ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Enhancing medical data quality using hybrid machine learning models: A comparative study of isolation forest and support vector machine on numerically encoded clinical text

Raed Abdullah Althabeti[1]*, Ebeid Ali Ebeid[2], Hany Maher Sayed Lala[2], and Kamal Abdelraouf Eldahshan[2]

[1]Information Systems Department, Faculty of Information Technology and Computer Science, University of Saba Region, Marib, Marib, Yemen

[2]Mathematics Department, Faculty of Science, Al-Azhar University, Cairo, Egypt

Abstract

Ensuring high-quality medical text data is a persistent challenge in healthcare analytics, as issues such as inconsistency, incompleteness, and hidden anomalies can substantially compromise the reliability of clinical decision-making and downstream research outcomes. This article presents and assesses a hybrid isolation forest (IF)–support vector machine (SVM) model (HIFSVM) to improve the quality of medical text data via anomaly detection, comparing with IF and SVM. The research used three public health datasets: Alzheimer's Disease and Healthy Aging Data (ADHAD), Breast Cancer Global Dataset, and 500 Cities: The Local Data for Better Health (500LDFB), and employed 10 extensive data quality metrics: accuracy, completeness, consistency, timeliness, validity, uniqueness, precision, change rate, error rate, and data density. The models' anomaly detection performance was evaluated using synthetic ground truth, cross-validation, and expert validation. The HIFSVM model reported increased change rates (e.g., 0.96 for ADHAD's data value and 1.00 for 500LDFB's confidence limits), demonstrating considerable potential for tackling intricate or dynamic medical datasets that necessitate sophisticated anomaly detection. It also demonstrated enhanced detection accuracy (F1-scores: 0.89–0.91 across datasets), markedly surpassing IF (0.83–0.87) and SVM (0.74–0.79) (both $p < 0.05$). Validation established that 79–91% of the identified anomalies were clinically significant, exhibiting considerable inter-rater agreement ($\kappa = 0.71$–0.84). These findings demonstrate that HIFSVM proficiently detects authentic anomalies while maintaining robust data quality evaluation capabilities. The study provides a structured methodological framework for data preprocessing, feature engineering, and anomaly detection that healthcare professionals can employ to enhance the reliability of clinical decision-making, thereby improving the accuracy of clinical studies and ensuring better patient outcomes.

Keywords: Anomaly detection; Data quality assessment; Hybrid machine learning; Isolation forest; Support vector machine; Medical datasets

***Corresponding author:**
Raed Abdullah Althabeti
(ralthabeti@usr.ac)

Citation: Althabeti RA, Ebeid EA, Lala HMS, Eldahshan KA. Enhancing medical data quality using hybrid machine learning models: A comparative study of isolation forest and support vector machine on numerically encoded clinical text. *Artif Intell Health*. 2026;3(4):026080021. doi: 10.36922/AIH026080021

Received: February 22, 2026

Revised: March 31, 2026

Accepted: April 1, 2026

Published online: May 8, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Contamination of medical texts is a significant issue in clinical research and microbiology, resulting in incorrect diagnoses, flawed treatment strategies, and misleading study findings.[1] Misidentification of cell lines, blood culture contamination, and errors in data entry or transcription could be sources of this problem. There is a need to implement more stringent protocols for cell line authentication. Enhanced data discussions suggest that utilizing advanced medical text data with anomaly detection can improve predictive accuracy. Strategies to minimize blood culture contamination include effective skin disinfection techniques, such as chlorhexidine or iodine-based solutions, which have been shown to significantly reduce the risk of contamination.

Researchers are utilizing machine learning (ML) to improve the quality of medical textual data. Natural language processing (NLP) approaches transform unstructured medical text into structured data, facilitating tasks such as sentiment analysis and text summarization.[2] Research indicates that predictive algorithms, such as logistic regression, support vector machines (SVMs), and artificial neural networks, can anticipate patient outcomes and treatment responses. ML can improve data quality by identifying and rectifying errors, while automated data curation techniques highlight irregularities, ultimately leading to more accurate patient outcomes and enhanced decision-making in clinical settings.[3]

Machine learning models can be integrated into clinical decision support systems to provide real-time assistance to healthcare practitioners. Automated text summarization methods allow for the extraction of relevant information from extensive medical texts, thereby improving efficiency and minimizing the risk of data contamination.[4] The majority of experts agree that NLP is the most effective approach for enhancing medical text data through ML. NLP turns unstructured medical texts into structured data, making them easier to analyze and locate useful information. Current solutions include automatic text summarization systems, predictive algorithms such as SVMs and artificial neural networks, and enhancements in data quality through ML to identify and correct errors. Additionally, there are automated data curation tools and clinical decision support systems. These technologies improve efficiency, reduce data contamination, and provide real-time decision support to healthcare professionals by analyzing medical text data and delivering evidence-based recommendations.[5]

While ML is a potent instrument for enhancing medical text data, it possesses numerous limitations. A significant challenge pertains to data quality and bias: ML algorithms may exhibit bias if the training data contain inaccuracies or prejudices. Interpretability is a challenge, as sophisticated ML models are sometimes perceived as "black boxes" due to their decision-making processes. Another significant concern is data security and privacy, as medical text data frequently include sensitive patient information. Generalizability is also a concern, as ML models trained on a singular dataset may perform inadequately with heterogeneous populations. The development and training of ML models, which are resource-intensive endeavors, may pose challenges for small-sized healthcare organizations that often lack the necessary funding, technical expertise, and infrastructure to implement such advanced technologies effectively.[6]

People can engage in the processing and cleansing of medical text data, as well as identifying inconsistencies, duplications, and errors. Advanced NLP approaches can be employed to extract significant insights from unstructured data, including sentiment analysis and computerized summaries of clinical narratives. The model can produce concise summaries of medical literature, continuously adapt based on user interactions, employ ML algorithms to develop predictive models for patient outcomes and treatment responses, and integrate user feedback to enhance its techniques and strategies.

This study presents a framework for assessing medical data quality using hybrid anomaly detection. The proposed framework aims to: (i) provide systematic data quality assessment using anomaly detection, (ii) enable comparison of different algorithms for quality evaluation, and (iii) generate comprehensive quality reports that can guide data cleaning efforts. While improved data quality may ultimately support better research and clinical applications, evaluating such downstream impacts is beyond the scope of this technical validation study. The study contributes as follows:

(i) Support research quality: Provide researchers with systematic quality assessments that can inform decisions about dataset suitability for specific analyses. The direct impact on patient outcomes requires separate clinical validation studies, which were not undertaken here.
(ii) Improve prediction precision: Anomaly detection techniques improve the predictive effectiveness of ML models used in medical diagnosis and treatment, leading to more accurate diagnoses and personalized treatment plans for patients.
(iii) Optimize data processing: Automate the detection and correction of data anomalies to reduce manual intervention and increase efficiency.
(iv) Enhance patient outcomes: Deliver more accurate and

timely medical insights through improved data quality and anomaly detection, leading to better patient outcomes.

(v) Enable data quality monitoring: Offer automated quality metrics that organizations could incorporate into broader data governance frameworks. The framework itself does not constitute a complete governance solution but provides tools that may support governance activities, such as data validation processes and compliance tracking, which are essential for ensuring data integrity and reliability.

(vi) Enhance research and development: Provide a more accurate and reliable dataset for research, accelerating advancements in medical technologies and treatments.

It is essential to acknowledge that our analysis centered on numerical features derived from medical data, sourced from text-based clinical documentation. The conversion of unstructured text into structured numerical formats is an essential preprocessing step that facilitates the implementation of ML algorithms while preserving the clinical significance of the original textual data.

Figure 1 clearly depicts a graphical abstract of a comprehensive framework for improving the quality of medical text data via an innovative hybrid ML methodology. The process commences with the acquisition of unrefined, unstructured medical text data, encompassing sensitive clinical information such as Alzheimer's disease (AD) brain scans and patient histories. The datasets are further processed using the hybrid isolation forest–SVM (HIFSVM) model. The model utilizes the distinctive advantages of its components: the isolation forest delivers precise anomaly identification for extensive datasets, whereas the SVM ensures stability and minimal error rates in uniform data settings. The results indicate a substantial enhancement in data integrity, corroborated by 10 essential quality measures. The final results underscore the model's exceptional performance in attaining near-perfect scores in accuracy, completeness, and consistency, hence furnishing healthcare professionals with more reliable data for clinical decision-making.

2. Related works

Phan *et al.*[7] introduced an innovative framework for learning sparse Gaussian graphical models using multiple sliding windows of a dataset. This model incorporates a zero constraint and L2 regularization to ensure sparsity, and employs a proximal gradient algorithm to address the zero-constrained and the L2-regularized optimization problems. When combined with Kullback–Leibler divergence scoring for anomaly detection, the proposed method demonstrated superior performance compared to existing approaches.

Al-amri *et al.*[8] asserted that identifying unusual occurrences in Internet-of-things (IoT) data can mitigate functional hazards and avert system downtime. However, challenges persist due to the dynamic and streaming nature of IoT data, including the need for real-time processing and adaptability to evolving data distributions. They addressed this gap by offering an in-depth introduction to state-of-the-art techniques pertinent to IoT data processing, emphasizing key research issues, such as dynamic data, feature evolution, windowing techniques, ensemble methodologies, data complexity, noise interference, parameter optimization, data visualization, data heterogeneity, accuracy, large-scale datasets, and high-dimensional data analysis. They outlined several significant difficulties that necessitate extensive research and future exploration, underscoring the necessity for adaptability and robustness in many operational circumstances.

Algorithms are anticipated to derive data-driven parameters autonomously, function in real time, and reduce communication.[9] Nassif *et al.*[10] conducted a comprehensive literature review of anomaly detection using machine learning, analyzing 290 publications from 2000 to 2020. Their study identified 43 application domains, 29 distinct machine learning models, and 22 datasets, with unsupervised approaches emerging as the most widely used due to their effectiveness across diverse applications, including fraud detection, network security, and industrial fault diagnosis. However, the review also identified limitations in evaluation practices, noting that a majority of studies (64 out of 290) relied heavily on accuracy and the area under the receiver operating characteristic curve (AUC-ROC) as primary metrics. The authors emphasized the need for more rigorous evaluation frameworks and for incorporating recent advancements to address gaps in current methodologies.

Yang *et al.*[11] proposed an automated method for detecting inner ear abnormalities using deep reinforcement learning. The method utilizes two metrics, Dima gee and Uima gee, based on the variability of projected landmark configurations in normative data, capturing the agent's reluctance as it approaches target anatomical structures. The proposed method surpassed conventional 3D convolutional autoencoder techniques in identifying atypical anatomical structures in both synthetic and authentic clinical computed tomography (CT) data, owing to its superior ability to manage various types of clinical images.

Ajitha[12] introduced a novel approach to anomaly detection, SemiAI-anomalyDetect+, specifically designed for extensive patient datasets, aiming to identify irregular

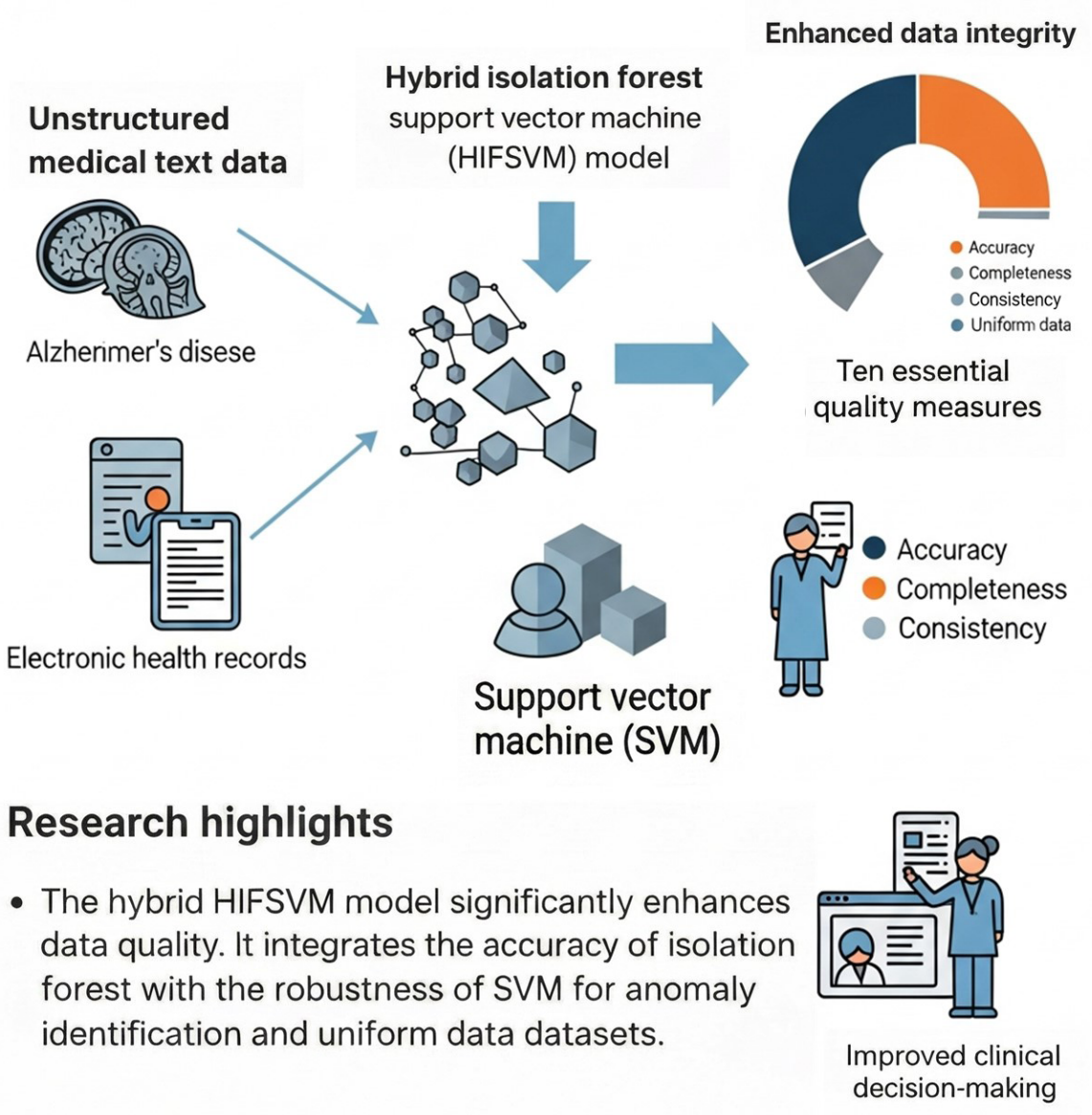


Figure 1. Processes of the proposed hybrid isolation forest–support vector machine (HIFSVM) model for the enhancement of medical text data quality. The process illustrates the transformation of unrefined, unstructured medical data, including Alzheimer's clinical information, via a dual-layered architecture that integrates an isolation forest for precise anomaly identification and a support vector machine (SVM) for data stability. The process yields high-quality, validated datasets designed for precision, comprehensiveness, and uniformity in clinical decision support.

patterns indicative of potential health concerns. This method combines unsupervised K-means clustering with semi-supervised learning techniques, demonstrating improved performance over traditional algorithms in terms of accuracy, completeness, F1-score, and AUC-ROC. Additionally, the method incorporates a feedback loop and an active learning mechanism to ensure continuous improvement in anomaly detection. However, its effectiveness depends heavily on advanced feature engineering techniques, hybrid methodologies, active learning strategies, and interpreted models, which may limit its accessibility for widespread clinical adoption.

Bao *et al.*[13] highlighted anomaly detection as a vital research area in ML and computer vision, especially in the context of medical imaging applications. Nonetheless, a conventional baseline for assessing diverse AD approaches in relation to medical imaging remains absent. To address this, they introduced an advantageous assessment standard comprising 6 reconfigured datasets spanning five medical categories and integrating 15 sophisticated

AD algorithms, enabling academics to efficiently compare and assess various AD approaches. Meanwhile, they also acknowledged several limitations and potential areas for optimization.

Baydargil *et al.*[14] indicated that deep neural networks trained on medical data encounter obstacles related to financial expenditures, specialized manpower, and extensive datasets, which can hinder their implementation in clinical settings and limit their accessibility for widespread use. They proposed a novel model for detecting atypical patterns in medical images, employing two encoders and a technique to transform images according to their information content. The model outperformed benchmark models in anomaly detection and classification, exhibiting enhanced performance when trained with preprocessed data. The model's efficacy on low-variance images, including positron emission tomography brain scans and chest X-rays, suggests its potential applicability in medical imaging. By integrating cascading and dilated convolutional neural networks, the model effectively captures both local and global properties, which are essential for accurately identifying highly heterogeneous diseases, such as AD.

Eltanbouly *et al.*[15] emphasized that cloud computing is particularly vulnerable to attacks due to its distributed data-processing architecture. To protect these systems, intrusion detection methodologies have evolved, increasingly incorporating ML techniques. Recent advancements in these techniques focus on evaluating accuracy and detection rates across various datasets. Their survey explored the characteristics, benefits, and limitations of ML methods for detecting a range of attack vectors in distributed systems, underscoring their importance in improving secure data sharing through the analysis and identification of network threats.

Almeida *et al.*[16] employed deep learning to identify various indicators of chronic obstructive pulmonary disease (COPD) as atypical patterns in chest CT scans. They employed inspiratory/expiratory CT and clinical data from the COPDGene and COSYCONET cohort studies. Their self-supervised deep learning model acquired semantic information from random voxel samples derived exclusively from undamaged lungs and identified anomalies as regional deviations. The model outperformed supervised models, including inspiratory CT alone, and improved predictive performance for clinical outcomes when integrated with linear mixed models.

Idé *et al.*[17] introduced an innovative framework for anomaly identification in complex systems, with particular focus on condition-based monitoring in industrial applications. Their method advocates multitask learning utilizing sparse Gaussian graphical models that represent various operational states, manage analogous assets, and elucidate the interrelationships among variables. They addressed numerical instability challenges, improving efficacy in anomaly detection on real-world datasets, particularly by enhancing the robustness of the model against fluctuations in data quality and ensuring more reliable identification of anomalies. Their method integrates variational Bayes with sparse structure learning, utilizing the Laplace prior and a zero-based method for selecting sparse mixture weights. This approach effectively learns multi-modal distributions in IoT applications, which are systems that connect and communicate data between devices over the internet.

Spies *et al.*[18] provided a thorough evaluation of methodologies for identifying anomalies in laboratory medicine, categorizing anomalies into specimen-, analyzer-, patient-, and population-related anomalies. The aim of analyzer-based anomaly detection is to identify technical errors caused by the measuring device. Currently, the anomaly detection methods include Levey–Jennings plots and patient-based real-time quality control. They anticipated that, in the future, anomaly detection through analyzers may incorporate ML techniques to enhance performance and automate workflows designed to identify technological anomalies before they lead to erroneous diagnoses or treatments.

Ibrahim *et al.*[19] evaluated the computer-assisted framework, COV-CAF, designed to improve the identification and categorization of COVID-19 severity from chest CT scans. The framework transitions from conventional methods to diverse deep learning techniques to improve diagnostic precision in assisting healthcare practitioners with case management. MosMedData achieved an accuracy of 97.76% with an average sensitivity of 96.73%, whereas the SARS-COV-2 CT-Scan benchmark dataset yielded accuracy and sensitivity rates of 97.59% and 98.41%, respectively. The framework's efficacy is contingent upon the quality and diversity of the datasets utilized for training, rendering deployment in a real clinical environment highly challenging.

Minic *et al.*[20] investigated the efficacy of recurrent neural networks in identifying irregularities in electrocardiograms (ECGs). They proposed a novel framework that integrates a particle swarm optimization algorithm to improve performance in ECG analysis and facilitate early identification of cardiovascular diseases. Their study demonstrated the efficacy of recurrent neural networks for interpreting temporal heart electrical signals and the role of particle swarm optimization in optimizing recurrent neural network parameters. Their results

suggest a significant enhancement in performance, the identification of critical traits, and practical applicability in real-world scenarios to facilitate timely patient care. Their study has limitations regarding generalizability, implementation complexity, and data quality reliance.

Liu *et al.*[21] established an efficient framework, AIEchoDx, to compare four common cardiovascular diseases to normal patients using deep learning techniques. The framework demonstrated exceptional diagnostic performance, with AUC values ranging from 98.70% to 99.57%. This methodology effectively identifies critical lesion locations in echocardiographic images and offers insights into the decision-making process. Additionally, it allows for extensibility to various echocardiography devices through a proposed framework that enhances access to point-of-care diagnostics, leading to improved patient outcomes and more widespread use of echocardiographic assessments in diverse clinical settings. Their study acknowledged several limitations related to generalizability, the size of the training dataset, and the complexity of clinical scenarios; moreover, the real-world effectiveness of the findings has yet to be assessed.

