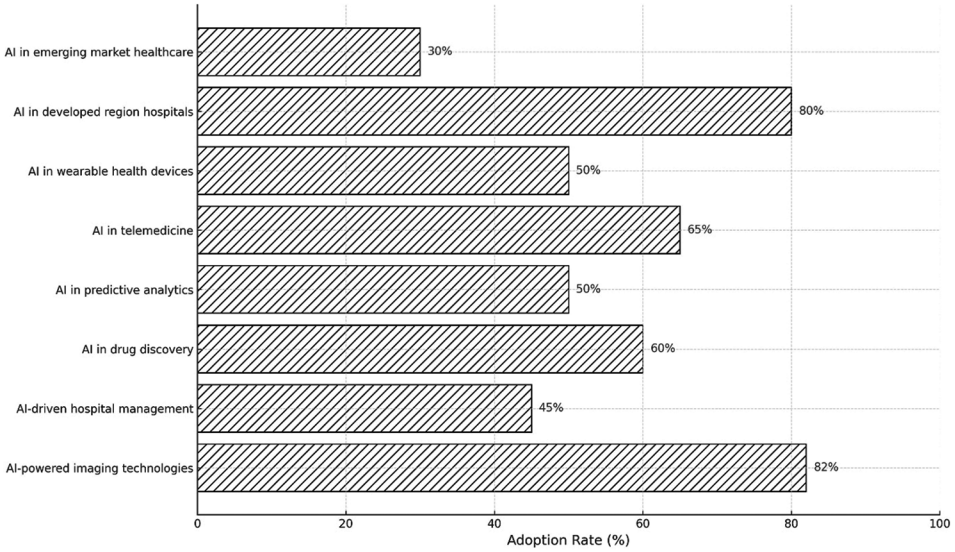



**Figure 3.** Artificial intelligence (AI) adoption trends in different healthcare sectors (2023 estimates). This figure illustrates the extent to which AI technologies have been integrated across various sectors of healthcare, including wearable health devices, telemedicine, predictive analytics, drug discovery, hospital management, and imaging technologies. Adoption trends refer to the percentage of healthcare institutions or organizations implementing AI-based tools or systems within their operational workflows. Data sourced from Global Growth Insights[35] and Fishchuk.[36]

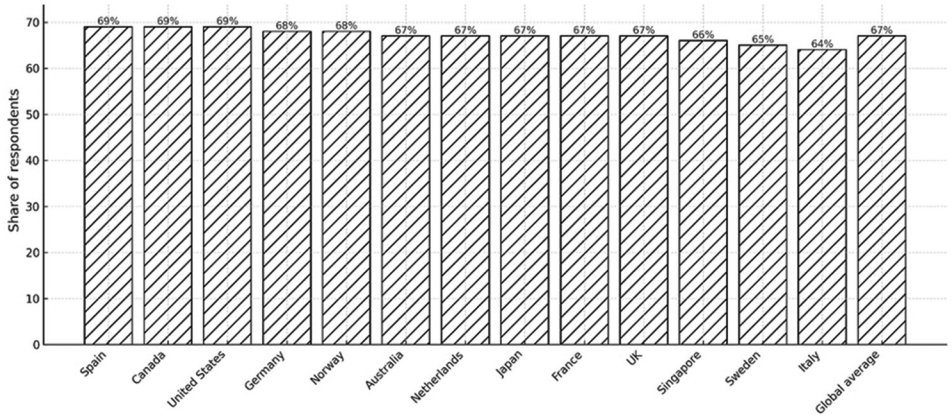



**Figure 4.** Consumer perceptions of artificial intelligence (AI) in healthcare across countries (Adapted from Statista, 2023). This chart displays the percentage of surveyed individuals in selected countries who consider AI-generated medical opinions helpful. Perception rates refer to public trust and perceived usefulness of AI recommendations in clinical or diagnostic settings. Data derived from a global consumer sentiment survey on AI applications in healthcare, conducted by Statista in 2023.[37]

findings emphasize the necessity of structured innovation management strategies to ensure the ethical, efficient, and impactful integration of AI in healthcare and biopharma (Figure 5).