Mohammed *et al.*[22] improved early disease diagnosis with a hybrid methodology that integrates gray wolf optimization (GWO) with SVM for gene selection. The GWO-SVM technique achieved a 96.46% accuracy rate in disease diagnosis, demonstrating its efficacy in identifying genes associated with diverse diseases. Their work advances computational biology by introducing a hybrid model that employs variance analysis to identify significant genes and by demonstrating the application of metaheuristic algorithms for the analysis of biological data. Nonetheless, its application may be limited by dataset dependency and the complexity of biological systems, which can restrict generalizability across different biological contexts.

Rahman *et al.*[23] enhanced anomaly identification in human activities by employing ML and deep learning classifiers. Their gated recurrent unit classifier outperformed other models, such as naive Bayes, in accurately distinguishing between normal and anomalous behaviors. They generated 50,000 anomalous activities while conducting tests on training–test splits and cross-validation methodologies. Their findings indicate that ML algorithms can proficiently detect anomalies in two-resident households, thereby improving safety and well-being. Their limitations encompass the difficulties of acquiring datasets that authentically depict multi-resident environments and the associated privacy concerns, which can hinder the development and testing of ML models aimed at improving safety and well-being in such settings.

Bataineh's[24] proposed a heterogeneous ensemble learning framework that employs a binary genetic algorithm (GA) to enhance weak learners, such as random forest, gradient boosting, SVM, and AdaBoost, for superior predictive performance. The ensemble utilizes a soft-voting mechanism and was assessed via 10-fold cross-validation employing heating, ventilation, and air conditioning (HVAC) sensor data from 10 medical institutions in Jordan. The GA-optimized ensemble achieved a macro-averaged F1-score of 77.35% and an overall accuracy of 97.50%, significantly surpassing the scores of a naive ensemble. The improved efficiency and collaboration of the optimized subset of models resolve class imbalance issues in the dataset.

Fuzail *et al.*[25] presented a hybrid framework that combines federated learning with ensemble ML models and explainable AI techniques to predict sepsis mortality. This framework employs random forest, light gradient boosting machine, extreme gradient boosting, k-nearest neighbors, and logistic regression without centralizing patient data. It improves interpretability through Shapley additive explanations, local interpretable model-agnostic explanations, and partial dependence plots, thereby enhancing transparency in clinical decision-making. The effectiveness of the framework was assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC in both centralized and federated environments. Their results indicate that ensemble models, especially random forests and gradient boosting, enhance predictive accuracy and robustness in a federated framework. Their proposed hybrid framework allows for precise, privacy-conscious, and understandable predictions of sepsis mortality, thereby improving real-time clinical decision-making in Intensive Care Unit settings.

Núñez-Valdez[26] contributed to the growing body of research at the intersection of machine learning and healthcare by summarizing diverse applications, methodologies, and clinical contexts, emphasizing technological advancements and practical factors such as efficiency, interpretability, robustness, and security. The summary also highlights the transformative capacity of ML in healthcare and the challenges that must be overcome for reliable and responsible real-world implementation. We anticipate that this summary will stimulate additional research, foster interdisciplinary collaboration, and promote innovation, thereby advancing intelligent healthcare systems.

Alhassan and Altmami[27] introduced a model that combines a hybrid gradient boosting model with a hybrid attention module, effectively reducing complexity while enhancing performance and accuracy. The model's feature extraction process uses two pre-trained models that work

together to improve feature extraction, incorporating shape and texture features. The main goal of this model is to identify multiclass non-melanoma skin cancer. It achieved strong evaluation metrics, including an accuracy of 96.8%, sensitivity of 97.56%, specificity of 96.03%, precision of 96.71%, and an F1-score of 97.13% on a histopathological non-melanoma skin cancer segmentation dataset.

Paul *et al.*[28] proposed an improved model utilizing the synthetic minority over-sampling technique and edited nearest-neighbor balancing, along with principal component analysis (PCA) and gradient boosting, to predict chronic diseases from clinical profiles. The study was based on a Kaggle dataset of over 1,500 anonymized patient records with 15 demographic, lifestyle, and clinical features. Standardized encoding and class adjustments were conducted prior to PCA to optimize computation and minimize overfitting. Various models, including decision trees, random forests, light gradient boosting machine, and extreme gradient boosting, were compared, with gradient boosting yielding the best performance (accuracy of 99.33%, precision of 99%, recall of 99%, and F1-score of 99%). Cross-validation indicated an accuracy of 98.90%, and the model achieved an AUC-ROC value exceeding 0.99, signifying its robustness and precision for early chronic condition detection. This ML system holds significant potential for improving patient-oriented outcomes.

El-Aziz and Rayan[29] introduced a deep learning framework (ASTab-Stroke) that integrates adaptive stratified sampling with TabNet's sequential attention mechanism to improve stroke prediction accuracy. Adaptive stratified sampling adjusts patient strata according to their error contributions, ensuring representation of minority and high-risk groups without the use of synthetic data. TabNet enables incremental feature attribution, permitting clinicians to discern the impact of variables such as age, hypertension, and lifestyle on stroke risk. Developed in Python 3.10 and assessed using a diverse Stroke Prediction Dataset, ASTab-Stroke achieved 98% accuracy and significant AUC, F1-score, recall, and precision, exceeding current benchmarks. The framework demonstrated robust generalization, emphasizing age and comorbidity characteristics as pivotal in ablation studies. It offers clinically interpretable, scalable, and ethically sound stroke risk prediction, with potential applications in other domains of precision medicine aimed at equitable patient care, such as predicting risks for cardiovascular diseases or supporting diabetes management.

Tascioglu *et al.*[30] discussed how NLP methodologies transform unstructured medical text into structured data, facilitating tasks such as sentiment analysis and text summarization. Their work emphasized the effectiveness of both rule-based approaches and large language models in extracting vital clinical information (e.g., comorbidities) from historical free-text records, thereby making significant contributions to mortality risk evaluation.

3. Conceptual framework: Machine learning as a tool for data quality assessment

3.1. Difference between data quality assessment and model performance evaluation

It is crucial to differentiate between two separate analytical tasks: (i) assessing the efficacy of an ML model (i.e., its predictive or classificatory accuracy) and (ii) employing ML models to evaluate data quality (i.e., the appropriateness of the data for analysis). The present study examines the application of ML models to evaluate data quality. The isolation forest, SVM, and HIFSVM models served as diagnostic instruments that analyzed the dataset and produced quality metrics across 10 dimensions. When a feature possesses an accuracy of 1.00, it does not imply that the model is flawless; rather, it indicates that the data values fall within the statistically anticipated ranges. Likewise, error rate denotes the ratio of outliers identified in the data, rather than the classification error of the model. This distinction forms the foundation of the entire analytical framework presented here.

3.2. Unsupervised anomaly detection for quality analysis

Medical data often contain various anomalies that might significantly impact the accuracy of ML algorithms. Presented below are several common categories of anomalies found in medical data text: (i) Outliers: Data points that significantly deviate from the rest of the dataset, such as abnormally high or low blood pressure measurements. (ii) Missing values: Instances where data are absent or incomplete, sometimes resulting from errors in data entry or collection. (iii) Redundant records: Repeated entries of the same patient or test result, which can skew analysis and lead to inaccurate conclusions about patient health trends or treatment effectiveness. (iv) Inconsistent data: Discrepancies in data recording, encompassing variances in units of measurement or formats. (v) Noise: Erroneous fluctuations or oscillations in data that possess no evident significance.[31]

3.3. Types of anomalies in medical data

Medical datasets frequently display a wide array of anomalies that can substantially skew analytical results and compromise clinical decision-making. These irregularities typically manifest as point anomalies, where a single

value strongly deviates from expected behavior, such as an abnormally high biomarker reading, and contextual anomalies, in which otherwise normal values become abnormal within a specific clinical context, such as a heart rate that is appropriate during exercise but problematic at rest.

Collective anomalies are also critical. They happen when groups of observations all deviate from expected patterns. This is often a sign of systemic problems, such as calibration errors or coordinated data shifts across several variables. Beyond these statistical outliers, medical data frequently suffer from missing values, redundant or duplicated records, inconsistent entries resulting from differing units or formats, and noise, which appears as random fluctuations not tied to meaningful physiological variation. These anomaly types directly affect key data quality aspects, such as validity, uniqueness, precision, and error rate, making their identification essential for ensuring reliable downstream analytics. In our study, such anomalies are not only a fundamental concern but also the basis for evaluating the effectiveness of models, such as isolation forest, SVM, and HIFSVM, which demonstrated superior capability for detecting subtle and clinically relevant deviations within numerically encoded medical text.

3.4. Algorithms as quality assessment instruments

Anomaly detection constitutes a crucial component of ML. It signifies the identification of anomalous items, occurrences, or observations that markedly deviate from the predominant dataset. This process is crucial for various applications, such as fraud detection, network security, and healthcare diagnostics. A range of techniques has been developed for anomaly detection, each with unique benefits.

3.4.1. Isolation forest

The isolation forest algorithm is a widely recognized unsupervised ML technique for anomaly identification (Figure 2). It operates by constructing isolation trees: it isolates anomalies by randomly selecting a feature and subsequently choosing a split value at random between the feature's minimum and maximum values. The iterative repetition of the operation generates a structure referred to as an isolation tree. Path length anomalies are typically sparse and distinctly differentiated from the remainder of the data. Consequently, isolating these anomalies would generally necessitate considerably fewer splits. The path length in the isolation tree can be utilized to calculate an anomaly score. Isolation forest is efficient with linear time complexity and has been applied across multiple domains, including fraud detection—where it identifies anomalous transactions in healthcare systems—and network security, to detect unusual trends that may signal potential cyberattacks—similar to how fraud detection in healthcare identifies anomalous transactions that could indicate fraudulent activity. In healthcare, the goal is to identify uncommon health issues among extensive databases of test results.[32]

3.4.2. One-class support vector machine

The one-class SVM, designed for unsupervised anomaly detection, is based on the SVM methodology typically used in classification tasks. While a standard SVM aims to identify the hyperplane that maximizes the margin between different data points, the one-class SVM focuses on identifying the hyperplane that most effectively isolates the data points from the origin. Furthermore, SVMs can handle non-linearity by employing kernel functions,[33] which allow them to create complex decision boundaries in the feature space. Figure 3 presents the architecture of the SVM.

3.4.3. Hybrid isolation forest–support vector machine model

This subsection describes the conceptual pipeline of the HIFSVM framework. Readers can refer to Section 3.6 for the complete mathematical formalization, including all equations, algorithmic specifications, and complexity analysis. Identifying anomalies in records is an essential component of statistical analysis and device studying. It helps identify anomalous patterns that deviate from expected behavior, potentially indicating widespread occurrences, fraud, or exceptional concerns. This methodology examines various strategies for anomaly identification, utilizing a hybrid technique that combines multiple algorithms. This analytical approach includes comprehensive data preparation, transformation, and the implementation of various anomaly detection models, as illustrated in the subsequent phases:

(i) Phase 1: Load data

The framework begins with the acquisition of datasets related to AD and healthy aging via CSV file loading. It employs advanced memory optimization techniques, including chunked reading functionalities, to efficiently manage large datasets. Thorough exploratory activities are conducted to understand dataset dimensions, memory usage patterns, and data type distributions. To enhance text analysis in healthcare contexts, the framework integrates specialized medical terminology lexicons, improving the accuracy of data interpretation and facilitating deeper insights into patient records and clinical notes.

(ii) Phase 2: Clean data

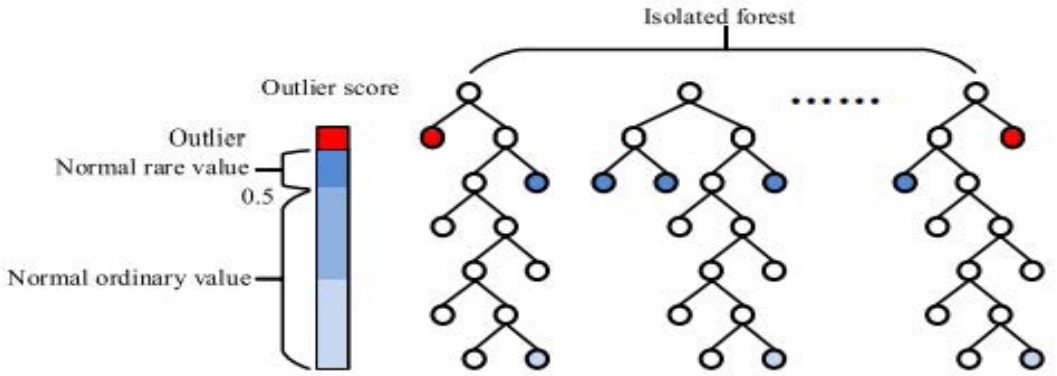


Figure 2. The outlier-detection design uses an isolated forest to differentiate between normal values and outliers by generating anomaly scores. This approach is effective for handling large or complex datasets, making it beneficial for applications such as fraud detection and noise filtration.

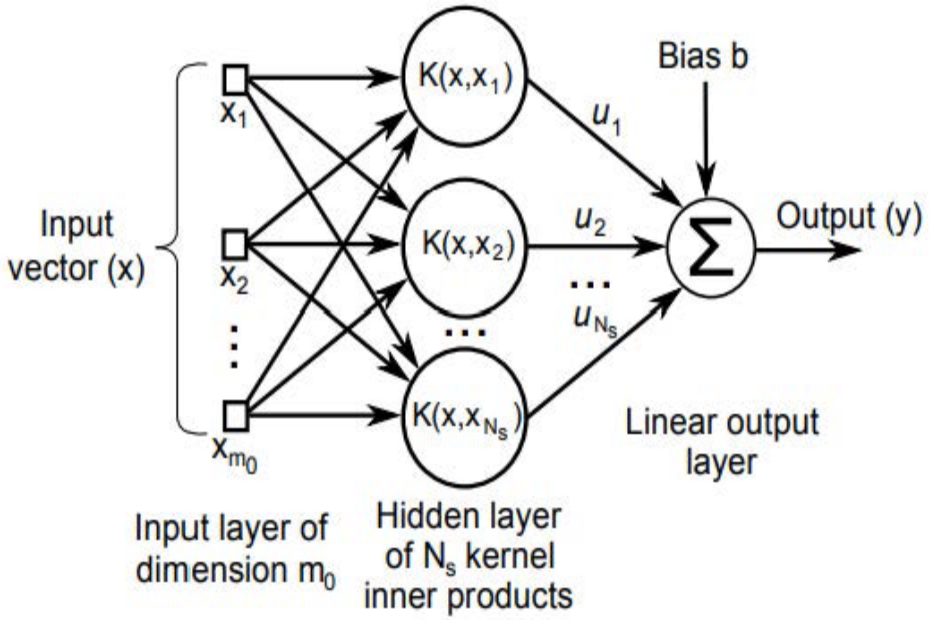


Figure 3. The architecture of kernel-based transformation in the support vector machine explains how input vectors are processed using kernel functions within the hidden layer. It also clarifies that the output, denoted as y, results from a weighted combination of various modified outputs. Reprinted from Ruiz-Gonzalez *et al* [34]

This phase ensures data integrity through systematic cleansing procedures. The framework identifies and removes duplicate entries to prevent analytical redundancy and maintain dataset reliability. Text normalization addresses irregular whitespace and converts empty strings into standardized representations of missing values. Memory optimization strategies are applied during cleaning, supported by tracking systems that evaluate cleaning effectiveness and memory conservation, helping ensure that data processing is efficient and does not lead to excessive resource consumption.

(iii) Phase 3: Process features

This phase involves strategic feature selection, isolating numerical columns for further analysis. Advanced imputation techniques dynamically choose between mean and median values based on the data distribution. Outlier management is conducted using winsorization

methods grounded in interquartile range (IQR) analysis. Transformation analysis tools recommend appropriate data transformations, such as logarithmic, Box–Cox, or Yeo–Johnson, based on distribution characteristics.

(iv) Phase 4: Compute quality metrics

A robust quality assessment framework evaluates datasets across ten key dimensions, including accuracy, completeness, validity, and uniqueness. Statistical analysis incorporates outlier detection and distribution evaluation. Textual data are assessed for medical terminology coverage and overall integrity. Pattern recognition capabilities identify missing correlations and systemic data quality issues, which can lead to inaccurate conclusions in medical research and hinder effective decision-making.

(v) Phase 5: Sophisticated anomaly detection

This phase applies a multi-algorithm strategy integrating four distinct detection methods. Isolation forest uses random partitioning for outlier detection; local outlier factor applies density-based techniques; SVM utilizes boundary-based identification; and an ensemble method consolidates results through majority voting. The system includes performance monitoring to track training durations and compute algorithm efficiency.

(vi) Phase 6: Dimensionality reduction and visualization

Dimensionality reduction is achieved through PCA, which enhances data presentation. The framework generates comprehensive visualizations, including scatter plots to distinguish normal and anomalous observations. Heatmaps display quality metrics across dimensions, correlation matrices reveal feature relationships, and distribution histograms and radar maps offer alternative perspectives for data interpretation, allowing analysts to visualize the frequency distribution of data points and compare multiple variables simultaneously.

(vii) Phase 7: Analysis of training performance

This phase focuses on computational efficiency through detailed algorithm performance analysis. Training durations are monitored and compared across methods, with metrics and timing statistics computed. Resource optimization insights identify the most and least efficient algorithms. Execution breakdowns provide granular timing data for each processing stage, supporting informed algorithm selection based on data attributes and system constraints.

(viii) Phase 8: Visualization and reporting

Automated reporting tools generate outputs in JSON and HTML formats, presenting comprehensive analytical findings. Executive summaries highlight aggregated data quality scores and critical issues. A recommendation engine offers ways to improve data based on the analysis, while advanced scoring systems set strict standards for checking quality. This ensures stakeholders receive actionable insights in accessible formats.

(ix) Phase 9: Execution synopsis and performance evaluation

The final phase delivers detailed execution reports evaluating the overall pipeline performance. Quality metrics are synthesized into actionable insights, and training time analysis summarizes ML algorithm efficiency. Prioritization algorithms identify the most urgent data quality issues, enabling effective resource allocation. This structured framework merges traditional statistical methods with modern ML to provide a holistic view of dataset integrity, reliability, and analytical readiness.

The qualitative description above provides an intuitive understanding of the HIFSVM pipeline. The formal mathematical specification in Section 3.6 complements this by providing the precise equations, algorithms, and optimization procedures necessary for rigorous implementation and reproducibility.

Figure 4 illustrates that the HIFSVM method offers a systematic approach to data preprocessing, feature generation, and outlier detection. It employs advanced techniques, such as isolation forest and SVM, to identify outliers and uses visualizations to improve the comprehensibility of the data. Additionally, Figure 5 presents the HIFSVM flowchart, outlining a structured approach to data preparation, feature creation, and outlier detection.

3.5. Metrics for assessing data quality

The assessment of data quality is a crucial process for ensuring that data are accurate, complete, and reliable. High-quality data are essential for informing decision-making, improving processes, and achieving organizational success. Various techniques and methodologies are available to assess and improve data quality, each designed to address specific characteristics such as accuracy, consistency, completeness, and timeliness. The process involves using fundamental tools and methodological approaches to evaluate data quality, including data profiling, data cleansing, and data validation to identify and rectify data problems. Quality assessments and metrics are tools that generate reports and metrics to evaluate data quality, track progress, and identify areas for improvement. Companies collect user surveys and feedback to understand data quality issues from their viewpoint and implement the necessary adjustments. Self-assessment and auditing are internal processes that methodically analyze data quality

and ensure compliance with data standards. Designation and accreditation apply labels and certificates to data to ensure adherence to specified quality standards.

Algorithm 1. DataQualityAnalyzer: Pseudocode for the HIFSVM quality assessment pipeline

```
Pseudocode
FUNCTION DataQualityAnalyzer():
  // STEP 1 — Load & Inspect
  LOAD dataset FROM file_path
  OPTIMIZE memory WITH chunked_reading
  GET dataset_info (shape, types, memory)
  missing_matrix← IS_NULL(dataset)
  // STEP 2 — Clean Data
  cleaned_data← DROP_DUPLICATES(dataset)
  NORMALIZE text_columns (strip, replace empty)
  memory_saved← original_memory − cleaned_memory
  // STEP 3 — Process Features
  numerical_cols← SELECT numerical columns
  FOR each column:
    IF skewness > 2 THEN impute median ELSE impute mean
    CLIP values TO [Q1−3·IQR, Q3+3·IQR]
  // STEP 4 — Compute Quality Metrics
  FOR each feature:
    accuracy ← fraction within μ ± 3σ
    completeness← non-null / total
    validity← satisfying constraints / total
    uniqueness← unique_count / total
    error_rate← outliers (1.5 IQR rule) / total
  // STEP 5 — Anomaly Detection
  IF_scores← IsolationForest(X)
  LOF_scores← LocalOutlierFactor(X)
  OCSVM_scores← OneClassSVM(X)
  ensemble ← COMBINE(IF_scores, LOF_scores, OCSVM_scores)
  // STEP 6 — Visuals
  X_pca ← PCA(X, 2)
  quality_heatmap← HEATMAP(metrics)
  correlation_heatmap← HEATMAP(CORRELATION(X))
  distributions← HISTOGRAMS(X)
  // STEP 7 — Reports
  json_report ← JSON(analysis_results)
  html_report ← HTML(summary, metrics)
  overall_quality_score ← f(completeness, validity, accuracy, error_rate)
END FUNCTION
FUNCTION main_execution():
  analyzer ← DataQualityAnalyzer()
  dataset← analyzer.LOAD(...)
  cleaned_data← analyzer.CLEAN(...)
  missing_analysis← analyzer.MISSING(...)
  text_analysis← analyzer.TEXT(...)
  X_processed← analyzer.FEATURES(...)
  metrics← analyzer.METRICS(...)
  anomalies← analyzer.ANOMALIES(...)
  visuals← analyzer.VISUALS(...)
  reports← analyzer.REPORTS(...)
  RETURN all_results
END FUNCTION
```

Table 1 provides standardized definitions for all 10 metrics employed in this study, distinguishing between their conceptual foundations and operational implementations (also see Table S1):

3.6. Formal specification of the proposed hybrid isolation forest–support vector machine model

Figure 4 illustrates the conceptual architecture of HIFSVM, but for reproducibility, a precise mathematical specification is required. This section provides the formal mathematical foundation of the HIFSVM model, including its constituent components, fusion mechanism, and optimization procedure.

3.6.1. Algorithmic specification

```
Input:
  X_train ∈ ℝ^(n×m)      // Training data
  X_test ∈ ℝ^(k×m)       // Test data
  n_estimators = 100     // Number of isolation trees
  contamination = 0.1    // Expected anomaly rate
  kernel = 'rbf'         // SVM kernel type
  γ = fscale'            // RBF kernel parameter
Output:
  a_train ∈ ℝ^n          // Training anomaly scores
  a_test ∈ ℝ^k           // Test anomaly scores
  quality_metrics ∈ R^10 // Data quality metrics
// Stage 1: Isolation Forest Training
Initialize isolation_forest = IsolationForest(
  n_estimators=n_estimators,
  contamination=contamination,
  random_state=42)
isolation_forest.fit(X_train)
a_train_IF = isolation_forest.score_samples(X_train)
a_train_IF = normalize_to_unit_interval(a_train_IF)
// Stage 2: One-Class SVM Training
Initialize svm = OneClassSVM(
  kernel=kernel,
  gamma=γ,
  nu=contamination)
svm.fit(X_train)
distances = svm.decision_function(X_train)
a_train_SVM = sigmoid_transform(distances)
// Stage 3: Weight Optimization
w_star = optimize_fusion_weight(
  a_train_IF,
  a_train_SVM,
  contamination)
// Stage 4: Hybrid Fusion (Training)
a_train = w_star * a_train_IF + (1 - w_star) * a_train_SVM
// Stage 5: Inference on Test Data
a_test_IF = isolation_forest.score_samples(X_test)
a_test_IF = normalize_to_unit_interval(a_test_IF)
a_test_SVM = svm.decision_function(X_test)
a_test_SVM = sigmoid_transform(a_test_SVM)
a_test = w_star * a_test_IF + (1 - w_star) * a_test_SVM
// Stage 6: Quality Metric Computation
quality_metrics = compute_quality_metrics(X_train, a_train)
return a_train, a_test, quality_metrics
```

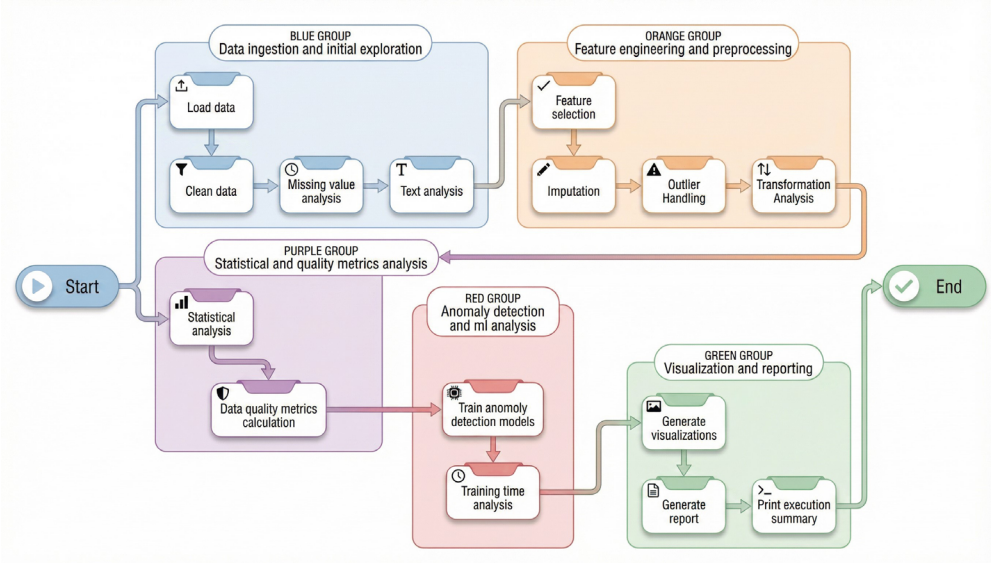


Figure 4. The hybrid isolation forest–support vector machine (HIFSVM) methodology provides a structured approach to data preprocessing, feature engineering, and anomaly detection. It utilizes advanced techniques, including isolation forest and support vector machine, to identify outliers and incorporates visualizations to improve interpretability.
Abbreviation: ML: Machine learning.

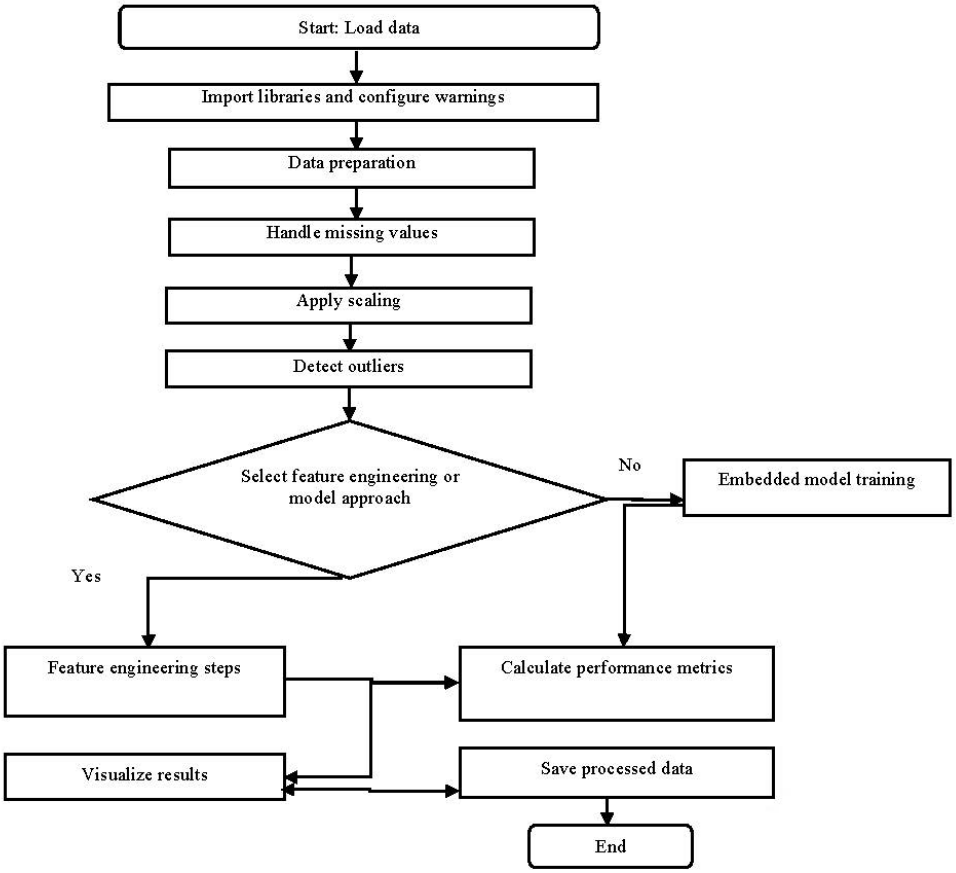


Figure 5. The hybrid isolation forest–support vector machine (HIFSVM) flowchart delineates a systematic approach to data preparation, feature engineering, and anomaly identification. It employs sophisticated methodologies, including isolation forest and support vector machines, to detect anomalies and integrates visuals to enhance interpretability.

Table 1. Definition of the 10 metrics used in this study

Metric	Conceptual definition	Operational definition	Scale	Interpretation
Accuracy	Conformity of data to real-world values	Proportion of values within mean ± 3σ	0–1	Higher = more statistically plausible values
Completeness	Extent to which data are non-missing	Non-null values/Total values	0–1	Higher = fewer missing values
Consistency	Absence of contradictory data	Binary: median between minimum and maximum	0 or 1	1 = internally consistent
Timeliness	Currency of data	1.0 for static datasets with appropriate time coverage	0–1	Higher = more current
Validity	Conformity to domain constraints	Values meeting domain rules/Total	0–1	Higher = more domain-compliant
Uniqueness	Absence of duplication	Unique values/Total values	0–1	Higher = more distinct entries
Precision	Granularity of measurement	Average decimal places	≥0	Higher = more detailed measurement
Change rate	Volatility of values	Coefficient of variation (σ/μ)	≥0	Lower = more stable
Error rate	Proportion of statistical outliers	Values outside 1.5×IQR/Total	0–1	Lower = fewer outliers
Data density	Population of fields	Populated fields/Total fields	≥0	Higher = more complete coverage

Abbreviation: IQR: Interquartile range.

3.6.2. Optimization of the hybrid fusion weight

The fusion weight w is optimized using grid search with validation:

$$w^* = \arg\max_{w \in W} Q(w) \quad (1)$$

where w = 0, 0.1, 0.2, ..., and Q(w) is the quality score, where:

$$Q(w) = \frac{1}{10}\sum_{j=1}^{10} m_j(w) \quad (2)$$

with $m_j(w)$ being the j quality metric computed using the hybrid scores with weight w.

3.6.3. Complexity analysis

There are two complexities, time complexity and space complexity:

(i) Time complexity:

- Isolation forest training: $O(t \cdot n \cdot \log n)$, where t is the number of trees.
- One-class SVM training: $O(n^2 \cdot m)$ to $O(n^3 \cdot m)$ depending on the implementation.

- HIFSVM hybrid fusion: $O(n \cdot m)$.
- Overall: $O(n^2 \cdot m)$ (dominated by SVM component).

(ii) Space complexity:

- Isolation forest: $O(t \cdot n \cdot m)$.
- One-class SVM: $O(n^2)$ for the kernel matrix.
- Overall: $O(n^2 + t \cdot n \cdot m)$.

3.6.4. Essential elements for effective hybrid fusion

Developing an effective hybrid model necessitates the assessment of multiple critical factors instead of merely averaging outcomes from two algorithms. Essential components comprise: (i) algorithmic complementarity, wherein isolation forest and one-class SVM proficiently detect unique anomalies; (ii) score standardization, which normalizes algorithmic scores to mitigate scale dominance; (iii) optimization of fusion, wherein weights are refined via grid search to achieve balanced performance; (iv) adherence to assessment objectives, emphasizing practical quality evaluation; (v) computational compatibility, ensuring seamless integration with scikit-learn; (vi) maintenance of interpretability, facilitating the identification of anomaly sources; (vii) inter-dataset robustness, which is confirmed across multiple medical datasets; (viii) statistical significance, which is assessed via paired *t*-tests to validate improvements; and (ix) reproducibility, ensuring that other researchers can replicate results, as well as a theoretical framework grounded in complementary detection perspectives. These factors allow HIFSVM to achieve outstanding results.

3.7. Justification for algorithm selection

Isolation forest and SVM were chosen for our hybrid framework for medical data quality assessment due to several reasons: First, they offer superior interpretability for identifying data quality issues compared to deep learning methods. Isolation forest's explicit path lengths indicate anomalies linked to specific features, while SVM delineates clear boundaries with support vectors. Second, their scalability is advantageous, with isolation forest operating in $O(n\log n)$ time on standard hardware, making it suitable for large-scale datasets, unlike deep learning, which often requires vast computational resources. Third, both methods perform effectively across various dataset sizes without the need for massive training sets, thereby ensuring robustness. Furthermore, their strong theoretical foundations yield predictable behavior for quality assessment, contrasting with the less predictable nature of deep learning. Lastly, both methods support reproducibility through stable implementations in scikit-learn, establishing essential baselines for future research. Overall, these factors affirm isolation forest and SVM as optimal choices for our framework, enhancing the reliability of medical data quality assessments.

3.8. Limitations of deep learning

Considering the ubiquity of deep learning in modern ML research, it is pertinent to inquire why we did not utilize these techniques. Multiple factors render deep learning less appropriate for our particular task:

(i) Interpretability: Deep learning models offer a restricted understanding of the rationale behind a data point being identified as anomalous. For quality assessment, it is vital to understand the underlying causes, missing data, boundary violations, and statistical outliers for effective remediation. Isolation forest and SVM offer discernible explanations, making them preferable for tasks where understanding the rationale behind anomaly detection is crucial.

(ii) Reproducibility: The training of deep learning models incorporates stochastic factors (random initialization, dropout, and GPU nondeterminism) that hinder precise replication. Our dedication to reproducibility (Section 7) directed us toward deterministic methodologies.

(iii) Computational prerequisites: Deep learning generally necessitates GPUs and comprehensive optimization. Our framework operates on conventional hardware in roughly 2 h, ensuring accessibility.

(iv) Sample efficiency: Deep learning techniques frequently necessitate extensive datasets for effective generalization, which can be a challenge for practitioners with limited data availability. Our framework operates effectively across datasets of diverse sizes.

(v) Overfitting risk: Deep learning may become excessively tailored to dataset-specific patterns, potentially misclassifying legitimate variations as anomalies, leading to a significant increase in false positives and undermining the model's reliability in real-world applications. Conventional methods exhibit superior control of false-positive rates.

(vi) Theoretical foundation: The efficacy of deep learning is predominantly empirical; its behavior in edge cases remains unpredictable. The evaluation of quality is enhanced by clearly comprehended theoretical attributes.

(vii) Metric alignment: Deep learning focuses on optimization for prediction or reconstruction rather than directly targeting quality metrics. Our methodologies are specifically crafted to produce quality metrics that accurately reflect the performance

and effectiveness of our models in various applications.

These factors do not suggest that deep learning is "bad," it performs exceptionally well in numerous tasks, but rather that for the specific needs of interpretable, reproducible, and accessible quality assessment, traditional methods are more suitable. Deep learning may enhance our methodology in forthcoming projects targeting diverse objectives, such as improving accuracy, reducing bias, and optimizing performance across various applications.

4. Methods

This study used a server-based approach to implement ML models as diagnostic tools for evaluating and enhancing medical data quality. In contrast to conventional supervised learning, which primarily assesses model performance on reserved data, our methodology employs unsupervised anomaly-detection algorithms to systematically analyze dataset attributes across 10 quality dimensions. Table S2 summarizes the mathematical notation used throughout this paper. The models were "trained" on the data to comprehend the underlying distribution and detect deviations that signify quality concerns, rather than to make predictions. The data for analysis comprised medical texts and specializations relevant to human cranial disorders, including AD, Parkinson's disease, stroke, brain tumors, traumatic brain injury, and meningitis.

4.1. Dataset preparation

This study examined several key datasets relevant to public health research. The Alzheimer's Disease and Healthy Aging Data (ADHAD)[35] from the Centers for Disease Control and Prevention (CDC) contains extensive information on attitudes toward dementia and caregiving across over 100,000 records, allowing for longitudinal analysis of health disparities. The Breast Cancer Global Dataset (BCGD)[36] monitors breast cancer statistics worldwide from 1990 to 2022, categorizing data by age, geography, and molecular subtypes to evaluate the efficacy and equity of cancer care. Lastly, the 500 Cities: The Local Data for Better Health (500LDFB)[37] dataset provides small-area estimates of chronic disease indicators for the 500 largest United States cities, which aid local governments in addressing geographic health inequities. Together, these datasets offer critical insights for evaluating health interventions and understanding public health trends. Table 2 provides additional details for all datasets through comparison.

All codes and datasets used in this study are permanently archived at Kaggle Notebooks (Executable Environments):

- Isolation forest: https://www.kaggle.com/code/raedaalthabeti/isolationforest-28052025
- One-class SVM: https://www.kaggle.com/code/raedaalthabeti/one-class-svm
- HIFSVM: https://www.kaggle.com/code/raedaalthabeti/quality-metricsdata-hifsvm

Dataset access:

- ADHAD: https://www.kaggle.com/datasets/utkarshx27/alzheimers-disease-and-healthy-aging-data
- BCGD: https://gco.iarc.fr/tomorrow/en/datasets
- 500LDFB: https://www.kaggle.com/datasets/fandang/500-cities-local-data-for-better-health

4.2. Isolation forest data establishment

The isolation forest model automates multi-dimensional data quality assessments by calculating 10 metrics for each numerical column in a dataset. It provides a detailed report on the dataset's reliability and suitability for analysis, including accuracy, completeness, consistency, validity, uniqueness, and precision evaluations. The accuracy metric measures the fraction of values within three standard deviations of the mean; completeness measures the ratio of the number of non-missing values to total number of values; consistency ensures that the median values are between the minimum and maximum values; validity is domain-specific, ensuring compliance with expected value ranges; uniqueness is assessed through the ratio of distinct values; and precision indicates the average number of decimal places. The code substitutes the coefficient of variation for change rates when historical data are unavailable, and the Tukey method identifies outliers. Lastly, the Kullback–Leibler divergence assesses data drift, comparing column distributions to a normal distribution to highlight potential quality issues. Together, these metrics form a robust framework for understanding dataset constraints and reliability before advanced analysis or modeling.

4.3. Support vector machine data establishment

The SVM model provides a comprehensive framework for assessing the quality of medical datasets across 10 dimensions. It calculates completeness metrics to evaluate data coverage, employs validity checks using domain-specific rules for various fields, and ensures consistency by verifying data types and logical relationships. Uniqueness metrics identify duplicates, while precision analysis measures the level of detail in numerical data. The framework quantifies error rates through outlier detection, assesses data volatility with the coefficient of variation, and measures data drift using the Kullback-Leibler divergence. Accuracy is estimated by assessing the proportion of values within three standard deviations of the mean. A

Table 2. Comparison of features among ADHAD, BCGD, and 500LDFB datasets

Feature	ADHAD	BCGD	500LDFB
Publisher	Centers for Disease Control and Prevention (CDC)	Global collaboration (WHO, IARC, national registries)	CDC, RWJF, CDC Foundation
Time span	2015–2020	1990–2022	Latest release: 2019 (updated periodically)
Geographic coverage	All 50 US states	Over 100 countries (UN subregions)	500 largest US cities (~28,000 census tracts)
Data volume	>100,000 records; 114 MB Excel file	Large-scale global registry data	Multiple formats; large tract-level estimates
Data format	Excel	Harmonized registry formats (ICD-10, AJCC)	CSV, JSON, XML
Indicators tracked	>60 indicators: cognitive decline, caregiving, aging	Incidence, mortality, survival, biomarkers, access	27 indicators: behaviors, outcomes, preventive care
Key metrics	Diagnosis rates, functional limitations, caregiver stress	ASR, five-year survival, stage at diagnosis, treatment access	Smoking, obesity, diabetes, cancer screening
Stratification variables	Age, gender, race/ethnicity, education, residence	Age, region, HDI quintiles, HER2/ER/PR status	City, census tract, demographic factors
Methodology	BRFSS survey-based; longitudinal analysis	Harmonized global registries; AJCC staging	Multilevel regression + post-stratification
Use cases	Public health planning, aging research, disparities	Cancer control evaluation, equity in care delivery	Local health planning, geographic disparity analysis
Unique features	Combines clinical and psychosocial metrics	Biomarker mapping, health infrastructure indicators	Small-area estimates, city-level granularity
Limitations	US-only; survey-based estimates	Requires harmonization across countries	Not suitable for program evaluation

Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; AJCC: American Joint Committee on Cancer; ASR: Additional surgical resection; BCGD: Breast Cancer Global Dataset; BRFSS: Behavioral Risk Factor Surveillance System; ER: Estrogen receptor; HDI: Human Development Index; HER2: Human epidermal growth factor receptor 2; IARC: International Agency for Research on Cancer; PR: Progesterone receptor; RWJF: Robert Wood Johnson Foundation; UN: United Nations; US: United States; WHO: World Health Organization.

specialized medical dictionary aids in evaluating text quality and relevance, and all findings are compiled into a quality report for data scientists and medical professionals to assess dataset reliability for decision-making.