## 4. Discussion

This systematic review emphasizes the transformative potential of AI in reshaping both healthcare delivery and pharmaceutical innovation. While AI has proven its ability to enhance diagnostics, personalize treatments, and accelerate drug discovery, its full potential depends on effective innovation management. This discussion synthesizes key implications, highlights strategic frameworks, and outlines challenges and opportunities for sustainable AI integration.

### 4.1. Implications of AI in healthcare and biopharma

AI applications have markedly improved diagnostic accuracy, early disease detection, and personalized care. For example, Topol[1] and Obermeyer and Emanuel[2] emphasized AI's role in reducing clinical errors and administrative burdens. In biopharma, AI-driven models

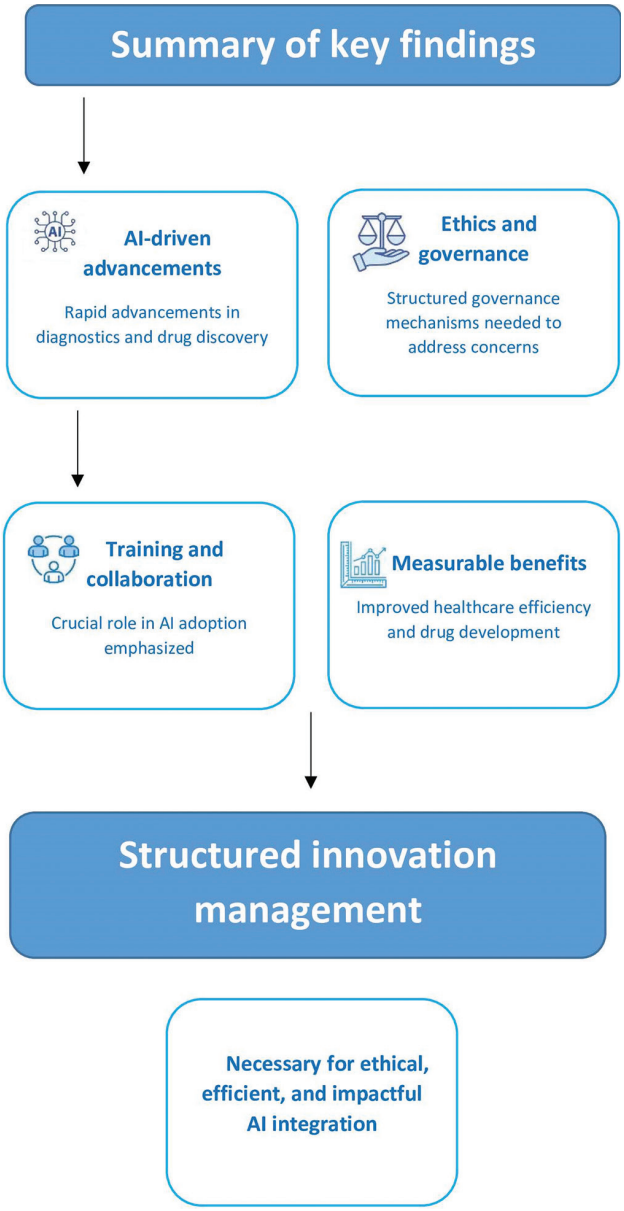



**Figure 5.** An overview of key findings on AI adoption in healthcare and biopharma. The diagram illustrates four central themes identified in the review: Rapid advancements in AI-driven diagnostics and drug discovery, the need for ethical and regulatory governance, the importance of workforce training and interdisciplinary collaboration, and the measurable benefits of AI integration. Collectively, these elements highlight the critical role of structured innovation management in guiding responsible and effective AI implementation across the healthcare and biopharmaceutical sectors.
Abbreviation: AI: Artificial intelligence.

expedited biomarker discovery and molecular design,[8] as illustrated by Moderna's use of AI in mRNA vaccine development during the COVID-19 pandemic.[20]

Such examples affirm that innovation management, through alignment of governance, training, and collaboration, enables AI to move from experimental phases to impactful clinical and industrial use.