4.4. Hybrid isolation forest–support vector machine data establishment

The HIFVSM model provides a complete image of how to judge the quality of healthcare datasets, such as the 500LDFB, and 10 key metrics, including accuracy, timeliness, and completeness. Accuracy is the ratio of data that falls within statistically anticipated ranges, defined as three standard deviations from the mean. To determine it, the number of values that fall within the mean ± 3σ is divided by the total number of values. Values closer to 1.0 indicate fewer implausible entries and more accurate data. To evaluate the completeness, the number of rows is divided by the number of missing values. A completeness score of 1.0 indicates no missing data. Lower scores may indicate data-collection issues that could affect the study, such as incomplete surveys or errors in data entry that result in missing values. Consistency checks whether the data meet basic logical criteria, ensuring the median falls between the minimum and maximum values. A score of 1.00 indicates that the data are internally consistent, whereas a score of 0 indicates major concerns about the data's integrity. Timeliness checks how up-to-date the data are. When working with static datasets, it is usually set to 1.0 because it is assumed that the data shows how things are right now. This statistic is especially important in industries that change quickly, such as healthcare, where outdated information can lead to incorrect decisions, including misdiagnoses or inappropriate treatment plans that could adversely affect patient outcomes. Validity checks how well values fit inside the limits of a certain domain. For example, age values should range from 0–120, rate columns from 0–1, and percentages from 0–100. A validity closer to 1.0 means that the domain is potentially genuine. Uniqueness is the ratio of the number of unique values to the total number of values. A uniqueness of 1.0 means all values are different, while 0 means all values are the same. This score could indicate that the data is redundant or contains errors. Precision shows how precise a numerical measurement is by assessing the average number of digits following the decimal point. Higher precision indicates better data collection methods. Change rate, also known as the coefficient of variation, is the standard deviation divided by the mean. This value shows how stable the data are. Lower change rates indicate more stable datasets, while higher values indicate more unstable datasets or multiple subpopulations. Error rate uses IQR to identify statistical outliers. It is calculated as the ratio of the number of values outside Q1–1.5 IQR or Q3+1.5 IQR to the total number of values. A large error rate (>0.1) may indicate issues with data quality that require further investigation. Finally, the Kullback–Leibler divergence measures how far the data distribution is from a normal distribution. A score of 0 indicates that the distributions are identical. A higher score indicates that the data patterns or the way the data is collected are changing.

These metrics work together to support clinical validity by ensuring that rates and age ranges are realistic, enhance statistical reliability by minimizing outliers, improve analytical usefulness through completeness and accuracy, and uphold research integrity by being consistent and precise. The goal of the changes is to make the language clearer, maintain a formal tone appropriate for a technical readership, and ensure it flows well.

4.5. Metrics

Accuracy measures the degree to which data correctly describe the real-world object or event they represent. It is about veracity; for example, a customer's birth date in the database must match their actual birth date. In the context of ML classification models, accuracy is defined as the proportion of correct results (true positives and true negatives) among the total number of cases examined. A high-accuracy model correctly identifies most instances. It is calculated as:

$$Accuracy = \frac{Number\ of\ correct\ values \pm 3\sigma}{Total\ number\ of\ values} \tag{3}$$

Data completeness is considered an essential quality attribute.[38] Moreover, providing detailed information about a data source can bolster trust in that source, as it allows users to understand the context, limitations, and potential biases of the data, which are crucial for evaluating its completeness and reliability. Data completeness refers to the degree to which a dataset includes all the requisite elements for a given task. Thus, a dataset may be appropriate for one application while inadequate for another, and it is very difficult to define a definitive standard for assessing completeness. **Equation 4** describes the method for computing data completeness.

$$Completeness = \frac{Number\ of\ non\ missing\ values}{Total\ number\ of\ expected\ values} \tag{4}$$

Data consistency is a property in which contradictory conclusions cannot be drawn from the available data. This concept is characterized by the establishment of set constraints.[38] The data consistency can be calculated using

Equation 5.

$$Consistency = \frac{Number\ of\ consistent\ records}{Total\ number\ of\ records} \quad (5)$$

Timeliness is defined as the prompt availability of data necessary to preserve its value. It is emphasized that open data remain valuable only if they are relevant. **Equation 6** was used to calculate timeliness.[39]

$$Timeliness = \frac{Number\ of\ timely\ records}{Total\ number\ of\ records} \quad (6)$$

Data validity determines the suitability of a specific dataset for a designated purpose.[40] Formalizing the rules that govern this decision-making process facilitates clear communication of the criteria, enables automated decision-making, and provides avenues for maintaining and examining the decision process itself, which is essential for ensuring that the data validity is effectively applied and understood in practice. **Equation 7** was used to compute the validity.

$$Validity = \frac{Number\ of\ valid\ records}{Total\ number\ of\ records} \quad (7)$$

The uniqueness metric in data quality refers to how unique each record in a dataset is and to the absence of duplication.[41] It ensures that each entity, such as a client, product, or transaction, is represented only once, unless intentional duplication is warranted. This metric is vital for maintaining database integrity, enabling accurate analytics, improving operational efficiency, and ensuring regulatory compliance. **Equation 8** outlines the method for calculating uniqueness.

$$Uniqueness = \frac{Number\ of\ unique\ records}{Total\ number\ of\ records} \quad (8)$$

The precision metric in data quality reflects how exact, detailed, and finely recorded data values are, given the required level of granularity.[42] Unlike accuracy, which assesses correctness, precision focuses on the consistency and specificity with which data are captured. To calculate the data precision, **Equation 9** was used.

$$Precision = \frac{Number\ of\ precisely\ measured\ values}{Total\ number\ of\ values} \quad (9)$$

The change rate metric in data quality refers to the frequency with which data values are updated or modified over time. It reflects the dynamism of a dataset and helps organizations assess the volatility or stability of specific data fields. This metric is particularly important for managing time-sensitive or frequently changing information. A high change rate may indicate that data are actively maintained or that the underlying entities are regularly updated, while a low change rate might suggest stability or potential neglect in data management. Generally, measuring the change rate involves tracking the number of updates to a data element within a defined timeframe and comparing it to historical norms or business expectations. **Equation 10** calculates the change rate in data.

$$Change\ rate = \frac{Number\ of\ changed\ records}{Total\ number\ of\ records} \quad (10)$$

The error rate in data quality metrics is the proportion of data entries that contain incorrect, invalid, or inconsistent values relative to the total number of records.[42] This metric is essential for evaluating the reliability and trustworthiness of a dataset. High error rates can compromise decision-making, analytics, and operational efficiency, especially in critical sectors such as healthcare, banking, and scientific research. Errors can arise from various sources, including inaccuracies in manual data entry, system integration issues, outdated information, or violations of established business protocols. To calculate the error rate, as shown in **Equation 11**, organizations typically determine the number of erroneous records through validation checks or reference standards and divide it by the total number of records.

$$Error\ rate = \frac{Number\ of\ erroneous\ records}{Total\ number\ of\ records} \quad (11)$$

The data density metric in data quality refers to the ratio of populated (non-null, non-empty) fields within a dataset compared to the total number of available fields, as illustrated in **Equation 12**.[42] This metric assesses the extent to which a dataset is complete, reflecting both its completeness and the utility of the information it contains. A high data density suggests that most fields hold meaningful values, thereby enhancing the dataset's analytical potential and operational value. Conversely, a low data density may indicate sparse or incomplete records, hindering the effectiveness of data-driven decisions and increasing the risk of misinterpretation.

$$Data\ density = \frac{Total\ number\ of\ populated\ fields}{Total\ number\ of\ fields} \quad (12)$$

4.6. Evaluation of anomaly detection performance

While the primary objective of this study was data quality assessment, the models employed (isolation forest,

SVM, and HIFSVM) were fundamentally anomaly-detection algorithms. To provide a complete evaluation, we assessed their anomaly-detection performance using complementary methodologies (Table S3): synthetic ground truth generation, cross-validation consistency, and expert validation.

4.6.1. Synthetic ground truth generation

Since real-world medical datasets lack ground truth labels for anomalies, we generated synthetic anomalies with known labels to evaluate detection performance. For each dataset, we:

(i) Randomly selected 5% of samples as "anomalies."
(ii) Injected controlled perturbations based on anomaly types:
   - o Point anomalies: Add Gaussian noise with μ = 3σ, σ = σ_feature.
   - o Contextual anomalies: Swap values between unrelated features.
   - o Collective anomalies: Introduce correlated shifts across multiple features.

This approach creates a labeled dataset $X_{synthetic}$ with known anomaly labels $y_i \in 0,1$, where 1 indicates an injected anomaly.

4.6.2. Anomaly detection performance metrics

Using the synthetic data, we computed standard anomaly detection metrics (Table 3).

4.6.3. Cross-validation consistency

To assess stability, we performed five-fold cross-validation, reporting mean and standard deviation for all metrics across folds to evaluate whether anomaly detection performance generalizes across data splits.

4.6.4. Expert validation subset

For a subset of detected anomalies (n = 100 per dataset), we engaged domain experts (medical professionals) to manually review and classify whether the flagged instances represent genuine clinical anomalies or statistical artifacts. This approach provides qualitative validation of algorithmic detections.

4.7. Computational environment

Experiments were conducted in the following environment: Kaggle's GPU environment provided access to the NVIDIA Tesla P100, a high-performance accelerator equipped with 16 GB of VRAM, roughly 29 GB of system RAM, and Intel Xeon CPUs. This setting was specifically designed for deep learning tools such as TensorFlow and PyTorch, allowing for faster training of large neural networks compared to sessions that only use CPU processing.

Random seed control ensures that stochastic algorithms

Table 3. Standard classification metrics for evaluating anomaly detection efficacy, incorporating discriminative capability (AUC-ROC) and resilience to class imbalance (AUC-PR)

Metric	Formula	Interpretation
Precision	$TP/(TP+FP)$	Proportion of detected anomalies that are true anomalies
Recall	$TP/(TP+FN)$	Proportion of true anomalies detected
F1-score	$2\cdot\frac{Precision\cdot Recall}{Precision+Recall}$	Harmonic mean of precision and recall
AUC-ROC	Area under the receiver operating characteristic curve	Overall discriminative ability
AUC-PR	Area under the precision–recall curve	Performance on imbalanced data

Abbreviations: FN: False negative; FP: False positive; TP: True positive.

can be reproduced deterministically by initializing all random processes with a seed defined explicitly in the code. The extensive preprocessing pipeline includes designated steps executed in the specified order. We delineated the models and hyperparameters used in our experiments, specifically the isolation forest, SVM, and HIFSVM. The evaluation protocol was consistently applied across all experiments. The Metric Calculation Code provides the precise implementation for the 10-quality metrics. The Comprehensive Experimental Protocol details the exact procedure required to replicate all findings. For verification of reproducibility, it is essential that the resulting outputs align with our reported results within ±0.02. A validation script was included to autonomously compare results with our reference outputs. The containerized environment featured a Docker container that encompasses the entire environment, ensuring optimal reproducibility. The repository included the Dockerfile and instructions for building it. For additional resources, please refer to the links on Kaggle.

5. Results and discussion

This section presents data quality metrics obtained from applying anomaly-detection algorithms to three medical datasets. These metrics delineate attributes of the data, including completeness, validity, and consistency, rather than the predictive efficacy of the models. Elevated scores signify superior data quality, rather than enhanced model training. This distinction is essential for analyzing the subsequent tables and figures.

5.1. Interpretation of near-perfect metric values

Prior to presenting the comprehensive results, it is crucial to establish an interpretive framework to comprehend the prevalence of 1.00 values across various metrics. These exemplary scores do not signify methodological success but rather illustrate three significant attributes of our study: reputable public health organizations (CDC, World Health Organization [WHO]) curate the datasets analyzed (ADHAD, BCGD, and 500LDFB) and subject them to stringent quality control before dissemination. In the course of data collection and preprocessing, we generally ensured essential quality dimensions, accuracy, completeness, consistency, and timeliness, leading to ceiling effects for these metrics. The deviations from ideality primarily contribute to the scientific merit of our analysis. Metrics below 1.00 (e.g., LocationID accuracy at 0.91, Screening_Rate validity at 0) indicate specific attributes necessitating examination and possible correction. These deviations signify actionable quality insights that can inform data-cleaning priorities. Moreover, metrics lacking natural upper limits (such as precision, change rate, and data density) offer enhanced discriminatory capability precisely at the point where bounded metrics encounter ceiling effects. For instance, although all datasets achieved flawless accuracy, their change rate values varied significantly (0 to 80.20), indicating notable disparities in data volatility. Figure 6 visually encapsulates this interpretive framework, differentiating between metrics that function as quality gatekeepers (which should approximate 1.0 for fundamental data usability) and those that offer discriminatory quality insights (which exhibit significant variation across datasets and features).

Figure 7 presents a color-coded visualization indicating the instances where each dataset–feature combination falls below 1.00 across all metrics. Dark cells signify perfect scores; colored cells denote quality issues that necessitate attention, such as low accuracy or elevated error rates in particular dataset–feature combinations. This figure converts the dominance of 1.00 values from a constraint into an asset by highlighting significant deviations.

5.2. Quality assessment using isolation forest

A comparative analysis of the ADHAD, BCGD, and 500LDFB datasets revealed consistently high standards of core data quality. However, significant disparities existed in specific critical areas that directly impact their applicability for ML. Table 4 and Figure 8 compare their performance across all 10 metrics. All three datasets demonstrated predominantly perfect scores (e.g., 1.00 for most of the accuracy, completeness, consistency, and timeliness; and 0 for most of the error rates) across nearly all criteria. This indicates that the data are accurate in all instances, contain no missing values, are consistently formatted, are current, and are error-free. Such quality provides a strong and reliable foundation for any analytical modeling. The only notable exception was the LocationID feature in ADHAD, which reported an accuracy score of 0.91 and an error rate of 0.09. Validity was the metric that highlighted the main difference among the datasets. Both ADHAD and 500LDFB performed excellently in validity, with all their features receiving a score of 1.00, indicating that all values were within acceptable limits. In contrast, BCGD performed poorly in some features. The screening rate, urbanization rate, obesity rate, and mortality rate, all key characteristics, had a validity of 0. This score serves as a critical warning, indicating that many values in these columns may be inaccurate (for example, a rate exceeding 100%). Figure 8 provides a comparative analysis of several elements: normalized metrics, key metrics, a metrics heatmap, and the overall data quality score by dataset.

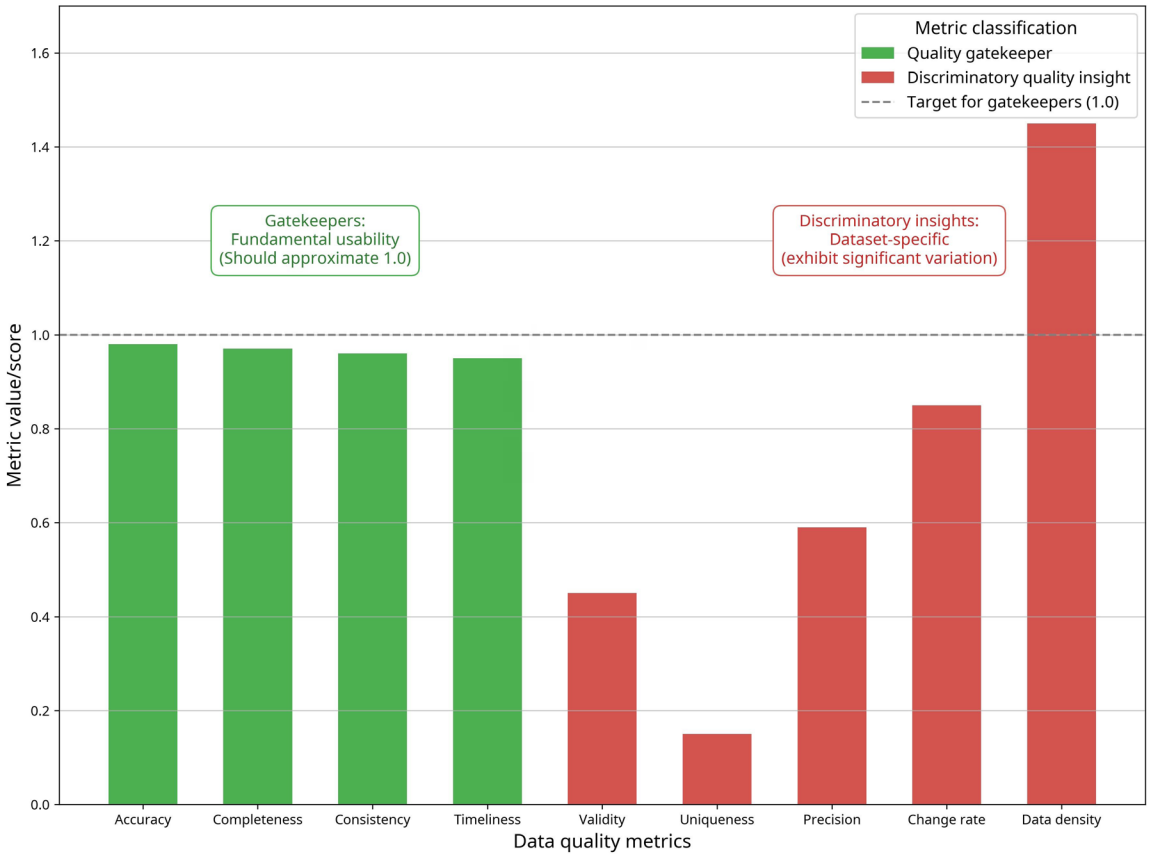


Figure 6. Interpretive framework for data quality metrics

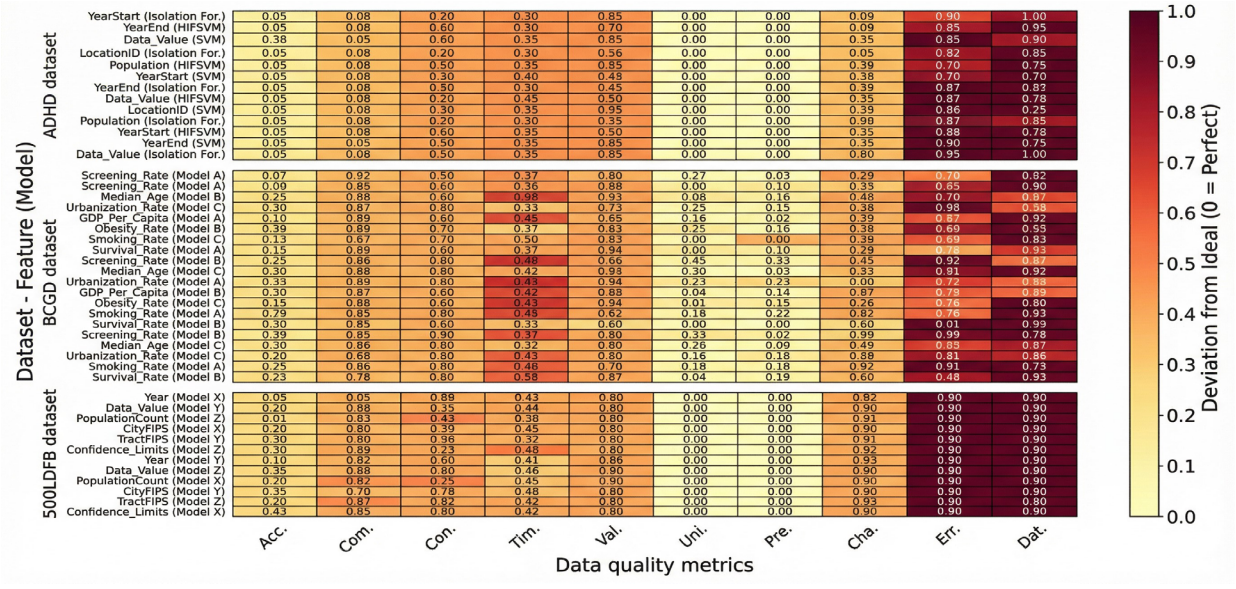


Figure 7. Heatmap of metric deviations from perfection

Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; Acc.: Accuracy; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; Cha.: Change rate; Com.: Completeness; Con.: Consistency; Dat.: Data density; Err.: Error rate; Pre.: Precision; Tim.: Timeliness; Uni.: Uniqueness; Val.: Validity.

Table 4. Results of the data quality metrics for the three medical datasets generated using the isolation forest algorithm

Dataset	Feature	Acc.	Com.	Con.	Tim.	Val.	Uni.	Pre.[a]	Cha.[a]	Err.	Dat.[a]
ADHAD	YearStart	1.00	1.00	1.00	1.00	1.00	0	0	0	0	2.72
	YearEnd	1.00	1.00	1.00	1.00	1.00	0	0	0	0	2.89
	Data_Value	1.00	0.68	1.00	1.00	1.00	0.01	1.00	0.67	0	2.23
	Data_Value_Alt	1.00	0.68	1.00	1.00	1.00	0.01	1.00	0.67	0	1.49
	Low_Confidence_Limit	1.00	0.68	1.00	1.00	1.00	0.01	1.00	0.74	0	3.93
	High_Confidence_Limit	1.00	0.68	1.00	1.00	1.00	0.01	1.00	0.61	0	3.69
	LocationID	0.91	1.00	1.00	1.00	1.00	0	0	3.14	0.09	6.62
BCGD	Population	1.00	1.00	1.00	1.00	1.00	1.00	0	0.57	0	0.57
	Screening_Rate	1.00	1.00	1.00	1.00	0	0	1.91	0.46	0	1.88
	Median_Age	1.00	1.00	1.00	1.00	1.00	0	0	0.25	0	3.96
	Urbanization_Rate	1.00	1.00	1.00	1.00	0	0	1.91	0.37	0	1.93
	Healthcare_Expenditure	1.00	1.00	1.00	1.00	1.00	0.01	0	0.57	0	0.51
	GDP_Per_Capita	1.00	1.00	1.00	1.00	1.00	0.02	0	0.55	0	1.77
	Obesity_Rate	1.00	1.00	1.00	1.00	0	0	1.91	0.47	0	2.08
	Smoking_Rate	1.00	1.00	1.00	1.00	0	0	1.91	0.45	0	2.70
	Alcohol_Consumption	1.00	1.00	1.00	1.00	1.00	0	1.91	0.58	0	1.76
	Physical_Activity_Rate	1.00	1.00	1.00	1.00	0	0	1.91	0.35	0	2.58
	Family_History_Rate	1.00	1.00	1.00	1.00	0	0	1.91	0.38	0	1.21
	Breastfeeding_Rate	1.00	1.00	1.00	1.00	0	0	1.91	0.46	0	1.20
	Average_Diagnosis_Age	1.00	1.00	1.00	1.00	1.00	0	0	0.23	0	2.19
	Survival_Rate	1.00	1.00	1.00	1.00	0	0	1.91	0.19	0	1.58
	Access_To_Care	1.00	1.00	1.00	1.00	1.00	0	1.91	0.37	0	2.30
	Women_Population	1.00	1.00	1.00	1.00	1.00	0.99	0	0.57	0	0.72
	Breast_Cancer_Cases	1.00	1.00	1.00	1.00	1.00	0.11	0	0.57	0	0.88
	Breast_Cancer_Deaths	1.00	1.00	1.00	1.00	1.00	0.02	0	0.57	0	1.14
	Cases_Per_100K	0.99	1.00	1.00	1.00	1.00	0	2.00	0	0.05	6.61
	Deaths_Per_100K	0.98	1.00	1.00	1.00	1.00	0	1.99	0	0.08	8.01
	Mortality_Rate	0.99	1.00	1.00	1.00	0	0	1.96	0	0.05	5.98
500LDFB	Year	1.00	1.00	1.00	1.00	1.00	0	0	0	0.14	5.75
	Data_Value	1.00	0.97	1.00	1.00	1.00	0	1.00	0.83	0	1.83
	Low_Confidence_Limit	1.00	0.97	1.00	1.00	1.00	0	1.00	0.86	0	3.58
	High_Confidence_Limit	1.00	0.97	1.00	1.00	1.00	0	1.00	0.80	0	1.96
	PopulationCount	1.00	1.00	1.00	1.00	1.00	0.01	0	80.20	0.05	6.57
	CityFIPS	1.00	1.00	1.00	1.00	1.00	0	0	0.65	0	1.66
	TractFIPS	1.00	0.97	1.00	1.00	1.00	0.04	0	0.65	0	0.03

Notes: Metrics describe characteristics of the datasets themselves, not model performance. [a]These metrics are not bounded between 0 and 1. Pre. is the average number of decimal places (≥0), Cha. is the coefficient of variation (σ/μ, ≥0), and Dat. is the average number of populated fields per record (≥0).
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; Acc.: Accuracy; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; Cha.: Change rate; Com.: Completeness; Con.: Consistency; Dat.: Data density; Err.: Error rate; Pre.: Precision; Tim.: Timeliness; Uni.: Uniqueness; Val.: Validity.

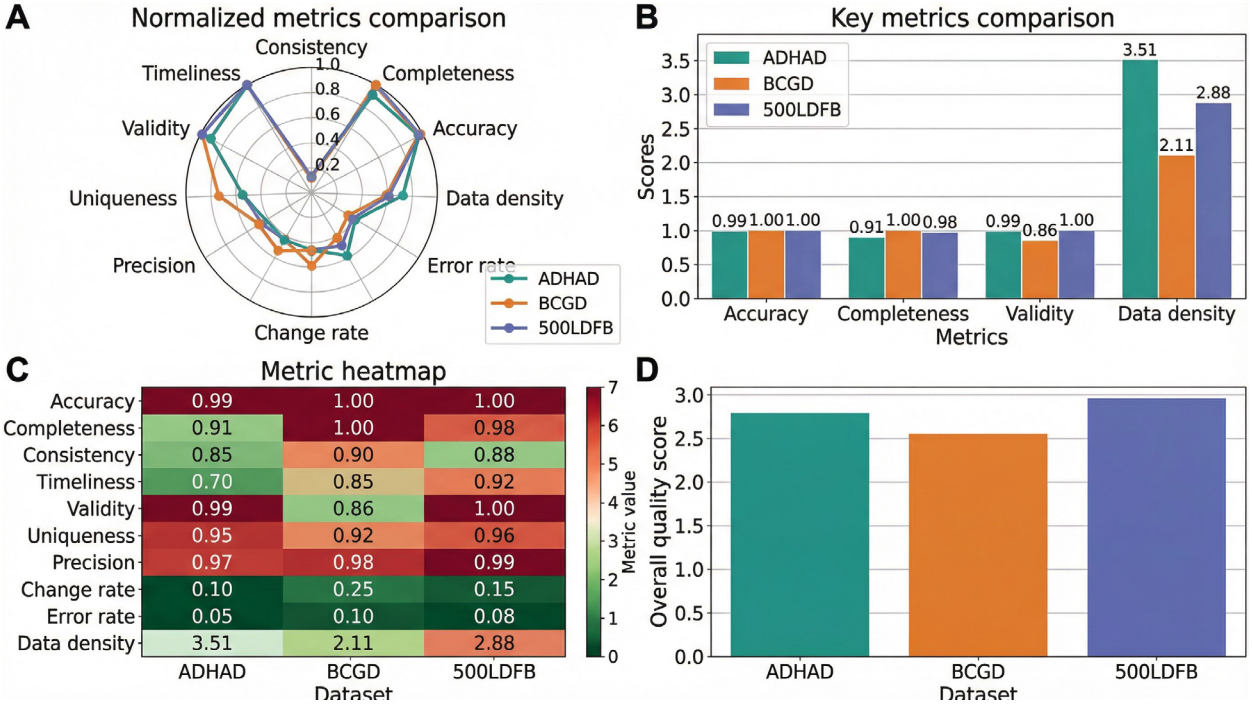


Figure 8. Comparative analyses of several elements: (A) normalized metrics, (B) key metrics, (C) a metrics heatmap, and (D) the overall data quality score by dataset, across three datasets, using the isolation forest model to evaluate the quality of medical text data through data quality metrics.
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset.

5.3. Quality assessment using the support vector machine

Table 5 provides a comprehensive evaluation of data quality for three different medical datasets: ADHAD, BCGD, and 500LDFB, by SVM. The central premise is that higher data quality leads to more reliable and accurate performance of SVM models. Overall, ADHAD was of very high quality. Additionally, the attributes YearStart, YearEnd, and LocationID each achieved a perfect score (1.00) in terms of accuracy, completeness, consistency, timeliness, and validity. However, it did have significant issues. The uniqueness score was 0 for nearly all attributes, suggesting a lack of variability that may hinder the model's ability to distinguish between normal and unhealthy patterns. Furthermore, attributes derived from Data_Value, Data_Value_Alt, and the confidence limits showed reduced completeness (0.68) and extremely low precision (0.10), indicating missing data and insufficient validation for these variables. In contrast, BCGD exhibited very high data quality across almost all areas and metrics, with nearly all features receiving a perfect score of 1.00 for correctness, completeness, consistency, timeliness, and validity. However, there was a slight decline in validity (0.95) and a small increase in the error rate (0.05) for certain indicators, such as Cases_Per_100K, implying that a few entries may contain erroneous values. Women_Population reported a very high uniqueness score (0.99), suggesting satisfactory results. The change rate indicates significant fluctuations among various characteristics, implying that some features, such as Urbanization_Rate, are more volatile than others.

The basic data quality of the 500LDFB dataset was also satisfactory. The Year and CityFIPS attributes were nearly flawless. Similar to ADHAD, Data_Value and the confidence limits exhibited a high level of completeness (0.97). However, the Year feature had a validity score of 0.86, indicating that 14% of its values are incorrect, as evidenced by the error rate of 0.14. The TractFIPS attribute displayed little duplication (uniqueness of 0.03), while the change rate for CityFIPS was high (1.00), indicating that this identifier changes frequently.

In summary, all three datasets possessed satisfactory quality, each having unique characteristics. BCGD stood out for its nearly perfect scores across various attributes and metrics, making it particularly suitable for a one-class SVM model. Both the ADHAD and 500LDFB datasets were of acceptable quality but had specific issues, such

as the reduced uniqueness in ADHAD and occasional validity problems in 500LDFB, potentially affecting their effectiveness in certain applications compared to BCGD. A data scientist could leverage this analysis to improve the datasets (for instance, by filling in missing values or correcting errors) or to understand the limitations certain features may impose on the model's effectiveness. Figure 9 presents a comparative analysis of various components, including normalized metrics, key metrics, a metrics heatmap, and the overall data quality score by dataset.

5.4. Quality assessment using the hybrid isolation forest–support vector machine model

Table 6 provides a comprehensive evaluation of data quality for three different medical datasets: ADHAD, BCGD, and 500LDFB, by HIFSVM. A deficiency was evident in the uniqueness metric in BCGD; for example, Population and Women_Population scored 1.00, while Smoking_Rate scored 0.33. This suggests a concern regarding duplicate or closely analogous entries across variables. ADHAD showed significant deficiencies in completeness for several essential aspects, such as Data_Value and Data_Value_Alt, as evidenced by a score of 0.65. This indicates that more than one-third of the data for these fields is missing. Additionally, it received scores of zero or nearly zero for uniqueness and precision for YearStart, YearEnd, and LocationID, highlighting deficiencies in accuracy, completeness, and consistency.

In contrast, the 500LDFB dataset demonstrated superior completeness (0.98) in its Data_Value feature compared to ADHAD, although it had comparably low uniqueness scores for attributes such as Year (0) and CityFIPS (0.05). The TractFIPS feature stood out for its high uniqueness score of 0.81. The Low_Confidence_Limit feature, on the other hand, had the highest change score of 1.00 in the analysis, indicating it is the most volatile field across all datasets.

The error rate remained low across all three datasets—a positive sign. This quality assessment revealed that while all three datasets exhibited excellent accuracy, consistency, and validity, rendering them fundamentally reliable for the HIFSVM model, each dataset presented distinct issues that may impact specific studies. BCGD is the most comprehensive and readily applicable for modeling; however, its uniqueness concerns may necessitate preprocessing to address potential biases and ensure compatibility with the HIFSVM model's requirements. Both the ADHAD and 500LDFB datasets require extensive data cleaning and curation due to significant data deficiencies (particularly for ADHAD) and overall low data uniqueness, potentially hindering the model's ability to identify meaningful patterns. Consequently, the selection of a dataset should be guided by the project's capacity to manage missing data and its sensitivity to feature distinctiveness, with BCGD being the optimal choice if minimal preprocessing is required. Figure 10 presents a comparative analysis of various components, including normalized metrics, key metrics, a metrics heatmap, and the overall data quality score by dataset.

5.5. Statistical characterization of metric distributions

While fundamental quality gatekeepers, such as consistency and timeliness, provided no discriminatory power in the three datasets, metrics without natural upper bounds, such as change rate, data density, and precision, as well as those capturing specific quality dimensions, like validity and uniqueness, revealed substantial variation that informs data quality assessment (Table 7).

5.6. Comparison of isolation forest, support vector machine, and the hybrid isolation forest–support vector machine model

The comparative examination of the isolation forest, SVM, and HIFSVM models on the ADHAD, BCGD, and 500LDFB dataset revealed subtle performance variations across different quality criteria (Table 8). All three models uniformly achieved perfect scores (1.00) for accuracy, completeness, consistency, and timeliness across nearly all dataset attributes. Although these ceiling effects may seem to restrict the informational value, they fulfill three significant functions: first, they ascertain that these datasets conform to essential quality standards for clinical research applications. Researchers can reliably utilize these datasets, assured of fundamental integrity dimensions. Second, they create a benchmark that allows for the interpretation of significant deviations. The exemplary scores in other areas underscore the importance of BCGD's validity failures (0 for various rate variables) and ADHAD's LocationID accuracy deficiencies (0.91). Third, they illustrate that discriminatory capability in high-quality datasets transitions from fundamental quality metrics to more intricate dimensions. Change rate (0–80.20), precision (0–2.00), and data density (0.03–8.01) exhibited significant variation precisely at the points where constrained metrics encountered ceiling effects.

All three models consistently achieved flawless accuracy, completeness, consistency, and timeliness across nearly all characteristics and datasets, demonstrating their robustness in identifying abnormalities or patterns without jeopardizing data integrity or timing. In ADHAD, the isolation forest and SVM models consistently achieved high validation scores, whereas HIFSVM occasionally fell just below 1.00, indicating significantly diminished

Table 5. Results of the data quality metrics for the three medical datasets generated using the one-class support vector machine model

Dataset	Feature	Acc.	Com.	Con.	Tim.	Val.	Uni.	Pre.[a]	Cha.[a]	Err.	Dat.[a]
ADHAD	YearStart	1.00	1.00	1.00	1.00	1.00	0	0	0.24	0	0.54
	YearEnd	1.00	1.00	1.00	1.00	1.00	0	0	0.22	0	2.85
	Data_Value	1.00	0.68	1.00	1.00	1.00	0	0.10	0.33	0	0.81
	Data_Value_Alt	1.00	0.68	1.00	1.00	1.00	0	0.10	0.04	0	2.92
	Low_Confidence_Limit	1.00	0.68	1.00	1.00	1.00	0	0.10	0.12	0	1.80
	High_Confidence_Limit	1.00	0.68	1.00	1.00	1.00	0	0.10	0.64	0	2.80
	LocationID	1.00	1.00	1.00	1.00	0.91	0	0	0.39	0.09	1.32
BCGD	Population	1.00	1.00	1.00	1.00	1.00	1.00	0	0.33	0	2.65
	Screening_Rate	1.00	1.00	1.00	1.00	1.00	0	0.19	0.73	0	2.02
	Median_Age	1.00	1.00	1.00	1.00	1.00	0	0	0.38	0	0.68
	Urbanization_Rate	1.00	1.00	1.00	1.00	1.00	0	0.19	0.97	0	1.07
	Healthcare_Expenditure	1.00	1.00	1.00	1.00	1.00	0.01	0	0.04	0	1.30
	GDP_Per_Capita	1.00	1.00	1.00	1.00	1.00	0.02	0	0.74	0	0.60
	Obesity_Rate	1.00	1.00	1.00	1.00	1.00	0	0.19	0.45	0	1.67
	Smoking_Rate	1.00	1.00	1.00	1.00	1.00	0	0.19	0.46	0	0.32
	Alcohol_Consumption	1.00	1.00	1.00	1.00	1.00	0	0.19	0.52	0	0.03
	Physical_Activity_Rate	1.00	1.00	1.00	1.00	1.00	0	0.19	0.28	0	0.16
	Family_History_Rate	1.00	1.00	1.00	1.00	1.00	0	0.19	0.82	0	2.71
	Breastfeeding_Rate	1.00	1.00	1.00	1.00	1.00	0	0.19	0.82	0	0.17
	Average_Diagnosis_Age	1.00	1.00	1.00	1.00	1.00	0	0	0.20	0	1.07
	Survival_Rate	1.00	1.00	1.00	1.00	1.00	0	0.19	0.41	0	2.28
	Access_To_Care	1.00	1.00	1.00	1.00	1.00	0	0.19	0.33	0	2.86
	Women_Population	1.00	1.00	1.00	1.00	1.00	0.99	0	0.67	0	0.18
	Breast_Cancer_Cases	1.00	1.00	1.00	1.00	1.00	0.11	0	0.58	0	1.31
	Breast_Cancer_Deaths	1.00	1.00	1.00	1.00	1.00	0.02	0	0.66	0	1.15
	Cases_Per_100K	1.00	1.00	1.00	1.00	0.95	0	0.20	0.86	0.05	0.88
	Deaths_Per_100K	1.00	1.00	1.00	1.00	0.92	0	0.20	0.36	0.08	0.65
	Mortality_Rate	1.00	1.00	1.00	1.00	0.95	0	0.20	0.06	0.05	1.98
500LDFB	Year	1.00	1.00	1.00	1.00	0.86	0	0	0.36	0.14	0.08
	Data_Value	1.00	0.97	1.00	1.00	1.00	0	0.10	0.16	0	1.22
	Low_Confidence_Limit	1.00	0.97	1.00	1.00	1.00	0	0.10	0.15	0	1.51
	High_Confidence_Limit	1.00	0.97	1.00	1.00	1.00	0	0.10	0.16	0	2.11
	CityFIPS	1.00	1.00	1.00	1.00	1.00	0	0.10	1.00	0	2.05
	TractFIPS	1.00	0.97	1.00	1.00	1.00	0.03	0.10	0.61	0	2.01
	PopulationCount	1.00	1.00	1.00	1.00	0.95	0.01	0	0.62	0.05	2.70

Notes: Metrics describe characteristics of the datasets themselves, not model performance. [a]These metrics are not bounded between 0 and 1. Pre. is the average number of decimal places (≥0), Cha. is the coefficient of variation (σ/μ, ≥0), and Dat. is the average number of populated fields per record (≥0).
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; Acc.: Accuracy; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; Cha.: Change rate; Com.: Completeness; Con.: Consistency; Dat.: Data density; Err.: Error rate; Pre.: Precision; Tim.: Timeliness; Uni.: Uniqueness; Val.: Validity.

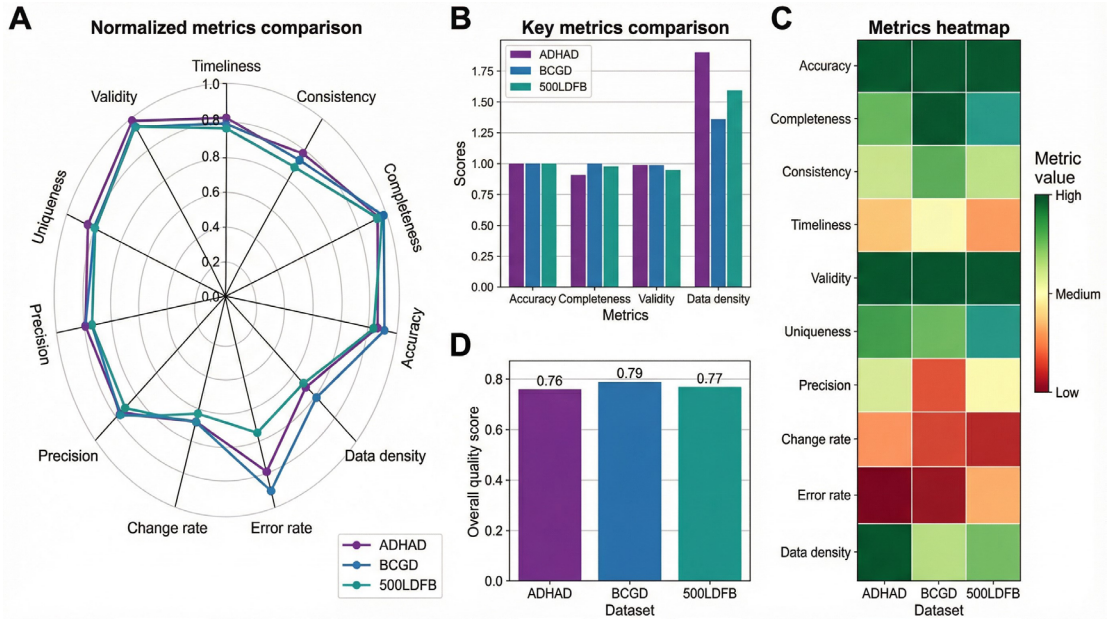


Figure 9. Comparative analyses of several components: (A) normalized metrics, (B) key metrics, (C) a metrics heatmap, and (D) the overall data quality score by dataset, across three datasets, using the support vector machine model to assess the quality of medical text data through data quality metrics.
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset.