### 4.2. Addressing ethical and regulatory challenges

Despite these benefits, ethical concerns and regulatory inconsistencies remain significant hurdles. The European Union's General Data Protection Regulation and proposed AI Act offer a model for risk-based AI governance.[38] However, many LMICs lack dedicated frameworks, resulting in fragmented oversight.[39] In addition, dominant ethical models shaped by the Global North may not adequately address localized priorities, prompting calls for "decolonized AI ethics" tailored to specific cultural and healthcare systems.[40]

To ensure equitable AI deployment, innovation strategies must include not only compliance mechanisms but also context-sensitive ethical guidance.

### 4.3. Overcoming technological and operational barriers

Infrastructure limitations, poor data access, and low digital literacy hamper AI adoption in many LMICs.[41] AI deployment in LMICs must be adapted to local realities, such as infrastructure gaps, limited internet access, and weak regulatory systems.[31,32] Unlike high-income countries, LMICs often face challenges with fragmented data and digital readiness. Real-world examples from Nigeria, Brazil, India, and the Philippines highlight that successful implementation depends on contextual adaptation, including mobile-based diagnostics and public–private partnerships.[33,34] To prevent widening health disparities, AI innovation strategies should include capacity-building, locally driven governance models, and culturally responsive ethical frameworks that reflect indigenous values and healthcare goals.[42] Countries, such as Thailand and Vietnam offer emerging models of regulatory adaptation.[43] Innovation management frameworks that incorporate stakeholder engagement, iterative feedback, and clinician involvement are key to overcoming such barriers.[38]

Institutions, such as Kaiser Permanente and initiatives, such as the United States Cancer Moonshot demonstrate how innovation management can enhance implementation fidelity and real-world impact, as long as over-reliance on unvalidated AI models is avoided.[44,45]

### 4.4. Strategies for effective innovation management

This review identified five core strategies essential to managing AI adoption:

(i) Governance frameworks to enable ethical oversight, accountability, and transparency[12,38]
(ii) Workforce training to build digital literacy among healthcare providers.[3]

(iii) Interdisciplinary collaboration to integrate perspectives from clinicians, data scientists, and regulators.[6]
(iv) Explainable and ethical AI to encourage adoption of interpretable models.[44,46]
(v) Regulatory alignment to advocate for international standards and harmonized policies.[11,43]

These strategies create an enabling ecosystem that supports both technological innovation and ethical integration.

### 4.5. Future directions and research opportunities

Technologies, such as blockchain, Internet of Medical Things, and federated learning are expanding the innovation landscape. Blockchain ensures data traceability in clinical trials, the Internet of Medical Things facilitates continuous patient monitoring, and federated learning enables secure, collaborative model training across institutions.

As the market for AI in healthcare grows, managing these innovations effectively – through agile, adaptive frameworks – will be critical to realizing their full potential.

### 4.6. Limitations of the study

This review is subject to certain limitations. The included studies varied widely in design, scope, and outcome measures, thereby reducing comparability. Moreover, articles not in English or not peer-reviewed were excluded. Some AI models reviewed lacked long-term performance data or robust validation. Hence, future work should explore meta-analytical approaches and consider grey literature and non-English sources for broader inclusivity.

### 4.7. Future directions

#### *4.7.1. Blockchain in healthcare*

Blockchain technology offers a decentralized and immutable ledger system, ensuring secure and transparent handling of sensitive health data. Key applications include:
(i) Electronic health records: Blockchain facilitates secure sharing and management of electronic health records, enhancing interoperability among healthcare providers.
(ii) Clinical trials: It ensures data integrity and transparency in clinical trial processes, reducing the risk of data manipulation.[47]
(iii) Supply chain management: Blockchain enhances the traceability of pharmaceuticals, combating counterfeit drugs.

#### *4.7.2. Internet of medical things*

The Internet of Medical Things refers to the network of interconnected medical devices and applications that collect and transmit health data. Its applications include:[48]
(i) Remote patient monitoring: Wearable devices track vital signs, enabling continuous monitoring and early detection of health issues.
(ii) Chronic disease management: Internet of Medical Things devices assist in managing conditions, such as diabetes and hypertension by providing real-time data to healthcare providers.
(iii) Emergency response: Connected devices can alert medical personnel during emergencies, improving response times.