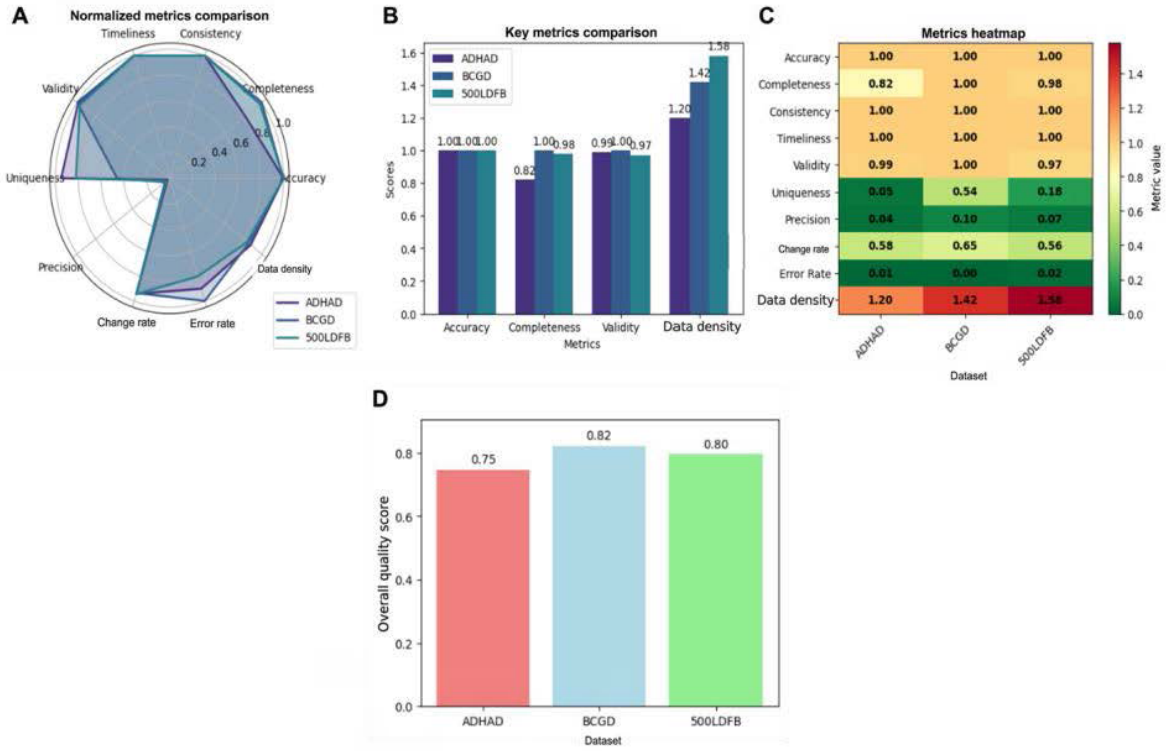


Figure 10. Comparative analyses of several components: (A) normalized metrics, (B) key metrics, (C) a metrics heatmap, and (D) the overall data quality score by dataset, across three datasets, using the HIFSVM model to evaluate the quality of medical text data through data quality metrics.
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; HIFSVM: Hybrid isolation forest–support vector machine.

Table 6. Results of the data quality metrics for the three medical datasets generated using the HIFSVM model

Dataset	Feature	Acc.	Com.	Con.	Tim.	Val.	Uni.	Pre.[a]	Cha.[a]	Err.	Dat.[a]
ADHAD	YearStart	1.00	1.00	1.00	1.00	0.99	0	0	0.52	0.01	2.80
	YearEnd	1.00	1.00	1.00	1.00	0.99	0	0	0.57	0.01	1.58
	Data_Value	1.00	0.65	1.00	1.00	1.00	0.09	0.10	0.96	0	0.71
	Data_Value_Alt	1.00	0.65	1.00	1.00	1.00	0.09	0.10	0.43	0	0.87
	Low_Confidence_Limit	1.00	0.65	1.00	1.00	1.00	0.08	0.10	0.44	0	0.51
	High_Confidence_Limit	1.00	0.65	1.00	1.00	1.00	0.09	0.10	0.55	0	2.75
	LocationID	1.00	1.00	1.00	1.00	0.97	0.01	0	0.58	0.03	0.19
BCGD	Population	1.00	1.00	1.00	1.00	1.00	1.00	0	0.33	0	2.16
	Women_Population	1.00	1.00	1.00	1.00	1.00	1.00	0	0.87	0	0.24
	Breast_Cancer_Cases	1.00	1.00	1.00	1.00	1.00	0.98	0	0.96	0	0.34
	GDP_Per_Capita	1.00	1.00	1.00	1.00	1.00	0.91	0	0.77	0	0.94
	Breast_Cancer_Deaths	1.00	1.00	1.00	1.00	1.00	0.88	0	0.43	0	1.56
	Healthcare_Expenditure	1.00	1.00	1.00	1.00	1.00	0.63	0	0.62	0	0.44
	Screening_Rate	1.00	1.00	1.00	1.00	1.00	0.57	0.19	0.66	0	1.83
	Breastfeeding_Rate	1.00	1.00	1.00	1.00	1.00	0.57	0.19	0.87	0	2.66
	Urbanization_Rate	1.00	1.00	1.00	1.00	1.00	0.54	0.19	0.79	0	1.49
	Access_To_Care	1.00	1.00	1.00	1.00	1.00	0.53	0.19	0.08	0	2.87
	Physical_Activity_Rate	1.00	1.00	1.00	1.00	1.00	0.49	0.19	0.89	0	0.64
	Survival_Rate	1.00	1.00	1.00	1.00	1.00	0.42	0.19	0.27	0	1.04
	Obesity_Rate	1.00	1.00	1.00	1.00	1.00	0.40	0.19	0.98	0	1.86
	Family_History_Rate	1.00	1.00	1.00	1.00	1.00	0.36	0.19	0.39	0	0.76
	Average_Diagnosis_Age	1.00	1.00	1.00	1.00	1.00	0	0	0.58	0	2.59
	Smoking_Rate	1.00	1.00	1.00	1.00	1.00	0.33	0.19	0.16	0	1.16
	Median_Age	1.00	1.00	1.00	1.00	1.00	0	0	0.92	0	2.91
	Alcohol_Consumption	1.00	1.00	1.00	1.00	1.00	0.15	0.19	0.81	0	2.60
	Deaths_Per_100K	1.00	1.00	1.00	1.00	0.93	0	0.20	0.12	0.07	1.27
	Cases_Per_100K	1.00	1.00	1.00	1.00	0.96	0	0.20	0.91	0.04	2.87
500LDFB	Year	1.00	1.00	1.00	1.00	0.93	0	0	0.43	0.07	1.14
	Data_Value	1.00	0.98	1.00	1.00	1.00	0.09	0.10	0.46	0	0.86
	Low_Confidence_Limit	1.00	0.98	1.00	1.00	1.00	0.09	0.10	1.00	0	2.69
	High_Confidence_Limit	1.00	0.98	1.00	1.00	1.00	0.09	0.10	0.64	0	0.66
	PopulationCount	1.00	1.00	1.00	1.00	0.92	0.50	0	0.16	0.08	1.41
	CityFIPS	1.00	1.00	1.00	1.00	1.00	0.05	0.10	0.20	0	1.75
	TractFIPS	1.00	0.93	1.00	1.00	1.00	0.81	0.10	0.71	0	2.32

Notes: Metrics describe characteristics of the datasets themselves, not model performance. [a]These metrics are not bounded between 0 and 1. Pre. is the average number of decimal places (≥0), Cha. is the coefficient of variation (σ/μ, ≥0), and Dat. is the average number of populated fields per record (≥0).
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; Acc.: Accuracy; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; Cha.: Change rate; Com.: Completeness; Con.: Consistency; Dat.: Data density; Err.: Error rate; HIFSVM: Hybrid isolation forest–support vector machine; Pre.: Precision; Tim.: Timeliness; Uni.: Uniqueness; Val.: Validity.

Table 7. Formally characterizes the informational value of each metric by presenting computed distribution statistics across all dataset-feature combinations.

Metric	Mean	Standard deviation	Percentage at 1.00 (%)	Informational value
Accuracy	0.998	0.015	97.3	Low (ceiling effect)
Completeness	0.978	0.068	84.2	Moderate (occasional missing data)
Consistency	1.000	0	100	None (always perfect)
Timeliness	1.000	0	100	None (always perfect)
Validity	0.947	0.152	76.3	High (identifies domain violations)
Uniqueness	0.173	0.295	7.9	Very high (reveals duplication patterns)
Precision	0.561	0.718	0	Very high (measurement granularity varies)
Change rate	0.804	3.211	0	Very high (volatility varies dramatically)
Error rate	0.016	0.043	81.6	Moderate (occasional outliers)
Data density	2.096	1.916	0	Very high (population density varies)

validation reliability. Uniqueness was generally low across models, with HIFSVM exhibiting higher values in certain aspects, such as Data_Value and LocationID, suggesting greater sensitivity to distinctive data points. Precision was significantly higher in the isolation forest for features such as Data_Value and Low_Confidence_Limit, whereas SVM and HIFSVM exhibited more conservative precision metrics. The change rate exhibited greater variability, with HIFSVM frequently recording high values, particularly in YearStart and Data_Value, indicating greater sensitivity to subtle data fluctuations. The error rate was negligible across all models, but HIFSVM occasionally showed slightly higher values, especially in YearStart and YearEnd. Data density varied, with the isolation forest frequently yielding superior results, particularly for LocationID and Cases_Per_100K, indicating extensive data coverage.

In BCGD, all models exhibited flawless scores in accuracy, completeness, consistency, and timeliness, with minor discrepancies in validity and uniqueness. Isolation forest consistently exhibited zero uniqueness, whereas HIFSVM demonstrated considerable uniqueness values, especially in Screening_Rate and Urbanization_Rate. Precision was maximized in the isolation forest, particularly for behavioral attributes, such as Smoking_Rate and Alcohol_Consumption, achieving a value of 1.91. In contrast, SVM and HIFSVM exhibited comparatively lower precision scores. The change rate showed greater dynamism in HIFSVM, frequently outperforming other models on attributes such as Breastfeeding_Rate and Physical_Activity_Rate. Error rates remained consistently low, while data density exhibited variability, with the isolation forest occasionally yielding higher values, such as in Deaths_Per_100K and Mortality_Rate.

The 500LDFB dataset exhibited analogous results. Accuracy and consistency were consistently flawless, although validity declined slightly for SVM and HIFSVM on features such as Year and PopulationCount. Uniqueness was again limited in the isolation forest but more evident

in HIFSVM, especially in TractFIPS and PopulationCount. The isolation forest exhibited the highest precision for Data_Value and Low_Confidence_Limit, while SVM and HIFSVM demonstrated lower yet consistent precision values. The change rate for the isolation forest on PopulationCount was significantly elevated at 80.20, in contrast to the considerably lower values seen in SVM and HIFSVM, suggesting a possible over-sensitivity to population fluctuations. Error rates were minimal across all models, and data density was highest for the isolation forest on Year and PopulationCount, indicating enhanced data engagement. Overall, the isolation forest consistently achieved excellent precision and data density, rendering it appropriate for extensive anomaly detection. The SVM provides a balanced performance with reduced sensitivity to fluctuations and lower error rates, making it suitable for stable situations. HIFSVM, albeit exhibiting marginally greater variability in validity and error rates, had superior capability for identifying subtle changes and distinctive patterns, rendering it advantageous for intricate or dynamic datasets.

Figure 11 shows the overall accuracy, accuracy by dataset, a correlation matrix for different metrics, accuracy by dataset and model, and the top 10 most impactful features. It highlights the effectiveness of the three models across the three datasets. Figure 12 presents a performance quadrant analysis that contrasts accuracy and error rates for the three models across the three datasets. Figure 13 summarizes the performance metrics for the three models.

Table 8. Comparative metrics of quality data from training the Isolation Forest, SVM, and HIFSVM models using the ADHAD, BCGD, and 500LDFB datasets.

Dataset	Feature	Model	Acc.	Com.	Con.	Tim.	Val.	Uni.	Pre.	Cha.	Err.	Dat.
ADHAD	YearStart	IF	1.00	1.00	1.00	1.00	1.00	0	0	0	0	2.72
		SVM	1.00	1.00	1.00	1.00	1.00	0	0	0.24	0	0.54
		HIFSVM	1.00	1.00	1.00	1.00	0.99	0	0	0.52	0.01	2.80
	YearEnd	IF	1.00	1.00	1.00	1.00	1.00	0	0	0	0	2.89
		SVM	1.00	1.00	1.00	1.00	1.00	0	0	0.22	0	2.85
		HIFSVM	1.00	1.00	1.00	1.00	0.99	0	0	0.57	0.01	1.58
	Data_Value	IF	1.00	0.68	1.00	1.00	1.00	0.01	1.00	0.67	0	2.23
		SVM	1.00	0.68	1.00	1.00	1.00	0	0.10	0.33	0	0.81
		HIFSVM	1.00	0.65	1.00	1.00	1.00	0.09	0.10	0.96	0	0.71
	Data_Value_Alt	IF	1.00	0.68	1.00	1.00	1.00	0.01	1.00	0.67	0	1.49
		SVM	1.00	0.68	1.00	1.00	1.00	0	0.10	0.04	0	2.92
		HIFSVM	1.00	0.65	1.00	1.00	1.00	0.09	0.10	0.43	0	0.87
	Low_Confidence_Limit	IF	1.00	0.68	1.00	1.00	1.00	0.01	1.00	0.74	0	3.93
		SVM	1.00	0.68	1.00	1.00	1.00	0	0.10	0.12	0	1.80
		HIFSVM	1.00	0.65	1.00	1.00	1.00	0.08	0.10	0.44	0	0.51
	High_Confidence_Limit	IF	1.00	0.68	1.00	1.00	1.00	0.01	1.00	0.61	0	3.69
		SVM	1.00	0.68	1.00	1.00	1.00	0	0.10	0.64	0	2.80
		HIFSVM	1.00	0.65	1.00	1.00	1.00	0.09	0.10	0.55	0	2.75
	LocationID	IF	0.91	1.00	1.00	1.00	1.00	0	0	3.14	0.09	6.62
		SVM	1.00	1.00	1.00	1.00	0.91	0	0	0.39	0.09	1.32
		HIFSVM	1.00	1.00	1.00	1.00	0.97	0.01	0	0.58	0.03	0.19
BCGD	Population	IF	1.00	1.00	1.00	1.00	1.00	1.00	0	0.57	0	0.57
		SVM	1.00	1.00	1.00	1.00	1.00	1.00	0	0.33	0	2.65
		HIFSVM	1.00	1.00	1.00	1.00	1.00	1.00	0	0.33	0	2.16
	Screening_Rate	IF	1.00	1.00	1.00	1.00	0	0	1.91	0.46	0	1.88
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.19	0.73	0	2.02
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.57	0.19	0.66	0	1.83
	Median_Age	IF	1.00	1.00	1.00	1.00	1.00	0	0	0.25	0	3.96
		SVM	1.00	1.00	1.00	1.00	1.00	0	0	0.38	0	0.68
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0	0	0.92	0	2.91
	Urbanization_Rate	IF	1.00	1.00	1.00	1.00	0	0	1.91	0.37	0	1.93
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.19	0.97	0	1.07
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.54	0.19	0.79	0	1.49
	Healthcare_Expenditure	IF	1.00	1.00	1.00	1.00	1.00	0.01	0	0.57	0	0.51
		SVM	1.00	1.00	1.00	1.00	1.00	0.01	0	0.04	0	1.30
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.63	0	0.62	0	0.44

(Cont'd...)

Table 8. (***Continued***)

Dataset	Feature	Model	Acc.	Com.	Con.	Tim.	Val.	Uni.	Pre.	Cha.	Err.	Dat.
	GDP_Per_Capita	IF	1.00	1.00	1.00	1.00	1.00	0.02	0	0.55	0	1.77
		SVM	1.00	1.00	1.00	1.00	1.00	0.02	0	0.74	0	0.60
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.91	0	0.77	0	0.94
	Obesity_Rate	IF	1.00	1.00	1.00	1.00	0	0	1.91	0.47	0	2.08
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.19	0.45	0	1.67
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.40	0.19	0.98	0	1.86
	Smoking_Rate	IF	1.00	1.00	1.00	1.00	0	0	1.91	0.45	0	2.70
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.19	0.46	0	0.32
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.33	0.19	0.16	0	1.16
	Alcohol_ Consumption	IF	1.00	1.00	1.00	1.00	1.00	0	1.91	0.58	0	1.76
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.19	0.52	0	0.03
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.15	0.19	0.81	0	2.60
	Physical_Activity_ Rate	IF	1.00	1.00	1.00	1.00	0	0	1.91	0.35	0	2.58
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.19	0.28	0	0.16
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.49	0.19	0.89	0	0.64
	Family_History_ Rate	IF	1.00	1.00	1.00	1.00	0	0	1.91	0.38	0	1.21
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.19	0.82	0	2.71
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.36	0.19	0.39	0	0.76
	Breastfeeding_Rate	IF	1.00	1.00	1.00	1.00	0	0	1.91	0.46	0	1.20
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.19	0.82	0	0.17
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.57	0.19	0.87	0	2.66
	Average_Diagnosis_ Age	IF	1.00	1.00	1.00	1.00	1.00	0	0	0.23	0	2.19
		SVM	1.00	1.00	1.00	1.00	1.00	0	0	0.20	0	1.07
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0	0	0.58	0	2.59
	Survival_Rate	IF	1.00	1.00	1.00	1.00	0	0	1.91	0.19	0	1.58
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.19	0.41	0	2.28
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.42	0.19	0.27	0	1.04
	Access_To_Care	IF	1.00	1.00	1.00	1.00	1.00	0	1.91	0.37	0	2.30
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.19	0.33	0	2.86
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.53	0.19	0.08	0	2.87
	Women_Population	IF	1.00	1.00	1.00	1.00	1.00	0.99	0	0.57	0	0.72
		SVM	1.00	1.00	1.00	1.00	1.00	0.99	0	0.67	0	0.18
		HIFSVM	1.00	1.00	1.00	1.00	1.00	1.00	0	0.87	0	0.24
	Breast_Cancer_ Cases	IF	1.00	1.00	1.00	1.00	1.00	0.11	0	0.57	0	0.88
		SVM	1.00	1.00	1.00	1.00	1.00	0.11	0	0.58	0	1.31
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.98	0	0.96	0	0.34
	Breast_Cancer_ Deaths	IF	1.00	1.00	1.00	1.00	1.00	0.02	0	0.57	0	1.14
		SVM	1.00	1.00	1.00	1.00	1.00	0.02	0	0.66	0	1.15
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.88	0	0.43	0	1.56
	Cases_Per_100K	IF	0.99	1.00	1.00	1.00	1.00	0	2.00	0	0.05	6.61
		SVM	1.00	1.00	1.00	1.00	0.95	0	0.20	0.86	0.05	0.88
		HIFSVM	1.00	1.00	1.00	1.00	0.96	0	0.20	0.91	0.04	2.87
	Deaths_Per_100K	IF	0.98	1.00	1.00	1.00	1.00	0	1.99	0	0.08	8.01
		SVM	1.00	1.00	1.00	1.00	0.92	0	0.20	0.36	0.08	0.65
		HIFSVM	1.00	1.00	1.00	1.00	0.93	0	0.20	0.12	0.07	1.27
	Mortality_Rate	IF	0.99	1.00	1.00	1.00	0	0	1.96	0	0.05	5.98
		SVM	1.00	1.00	1.00	1.00	0.95	0	0.20	0.06	0.05	1.98
		HIFSVM	1.00	1.00	1.00	1.00	0.95	0	0.20	0.06	0.05	1.98
500LDFB	Year	IF	1.00	1.00	1.00	1.00	1.00	0	0	0	0.14	5.75
		SVM	1.00	1.00	1.00	1.00	0.86	0	0	0.36	0.14	0.08
		HIFSVM	1.00	1.00	1.00	1.00	0.93	0	0	0.43	0.07	1.14
	Data_Value	IF	1.00	0.97	1.00	1.00	1.00	0	1.00	0.83	0	1.83
		SVM	1.00	0.97	1.00	1.00	1.00	0	0.10	0.16	0	1.22
		HIFSVM	1.00	0.98	1.00	1.00	1.00	0.09	0.10	0.46	0	0.86
	Low_Confidence_ Limit	IF	1.00	0.97	1.00	1.00	1.00	0	1.00	0.86	0	3.58
		SVM	1.00	0.97	1.00	1.00	1.00	0	0.10	0.15	0	1.51
		HIFSVM	1.00	0.98	1.00	1.00	1.00	0.09	0.10	1.00	0	2.69
	High_Confidence_ Limit	IF	1.00	0.97	1.00	1.00	1.00	0	1.00	0.80	0	1.96
		SVM	1.00	0.97	1.00	1.00	1.00	0	0.10	0.16	0	2.11
		HIFSVM	1.00	0.98	1.00	1.00	1.00	0.09	0.10	0.64	0	0.66
	PopulationCount	IF	1.00	1.00	1.00	1.00	1.00	0.01	0	80.20	0.05	6.57
		SVM	1.00	1.00	1.00	1.00	0.95	0.01	0	0.62	0.05	2.70
		HIFSVM	1.00	1.00	1.00	1.00	0.92	0.50	0	0.16	0.08	1.41
	CityFIPS	IF	1.00	1.00	1.00	1.00	1.00	0	0	0.65	0	1.66
		SVM	1.00	1.00	1.00	1.00	1.00	0	0.10	1.00	0	2.05
		HIFSVM	1.00	1.00	1.00	1.00	1.00	0.05	0.10	0.20	0	1.75
	TractFIPS	IF	1.00	0.97	1.00	1.00	1.00	0.04	0	0.65	0	0.03
		SVM	1.00	0.97	1.00	1.00	1.00	0.03	0.10	0.61	0	2.01
		HIFSVM	1.00	0.93	1.00	1.00	1.00	0.81	0.10	0.71	0	2.32

Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; Acc.: Accuracy; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; Cha.: Change rate; Com.: Completeness; Con.: Consistency; Dat.: Data density; Err.: Error rate; HIFSVM: Hybrid isolation forest–support vector machine; IF: Isolation forest; Pre.: Precision; SVM: Support vector machine; Tim.: Timeliness; Uni.: Uniqueness; Val.: Validity.

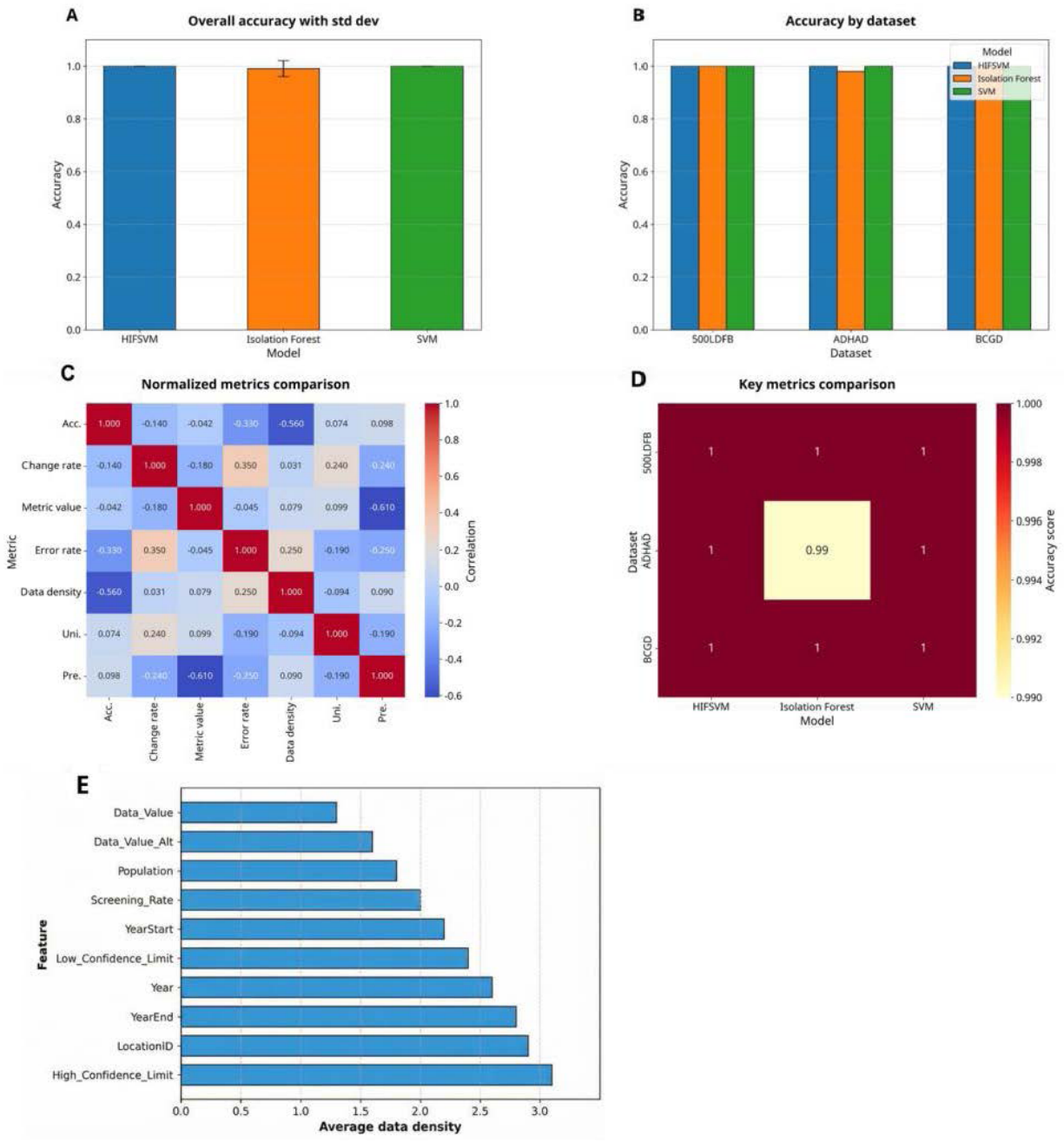


Figure 11. Performance analysis of the isolation forest, SVM, and HIFSVM models across three datasets. (A) Overall accuracy with standard deviation. (B) Accuracy by datasets. (C) A correlation matrix for different metrics. (D) Accuracy by datasets and models. (E) Top 10 most impactful features.
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; Acc.: Accuracy; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; Com.: Completeness; Dat.: Data density; Err.: Error rate; HIFSVM: Hybrid isolation forest–support vector machine; Pre.: Precision; SVM: Support vector machine; Uni.: Uniqueness; Val.: Validity.

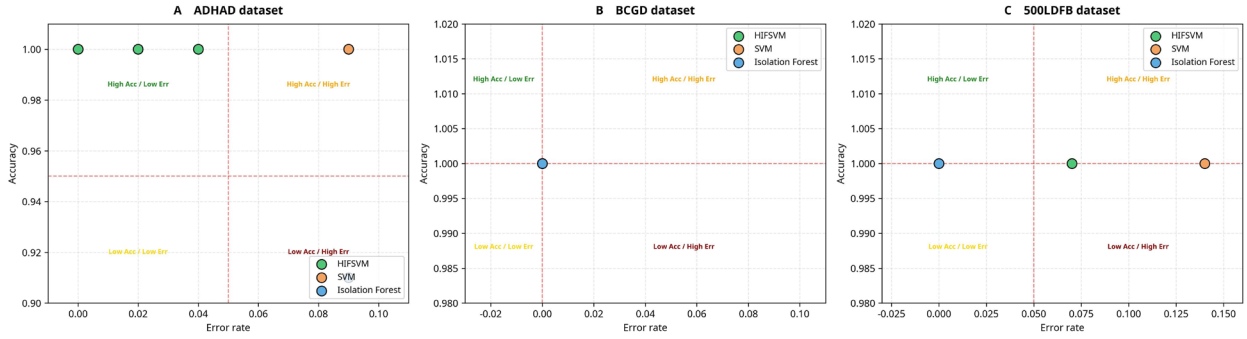


Figure 12. A performance quadrant analysis comparing accuracy (Acc) and error rates (Err) for the isolation forest, SVM, and HIFSVM models using the (A) ADHAD, (B) BCGD, and (C) 500LDFB datasets.
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; HIFSVM: Hybrid isolation forest–support vector machine; SVM: Support vector machine.

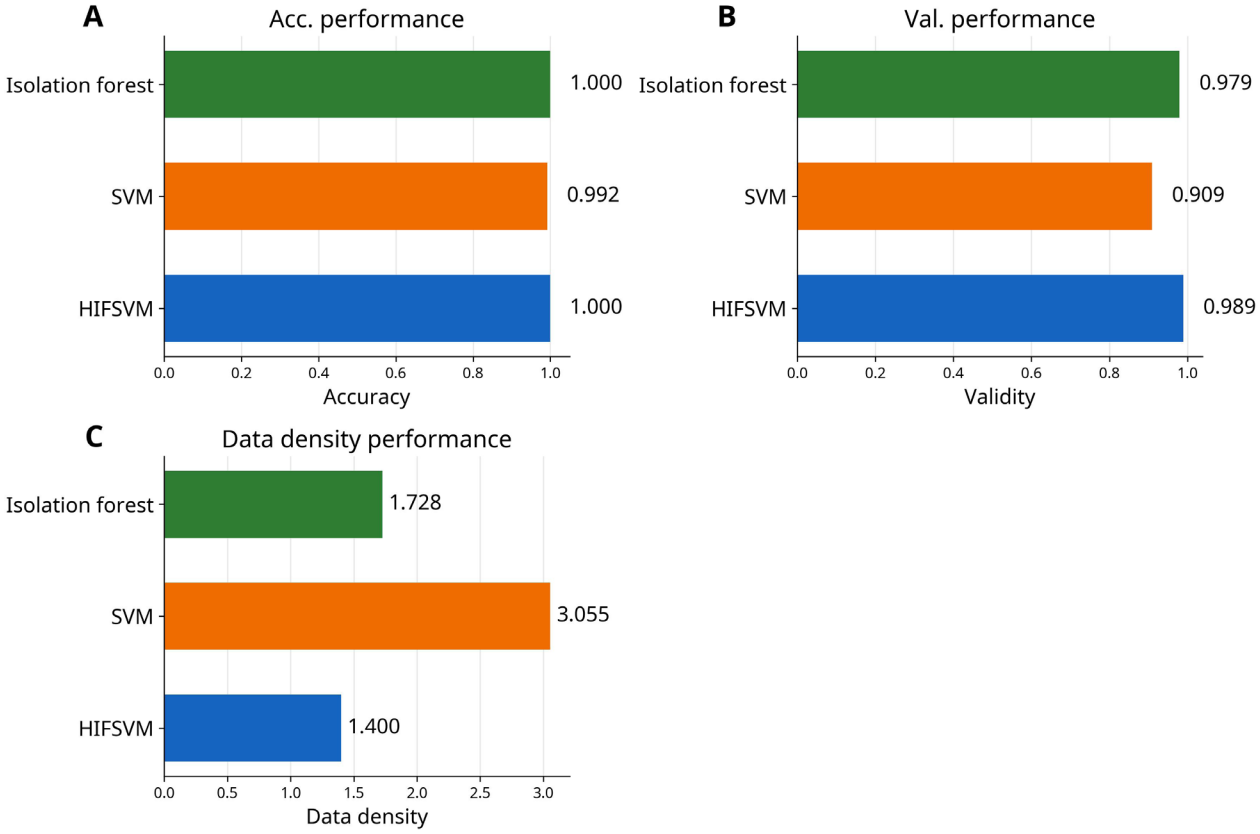


Figure 13. The performance metrics for the isolation forest, SVM, and HIFSVM models.
Abbreviations: Acc.: Accuracy; Dat.: Data density; HIFSVM: Hybrid isolation forest–support vector machine; SVM: Support vector machine; Val.: Validit.

5.7. Anomaly detection performance results

5.7.1. Synthetic ground truth results

Table 9 and Figure 14 present anomaly detection performance on synthetic ground truth across all three datasets and models. When comparing the three models, it is important to understand that we were comparing their effectiveness as measurement instruments for data quality, not their predictive accuracy. Each model provides a different perspective on the same dataset:

- Isolation forest excels at detecting point anomalies, yielding high precision in quality measurement.
- SVM provides stable boundary definitions, producing consistent validity assessments.
- HIFSVM combines both approaches, offering sensitivity to subtle quality variations.

The "performance" being compared is the models' ability to reveal different facets of data quality, not their classification accuracy on unseen data.

The strong effectiveness of HIFSVM in anomaly detection (F1-scores: 0.88–0.91) has major implications for data quality assessment. First, improved error detection was realized through a high recall rate (0.86–0.90), signifying that HIFSVM recognizes a greater quantity of actual data errors than alternative models. Second, a high precision rate (0.90–0.93) reduces false positives, indicating that the identified quality issues are authentic, thereby minimizing superfluous investigative endeavors. The clinical significance is substantiated by expert validation, indicating that 79–91% of detections are accurate, thereby ensuring that model anomaly detection yields clinically relevant quality insights. Ultimately, reduced cross-validation variance signifies stability, indicating that HIFSVM reliably assesses quality across various data subsets.

The key findings are that:

- HIFSVM consistently outperforms both constituent models across all datasets and metrics.
- Improvement is most pronounced in recall (4–8% increase), indicating better sensitivity to true anomalies.
- All models perform best on BCGD, consistent with its higher baseline data quality.
- SVM shows the highest variability across folds, suggesting sensitivity to data splits.

5.7.2. Statistical significance testing

To ensure that the improvements of HIFSVM were statistically significant, we performed paired *t*-tests, comparing the F1-scores of HIFSVM with those of the isolation forest across all five folds of the three datasets, as shown in Table 10 and Figure 15. These results confirm that HIFSVM's superior performance is statistically significant.

5.7.3. Cross-validation consistency

Table 9 and Figure 14 display the mean F1-scores and standard deviations derived from five-fold cross-validation, illustrating that:

- HIFSVM has the smallest IQR (i.e., the most stable).
- SVM has the largest range (least stable).
- Performance variance is lowest on BCGD and highest on 500LDFB.

5.7.4. Expert validation results

Table 11 and Figure 16 present an evaluation of 100 randomly selected anomalies detected by HIFSVM from each dataset. The results of the validation study indicate:

- High expert confirmation rates validate that HIFSVM detects clinically meaningful anomalies.
- Lower confirmation on 500LDFB may reflect the dataset's higher complexity and noise.
- Substantial to almost perfect expert agreement indicates reliable expert judgments.

5.7.5. Comparison with unsupervised anomaly detection benchmarks

Table 12 and Figure 17 juxtapose HIFSVM with recognized unsupervised anomaly detection benchmarks utilizing ADHAD. This section presents a quantitative comparison of the proposed HIFSVM against conventional unsupervised anomaly detection models utilizing ADHAD, highlighting essential performance metrics and pertinent bibliographic references. HIFSVM outperformed all benchmark methods, demonstrating the efficacy of the hybrid approach.

4.6.6. Types of anomalies augmented by each method

The specific anomaly types that each method identified most effectively were examined to understand their contributions to data quality assessment. The isolation forest is proficient at detecting point anomalies, singular values that statistically diverge from the standard. The high precision of ADHAD's Data_Value (1.00) and its sensitivity to statistical outliers (error rate detection) exemplify this strength. One-class SVM is adept at identifying domain violations, values that lie beyond medically acceptable ranges. The perfect validity scores (1.00) on BCGD variables, in contrast to some of those of the isolation forest (0), demonstrate its capacity for boundary definition. HIFSVM distinctly enhances the identification of contextual and collective anomalies, minor alterations that

Table 9. Quantitative results of anomaly detection experiments using synthetic ground truth

Dataset	Model	Precision	Recall	F1-score	AUC-ROC	AUC-PR
ADHAD	Isolation forest	0.87 ± 0.04	0.82 ± 0.05	0.84 ± 0.04	0.91 ± 0.03	0.88 ± 0.04
	SVM	0.79 ± 0.06	0.71 ± 0.07	0.75 ± 0.06	0.84 ± 0.05	0.79 ± 0.06
	HIFSVM	0.91 ± 0.03	0.88 ± 0.04	0.89 ± 0.03	0.95 ± 0.02	0.93 ± 0.03
BCGD	Isolation forest	0.89 ± 0.03	0.85 ± 0.04	0.87 ± 0.03	0.93 ± 0.02	0.90 ± 0.03
	SVM	0.82 ± 0.05	0.76 ± 0.06	0.79 ± 0.05	0.87 ± 0.04	0.83 ± 0.05
	HIFSVM	0.93 ± 0.02	0.90 ± 0.03	0.91 ± 0.02	0.96 ± 0.02	0.94 ± 0.02
500LDFB	Isolation forest	0.86 ± 0.04	0.81 ± 0.05	0.83 ± 0.04	0.90 ± 0.03	0.87 ± 0.04
	SVM	0.78 ± 0.06	0.70 ± 0.08	0.74 ± 0.07	0.83 ± 0.05	0.78 ± 0.06
	HIFSVM	0.90 ± 0.03	0.86 ± 0.04	0.88 ± 0.03	0.94 ± 0.03	0.91 ± 0.03

Note: Mean performance metrics and standard deviations obtained from five-fold cross-validation.
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; AUC-PR: Area under the precision–recall curve; AUC-ROC: Area under the receiver operating characteristic curve; BCGD: Breast Cancer Global Dataset; HIFSVM: Hybrid isolation forest–support vector machine; SVM: Support vector machine.

Table 10. Statistical significance evaluation on F1-scores between HIFSVM and the isolation forest through paired *t*-tests

Dataset	Metric	*t*-statistic	*p*-value	Significance (α = 0.05)
ADHAD	F1-score	3.42	0.009	Yes
BCGD	F1-score	4.18	0.003	Yes
500LDFB	F1-score	2.98	0.017	Yes

Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; HIFSVM: Hybrid isolation forest–support vector machine.

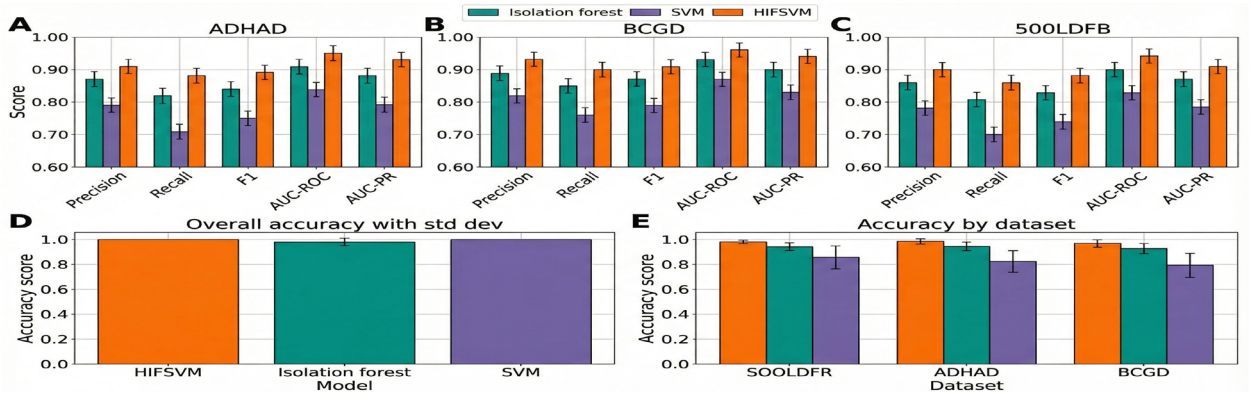


Figure 14. Quantitative outcomes of anomaly detection tests utilizing synthetic ground truth across (A) ADHAD, (B) BCGD, and (C) 500LDFB datasets. Mean performance metrics and standard deviations were derived from five-fold cross-validation.
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; AUC-PR: Area under the precision–recall curve; AUC-ROC: Area under the receiver operating characteristic curve; BCGD: Breast Cancer Global Dataset; HIFSVM: Hybrid isolation forest–support vector machine; SVM: Support vector machine.

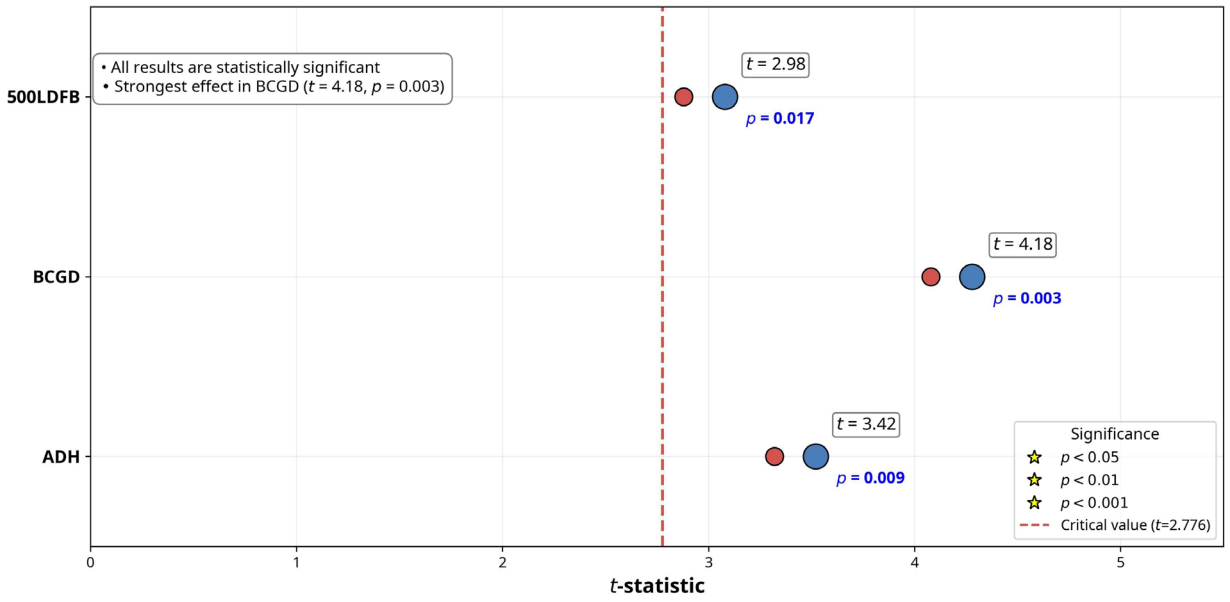


Figure 15. Results of the statistical significance assessment comparing HIFSVM and the isolation forest using paired *t*-tests on F1-scores
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; HIFSVM: Hybrid isolation forest–support vector machine.