#### *4.7.3. Federated learning in healthcare*

Federated learning enables multiple institutions to collaboratively train machine learning models without sharing raw data, preserving patient privacy. Applications include:[49,50]
(i) Collaborative research: Hospitals can jointly develop predictive models for disease diagnosis without compromising data security.
(ii) Personalized medicine: Federated learning supports the creation of individualized treatment plans by analyzing diverse datasets.
(iii) Pandemic response: It facilitates the rapid development of models to track and predict disease outbreaks across regions.

## 5. Conclusion

AI holds immense promise for reshaping healthcare and biopharma. However, this promise can only be fulfilled through deliberate, well-managed innovation that bridges the gap between cutting-edge algorithms and real-world patient outcomes. Policymakers, healthcare leaders, and technologists must work collaboratively to ensure that AI is not only advanced but also accessible, ethical, and transformative in the truest sense. For this to occur equitably innovation strategies must actively account for the unique challenges and opportunities present in LMICs. Investment in digital infrastructure, regional regulatory capacity, and context-aware governance frameworks is essential to prevent the exacerbation of global health disparities. The next wave of AI advancement must be both inclusive and adaptive, anchored in principles that prioritize equity, sustainability, and human-centered care across all healthcare systems.

## Acknowledgments

None.

## Funding

None.

## Conflict of interest

The authors declare that they have no competing interests.

## Author contributions

*Conceptualization:* All authors
*Visualization:* All authors
*Writing – original draft:* All authors
*Writing – review & editing:* All authors

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Availability of data

Not applicable.

## References


1. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York City: Basic Books; 2019.
2. Obermeyer Z, Emanuel EJ. Predicting the future-big data, machine learning, and clinical medicine. *N Engl J Med.* 2016;375(13):1216-1219.
   doi: 10.1056/NEJMp1606181
3. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The Practical implementation of artificial intelligence technologies in medicine. *Nat Med.* 2019;25:30-36.
   doi: 10.1038/s41591-018-0307-0
4. OECD. *AI in Health;* 2024. Available from: https://www.oecd.org/en/publications/ai-in-health_2f709270-en.html [Last accessed on 2025 Jul 17].
5. World Economic Forum. *Why AI has a Greater Healthcare Impact in Emerging Markets*; 2024. Available from: https://www.weforum.org/stories/2024/06/ai-healthcare-emerging-markets [Last accessed on 2025 Jul 17].
6. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115-118.
   doi: 10.1038/nature21056
7. Rajpurkar P, Irvin J, Zhu K, *et al*. *Radiologist-level Pneumonia Detection on Chest X-rays with Deep Learning*. arXiv [Preprint]; 2017. Available from: https://arxiv.org/abs/1711.05225 [Last accessed on 2025 Jul 17].
8. Zhavoronkov A. Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry. *Mol Pharm.* 2018;15(10):4311-4313.
   doi: 10.1021/acs.molpharmaceut.8b00930
9. OECD. Digital health at a glance. In: *Health at a Glance 2023: OECD Indicators*; 2023. Available from: https://www.oecd.org/en/publications/health-at-a-glance-2023_7a7afb35-en/full-report/digital-health-at-a-glance_86518984.html [Last accessed on 2025 Jul 17].
10. World Health Organization. *Global Initiative on AI for Health;* 2023. Available from: https://www.who.int/initiatives/global-initiative-on-ai-for-health [Last accessed on 2025 Jul 17].
11. London AJ. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Cent Rep.* 2019;49(1):15-21.
    doi: 10.1002/hast.973
12. Morley J, Machado CCV, Burr C, *et al*. The ethics of AI in health care: A mapping review. *Soc Sci Med.* 2020;260:113172.
    doi: 10.1016/j.socscimed.2020.113172
13. Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71.
    doi: 10.1136/bmj.n71
14. JBI. *JBI Manual for Evidence Synthesis;* 2024. Available from: https://jbi-global-wiki.refined.site/space/manual/4685837/jbi%20manual%20for%20evidence%20synthesis [Last accessed on 2025 Jul 18].
15. Sterne JAC, Savović J, Page MJ, *et al*. RoB 2: A revised tool for assessing risk of bias in randomized trials. *BMJ.* 2019;366:l4898.
    doi: 10.1136/bmj.l4898
16. Mela A, Rdzanek E, Tysarowski A, *et al*. The impact of changing the funding model for genetic diagnostics and improved access to personalized medicine in oncology. *Expert Rev Pharmacoecon Outcomes Res.* 2023;23(1):43-54.
    doi: 10.1080/14737167.2023.2140139
17. Mela A, Poniatowski ŁA, Drop B, *et al*. Overview and analysis of the cost of drug programs in Poland: Public payer expenditures and coverage of cancer and non-neoplastic diseases related drug therapies from 2015-2018. *Front Pharmacol.* 2020;11:1123.
    doi: 10.3389/fphar.2020.01123
18. Evans M. *CSL Taps AI to Power Drug Development*. Proactive Australia; 2024. Available from: https://www.proactiveinvestors.com.au/companies/news/1055325/csl-taps-ai-to-power-drug-development-1055325.html [Last accessed on 2025 Mar 18].
19. Behold AI. *AI Tool Shows Potential to Improve the Speed of Lung Cancer Detection*. Behold.ai News; 2020. Available from; https://www.behold.ai/news/ai-tool-shows-potential-to-improve-speed-of-lung-cancer-detection [Last accessed on 2025 Mar 18]
20. Moderna. *Moderna and Open AI Collaborate to Advance mRNA Medicine*. Moderna News; 2024. Available from: https://investors.modernatx.com/news/news-details/2024/

moderna-and-openai-collaborate-to-advance-mrna-medicine/default.aspx [Last accessed on 2025 Mar 18].

21. El Naqa I, Giger ML, Ten Haken RK. Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century. *Br J Radiol.* 2020;93(1106):20190855.

doi: 10.1259/bjr.20190855

22. Acosta JN, Dogra S, Adithan S. *The Impact of AI Assistance on Radiology Reporting: A Pilot Study Using Simulated AI Draft Reports.* arXiv [Preprint]; 2024. Available from: https://arxiv.org/abs/2412.12042 [Last accessed on 2025 Jul 18].

23. Washington Post. *AI Hasn't Killed Radiology, But It Is Changing It*; 2025. Available from: https://www.washingtonpost.com/health/2025/04/05/ai-machine-learning-radiology-software [Last accessed on 2025 Jul 18].

24. The Guardian. *Do I Think Doctors are Going to Be Out of A Job? Not at All': the Ex-radiologist Bringing AI to Healthcare*; 2025. Available from: https://www.theguardian.com/business/2025/mar/04/do-i-think-doctors-are-going-to-be-out-of-a-job-not-at-all-the-ex-radiologist-bringing-ai-to-healthcare [Last accessed on 2025 Jul 18].

25. Toxigon. *AI in Radiology: Real-World Case Studies and Insights.* Available from: https://toxigon.com/ai-in-radiology-case-studies [Last accessed on 2025 Jul 18].

26. RAND Corporation. *Artificial Intelligence in the COVID-19 Response: Volume 1, Applications Used in the Clinical and Public Health Response to COVID-19*; 2023. Available from: https://www.rand.org/pubs/external_publications/EP70086.html [Last accessed on 2025 Jul 18].

27. Sim JZ, Bhanu Prakash KN, Huang WM, Tan CH. Harnessing artificial intelligence in radiology to augment population health. *Front Med Technol.* 2023;5:1281500.

doi: 10.3389/fmedt.2023.1281500

28. Courier Mail. *CDU Probes How AI Could Help Diagnose Diseases;* 2025. Available from: https://www.couriermail.com.au/news/charles-darwin-university-has-begun-a-study-into-artificial-intelligence-and-how-it-can-help-diagnose-diseases/news-story/cb2be167c8f60cb519cbe4cd323e0536 [Last accessed on 2025 Jul 18].

29. Deloitte Insights. *Artificial Intelligence in Clinical Trials*; 2020. Available from: https://www2.deloitte.com/us/en/insights/industry/life-sciences/artificial-intelligence-in-clinical-trials.html [Last accessed on 2025 Jul 18].

30. Clinical Research Trends. *The Use of Artificial Intelligence in Clinical Trial Design.* Available from: https://www.clinicalresearchtrends.net/the-use-of-artificial-intelligence-in-clinical-trial-design [Last accessed on 2025 Jul 18].

31. Olatunji TA, Adedokun BO. Artificial intelligence-based mobile ultrasound screening in low-resource Nigerian settings. *BMC Health Serv Res.* 2022;22:745.

doi: 10.1186/s12913-022-08745-z

32. Morales HM, Guedes M, Silva JS, Massuda A. COVID-19 in Brazil-preliminary analysis of response supported by artificial intelligence in municipalities. *Front Digit Health.* 2021;3:648585.

doi: 10.3389/fdgth.2021.648585

33. Rajalakshmi R, Arulmalar S, Usha M, Narayanasamy A, Ramachandran A, Mohan V. Validation of smartphone-based retinal photography for diabetic retinopathy screening. *PLoS One.* 2015;10(9):e0138285.

doi: 10.1371/journal.pone.0138285

34. Del Rosario R. AI-enhanced telehealth access in the Philippine public sector during COVID-19. *Int J Med Inform.* 2022;160:104706.

doi: 10.1016/j.ijmedinf.2022.104706

35. Global Growth Insights. *Artificial Intelligence in Healthcare Market Size, Share, Growth, and Industry Analysis*; 2024. Available from: https://www.globalgrowthinsights.com/market-reports/artificial-intelligence-in-healthcare-market-104419 [Last accessed on 2025 Mar 19].

36. Fishchuk I. *Adopting AI in Healthcare: Benefits, Challenges and Real-Life Examples*; 2024. https://leobit.com/blog/adopting-ai-in-healthcare-benefits-challenges-and-real-life-examples [Last accessed on 2025 Mar 19].

37. Statista. *Share of Consumers Worldwide Who Believed That Medical Opinions or Suggestions from Generative AI Would Be Helpful as of 2023, by Selected Countries*; 2023. Available from: https://www.statista.com/statistics/1418167/trust-in-medical-advice-from-generative-ai-by-country [Last accessed on 2025 Mar 13].

38. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A Narrative review. *Heliyon*. 2024;10(4):e26297.

doi: 10.1016/j.heliyon.2024.e26297

39. Townsend BA, Sihlahla I, Naidoo M, Naidoo S, Donnelly DL, Thaldar DW. Mapping the regulatory landscape of AI in healthcare in Africa. *Front Pharmacol*. 2023;14:1214422.

doi: 10.3389/fphar.2023.1214422

40. Grancia MK. Decolonizing AI ethics in Africa's healthcare: An ethical perspective. *AI Ethics.* 2025;5:3129-3142.

doi: 10.1007/s43681-024-00650-z

41. Victor A. Artificial intelligence in global health: An unfair future for health in Sub-Saharan Africa? *Health Aff Sch*. 2025;3(2):qxaf023.

doi: 10.1093/haschl/qxaf023

42. Nweke CI, Adewopo J. Decolonizing AI ethics in Africa's healthcare: An ethical perspective. *AI Ethics.* 2024;4:98-111.

doi: 10.1007/s43681-024-00650-z

43. *AI Regulation in Healthcare around the World: What is the*

*Status Quo*? medRxiv [Preprint]; 2025. Available from: https://www.medrxiv.org/content/10.1101/2025.01.25.25321061v1.full [Last accessed on 2025 Jul 18].

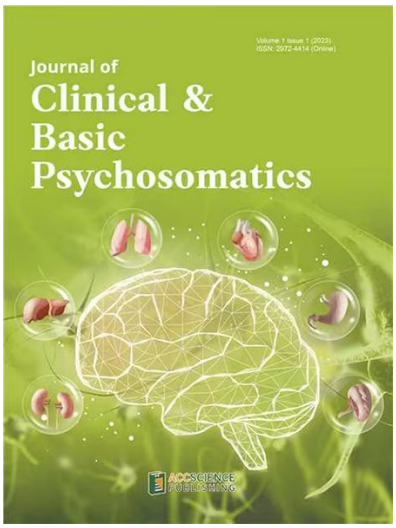
44. *What AI Can Do in Healthcare-and What It Should Never Do*. Wall Street Journal. Available from: https://www.wsj.com/tech/ai/what-ai-can-do-in-healthcareand-what-it-should-never-do-3e28a2b4 [Last accessed on 2025 Jul 18].
45. *How AI Can Make Cancer Treatment More Equitable.* Time. Available from: https://time.com/6325485/ai-cancer-moonshot [Last accessed on 2025 Jul 18].
46. Amann J, Blasimme A, Vayena E, Frey D, Madai IV. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20:310.

    doi: 10.1186/s12911-020-01332-6
47. Liu CZ, He D. Blockchain in healthcare applications: Research challenges and opportunities. *J Netw Comput Appl.* 2019;135:62-75.

    doi: 10.1016/j.jnca.2019.02.027
48. Al Khatib I, Shamayleh A, Ndiaye M. Healthcare and the internet of medical things: Applications, challenges, and solutions. *Informatics.* 2024;11(3):47.

    doi: 10.3390/informatics11030047
49. Joshi M, Pal A, Sankarasubbu M. *Federated Learning For Healthcare Domain-Pipeline, Applications And Challenges*. arXiv [Preprint]; 2022. Available from: https://arxiv.org/abs/2211.07893 [Last accessed on 2025 Jul 18].
50. OECD. Digital health. In: *Health at a Glance 2023: OECD Indicators;* 2023. Available from: https://www.oecd.org/en/publications/health-at-a-glance-2023_7a7afb35-en/full-report/digital-health_d79d912b.html [Last accessed on 2025 Jul 18].

# Our Journals



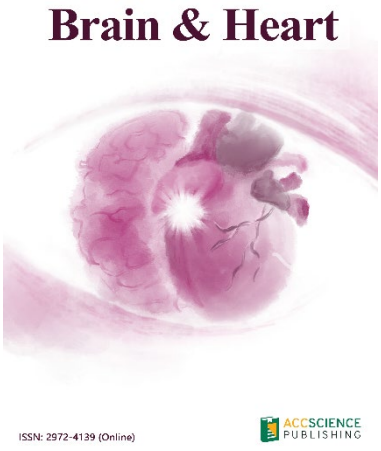

*Journal of Clinical and Basic Psychosomatics (JCBP)* is a quarterly journal focusing on clinical and basic research on symptoms, assessment, treatment, management, and the mechanism of psychosomatic disorders. *Journal of Clinical and Basic Psychosomatics* covers subject areas, including but not limited to the following:

- Conceptualization and classification of psychosomatic medicine
- Mechanism, biological markers, brain images, and treatment studies
- Psychosomatic reactions, syndromes, disorders, and diseases
- Psychosomatic disorders treated in general hospitals, including endocrinology, neurology, gastroenterology, dermatology, pain management, oncology, rheumatology, and other departments
- Psychological evaluation, management, rehabilitation, resilience training, and psychotherapy for general and specific populations during the pandemic
- Physiological disorders related to psychological factors (eating disorders, sleeping disorders, and sexual dysfunction)
- Somatic symptoms and related disorders and mental disorders due to somatic disease

*Brain & Heart* focuses on neurocardiology, a neurology and cardiology-based interdisciplinary subject that studies the circulatory mechanism of the human body, as well as the mechanisms of the interplay between the cardiovascular system and the nervous system. The journal's scope includes:

Clinical and basic research on diseases related to the circulatory and nervous systems, such as: orthostatic dizziness, orthostatic hypotension, autonomic dysfunction, and the relationship between the autonomic nervous system and the circulatory function in cerebral degeneration;

Heart-brain research on patients with syncope, autonomic dysfunction, cryptogenic stroke, and stroke with atrial fibrillation; research on the relationship between structural heart diseases and nervous system diseases, the correlation between cardiac electrophysiology and abnormal organizational structures and the pathogenesis of stroke, as well as new ways of diagnosis, treatment and prevention of unexplained stroke.




## Start a new journal

Write to us via email if you are interested to start a new journal with AccScience Publishing. Please attach your CV, professional profile page and a brief pitch proposal in your email. We shall inform you of our decision whether we are interested to collaborate in starting a new journal.

**Contact:** info@accscience.com

https://www.accscience.com/journal/AIH

ACCSCIENCE PUBLISHING

Contact
www.accscience.com
9 Raffles Place, Republic Plaza 1 #06-00 Singapore 048619
Email: editorial@accscience.com
Phone: +65 8182 1586