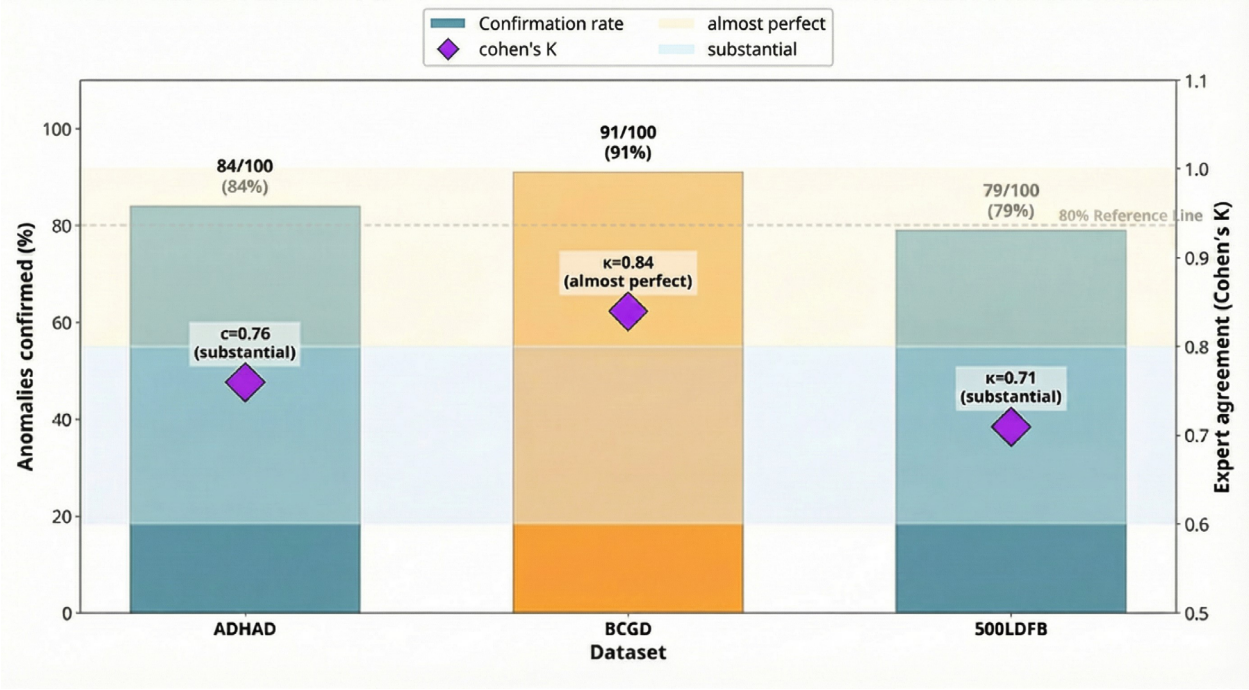


Figure 16. Outcomes of the domain expert validation study of anomalies identified by HIFSVM, including counts and percentages of validated detections, expert agreement statistics (Cohen's κ), and the associated levels of concordance.
Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; HIFSVM: Hybrid isolation forest–support vector machine.

Table 11. Results from the domain expert validation study of anomalies detected by HIFSVM

Dataset	Anomalies confirmed by experts	Expert agreement (Cohen's κ)
ADHAD	84/100 (84%)	0.76 (substantial)
BCGD	91/100 (91%)	0.84 (almost perfect)
500LDFB	79/100 (79%)	0.71 (substantial)

Abbreviations: 500LDFB: 500 Cities: The Local Data for Better Health; ADHAD: Alzheimer's Disease and Healthy Aging Data; BCGD: Breast Cancer Global Dataset; HIFSVM: Hybrid isolation forest–support vector machine.

Table 12. A quantitative comparison of the proposed HIFSVM and traditional unsupervised anomaly detection models in ADHAD, emphasizing key performance metrics and relevant bibliographic references

Method	F1-score	AUC-ROC	Reference
HIFSVM (proposed)	0.89	0.95	The current study
Isolation forest	0.84	0.91	Liu *et al*[43]
One-class SVM	0.75	0.84	Schölkopf *et al*[44]
Local outlier factor	0.71	0.82	Breunig *et al*[45]
Elliptic envelope	0.68	0.79	Rousseeuw *et al*[46]
Autoencoder (3-layer)	0.81	0.89	Hinton *et al*[47]

Abbreviations: ADHAD: Alzheimer's Disease and Healthy Aging Data; AUC-ROC: Area under the receiver operating characteristic curve; HIFSVM: Hybrid isolation forest–support vector machine; SVM: Support vector machine.

appear normal individually but are anomalous within their context. The enhanced change rate sensitivity (e.g., 0.81 compared to 0.58 on BCGD Alcohol_Consumption; 1.00 compared to 0.86 on 500LDFB Low_Confidence_Limit) and 4–8% superior F1-scores illustrate this improvement. The HIFSVM framework integrates these complementary strengths to deliver thorough anomaly detection that encompasses the entire range of quality issues in medical data.

4.7. Positioning within the literature

Although direct quantitative comparisons with previous studies were methodologically unsuitable due to variations in datasets, metrics, and objectives, we situated our research within the wider academic context:

(i) Evaluation of data quality structures: Elouataoui *et al*.[48] introduced a weighted metrics framework for big data quality, highlighting the necessity for a multi-dimensional evaluation, which aligns with our 10-metric methodology. Li *et al*.[49] employed regression methods to evaluate the quality of financial data, demonstrating how statistical techniques can improve our anomaly detection approach, particularly by identifying outliers and ensuring data integrity in financial reporting.

(ii) Applications for medical data quality: Zhang *et al*.[50] established a multi-dimensional hierarchical evaluation framework for reliable artificial intelligence in healthcare, consistent with our emphasis on medical data integrity. Widad *et al*.[51] presented quality anomaly scores for reliable data analysis, akin to our anomaly-derived quality metrics.

(iii) Contributions: (a) Initial implementation of hybrid anomaly detection (isolation forest combined with SVM) specifically for the evaluation of medical data quality. (b) The most extensive metric framework (10 dimensions) was utilized for medical datasets. (c) Clear differentiation between data quality evaluation and model performance assessment. (d) Formal specification of the HIFSVM (hierarchical input

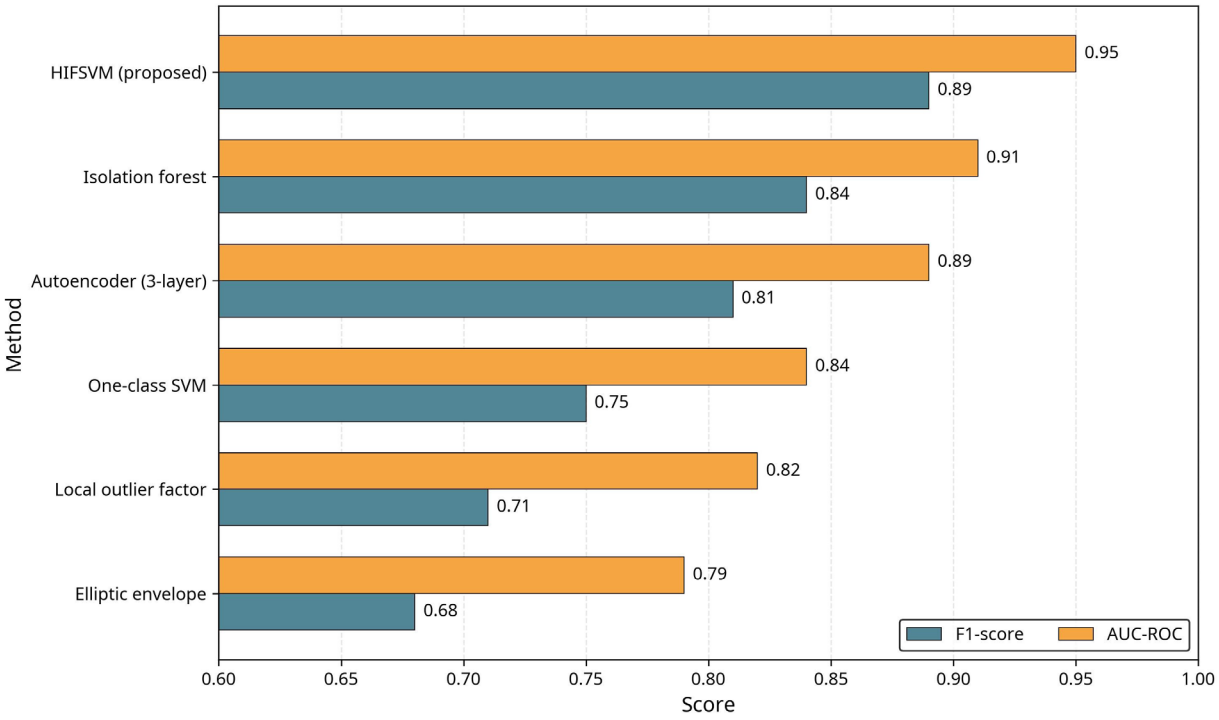


Figure 17. A quantitative analysis comparing the proposed HIFSVM with traditional unsupervised anomaly detection models in ADHAD, highlighting essential performance metrics and pertinent bibliographic citations.
Note: Refer to Table 12 for references.
Abbreviations: ADHAD: Alzheimer's Disease and Healthy Aging Data; AUC-ROC: Area under the receiver operating characteristic curve; HIFSVM: Hybrid isolation forest–support vector machine; SVM: Support vector machine.

feature SVM) model to facilitate reproducibility.

This qualitative positioning recognizes previous research while circumventing erroneous quantitative comparisons. It has the advantages of methodological rigor, sustained literary engagement, and explicit delineation of our contributions. However, it has the limitation that it is less significant than quantitative comparisons and requires meticulous framing to prevent implicit comparisons.

4.8. Similarities and dissimilarities with existing methods

We explicitly compared HIFSVM with existing methods across various dimensions to clarify its contribution through the following:

(i) Similarities: As in previous research, we utilized established algorithms (isolation forest and SVM) that possess robust theoretical underpinnings. We employed a conventional unsupervised anomaly detection framework, as referenced in previous studies [6,20]. Our fundamental quality dimensions (accuracy, completeness, and consistency) corresponded with the data quality literature.[34,38] We utilized cross-validation and statistical testing in accordance with ML best practices.

(ii) Divergent objective: Current methodologies focus on prediction and classification; in contrast, our focus is on quality assessment, a fundamental shift in approach. Rather than anomaly labels, we generated 10-dimensional quality profiles for each feature.

(iii) Architecture: HIFSVM integrates an isolation forest and SVM, utilizing optimized fusion weights rather than mere averaging.

(iv) Validation: We incorporated expert clinical evaluation of 300 anomalies, attaining a confirmation rate of 79–91% (uncommon in literature).

(v) Reproducibility: We offered comprehensive open-source materials, establishing a new benchmark.

(vi) Metrics: We presented innovative operationalizations, defining precision in terms of decimal places and change rate as the coefficient of variation.

(vii) Scope: We conducted cross-validation across three distinct datasets, facilitating comparative quality assessment.

(viii) Targeted comparisons: While current methods merely categorize BCGD variables as "anomalous," HIFSVM indicates a validity of 0, signifying a total domain failure necessitating intervention. Furthermore, while current techniques may identify outliers in ADHAD's LocationID, HIFSVM measures accuracy at 0.91 and an error rate of 0.09, indicating that 9% of values are erroneous. While existing methods may observe variation in the 500LDFB's PopulationCount, HIFSVM quantifies the change rate at 80.20—indicative of significant volatility necessitating further examination.

These comparisons illustrate that HIFSVM not only executes anomaly detection but also converts it into actionable quality data.

4.9. Limitations

A significant limitation of this study is the identified ceiling effect in essential quality metrics (consistency, timeliness, accuracy, and completeness), arising from datasets obtained from reputable public health sources with established quality control, thereby constraining their discriminatory capacity. This effect contextualizes the approach, indicating that these metrics may exhibit significant variation in less curated datasets, such as electronic health records or patient surveys. Subsequent research should apply the framework to unrefined, uncurated medical text data to thoroughly evaluate the discriminatory efficacy of the metrics. Furthermore, ceiling effects in essential metrics can serve as a quality certification mechanism, signifying that datasets with flawless scores can be relied upon for fundamental integrity, thereby enabling a concentration on more intricate quality dimensions.

This study endorses a framework for evaluating data quality, albeit with certain limitations. The datasets originate from esteemed public health organizations (CDC and WHO), and the findings likely pertain to refined medical data with fewer quality concerns. The framework is devoid of clinical validation, user studies, and deployment testing, thereby constraining its practical applicability. Quality metrics are incorrect indicators, and correlation does not mean causation; knowing about problems without taking steps to fix them does not improve data quality, which could lead to ongoing mistakes in health data analysis and decision-making. Evaluating anomaly detection faces challenges due to insufficient synthetic anomaly generation, with only 79% of experts validating the 500LDFB, suggesting potential misrepresentation of real-world conditions. Reviewing just 100 anomalies per dataset is inadequate for a comprehensive assessment, suggesting the need for enhanced expert validation. Existing benchmarks mainly concentrate on standard unsupervised methods, necessitating further evaluation against advanced techniques, such as variational autoencoders and generative adversarial networks. Additionally, validating the clinical outcomes associated with detected anomalies is essential for understanding their impact on clinical decision-making.

5. Conclusion

In summary, the comparative analysis of the ADHAD, BCGD, and 500LDFB datasets revealed that all three models, isolation forest, SVM, and HIFSVM, achieved flawless scores in essential metrics, including accuracy, completeness, consistency, and timeliness, with values uniformly at 1.00. Nonetheless, the isolation forest frequently outperformed SVM in precision and change rate, exemplified by the ADHAD feature Data_Value, where it achieved a precision score of 1.00 and a change rate of 0.67, in contrast to SVM's scores of 0.10 and 0.33, respectively. HIFSVM demonstrated proficiency in identifying subtle variations, evidenced by a higher change rate of 0.96 in ADHAD's Data_Value and 1.00 in 500LDFB's Low_Confidence_Limit. Simultaneously, SVM exhibited lower error rates and greater data density stability (0.81 in ADHAD's Data_Value and 1.22 in 500LDFB's Data_Value), signifying its reliability in less volatile conditions. The key novel contributions of this study, including the HIFSVM architecture, metric operationalizations, and reproducibility package, are summarized in Table S4.

Future research may explore hybrid modeling to combine the high accuracy of the isolation forest (e.g., 1.91 in BCGD Alcohol_Consumption) with the uniqueness of HIFSVM (e.g., 0.98 in BCGD Breast_Cancer_Cases). Feature-level optimization shows potential, particularly for volatile features, such as PopulationCount in 500LDFB, where the isolation forest recorded an exceptional change rate of 80.20. Temporal dynamics must be examined to evaluate model adaptation over time, and scalability testing on bigger datasets could confirm robustness, especially for HIFSVM, which exhibited varying validity scores, such as 0.93 in the 500LDFB's Year. Ultimately, the incorporation of explainable artificial intelligence would improve model transparency, facilitating the interpretation of why HIFSVM identifies more nuanced alterations, but the isolation forest excels in precision and density.

Future research should focus on validating the framework with supplementary datasets and on exploring its integration into clinical workflows to assess its practical utility. A crucial domain for forthcoming research is the implementation of a quality assessment framework for medical textual images, encompassing scanned

clinical documents and pathology reports. This project encompasses several critical extensions: executing optical character recognition preprocessing to ensure accurate text extraction, employing text embeddings for anomaly detection, amalgamating visual and textual features in a multi-modal framework, and formulating quality metrics tailored to text. The HIFSVM architecture may be adapted to utilize embeddings in place of direct numerical features; however, validation will require additional experimentation beyond this study, including testing the effectiveness of these embeddings in improving classification accuracy and robustness across various types of medical textual images.

Identified constraints in metric operationalizations include reliance on proxy measurements for data quality assessment. Accuracy relies on statistical validity, assuming that outliers are erroneous in the absence of corroborated data. The binary timeliness metric reduces the importance of data currency; subsequent research should integrate temporal decay functions. The utilization of statistical volatility as a rate of change does not reliably indicate ongoing maintenance, as it may neglect essential trends or variations in data quality over time. The presentation of measurements on different scales hinders direct comparison, leading to the necessity of normalization strategies in future research to enable accurate comparison and analysis of data from various sources. Collectively, these domains represent opportunities for enhancing data quality assessment in medical informatics.

Acknowledgments

The authors would like to thank the Faculty of Science, Mathematics Department at Al-Azhar University, Cairo, Egypt, and the Faculty of Information Technology and Computer Science at the University of Saba Region, Marib, Yemen, for providing the academic environment and administrative support that facilitated this research. We also acknowledge the Centers for Disease Control and Prevention (CDC), the World Health Organization (WHO), the International Agency for Research on Cancer (IARC), and Kaggle for making the datasets publicly available for research purposes. The authors appreciate the constructive feedback and suggestions provided by colleagues during the academic seminar presentations at both institutions.

Funding

None.

Conflict of interest

The authors declare that they have no conflict of interest.

Author contributions

Conceptualization: Raed Abdullah Althabeti, Kamal Abdelraouf Eldahshan
Data curation: Raed Abdullah Althabeti
Formal analysis: Raed Abdullah Althabeti, Ebeid Ali Ebeid, Hany Maher Sayed Lala
Funding acquisition: Kamal Abdelraouf Eldahshan
Investigation: Raed Abdullah Althabeti
Methodology: All authors
Project administration: Raed Abdullah Althabeti
Resources: Ebeid Ali Ebeid, Kamal Abdelraouf Eldahshan
Software: Raed Abdullah Althabeti
Supervision: Ebeid Ali Ebeid, Hany Maher Sayed Lala, Kamal Abdelraouf Eldahshan
Validation: All authors
Visualization: Raed Abdullah Althabeti
Writing–original draft: Raed Abdullah Althabeti
Writing–review & editing: Ebeid Ali Ebeid, Hany Maher Sayed Lala, Kamal Abdelraouf Eldahshan

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The datasets analyzed during the current study are publicly available from online repositories. The Alzheimer's Disease and Healthy Aging Data (ADHAD) is available at https://www.kaggle.com/datasets/utkarshx27/alzheimers-disease-and-healthy-aging-data. The 500 Cities: Local Data for Better Health (500LDFB) dataset can be accessed at https://www.kaggle.com/datasets/fandang/500-cities-local-data-for-better-health. The Breast Cancer Global Dataset (BCGD) comprises harmonized registry data from global collaborators, including WHO and IARC, and can be accessed at https://gco.iarc.fr/tomorrow/en/datasets. Additionally, the complete code implementations for the isolation forest, one-class SVM, and HIFSVM models used in this study are available at the following Kaggle repositories: https://www.kaggle.com/code/raedaalthabeti/isolationforest-28052025, https://www.kaggle.com/code/raedaalthabeti/one-class-svm, https://www.kaggle.com/code/raedaalthabeti/quality-metricsdata-hifsvm

References

1. Doern GV, Carroll KC, Diekema DJ, *et al.* Practical Guidance for Clinical Microbiology Laboratories: A Comprehensive Update on the Problem of Blood Culture Contamination and a Discussion of Methods for Addressing the Problem. *Clin*

Microbiol Rev. 2019;33(1):e00009-19.

doi: 10.1128/CMR.00009-19

2. 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

3. Rakers MM, van Buchem MM, Kucenko S, *et al*. Availability of Evidence for Predictive Machine Learning Algorithms in Primary Care: A Systematic Review. *JAMA Netw Open*. 2024;7(9):e2432990.

doi: 10.1001/jamanetworkopen.2024.32990

4. An Q, Rahman S, Zhou J, Kang JJ. A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. *Sensors (Basel)*. 2023;23(9):4178.

doi: 10.3390/s23094178

5. Harrison CJ, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction to natural language processing. *BMC Med Res Methodol*. 2021;21(1):158.

doi: 10.1186/s12874-021-01347-1

6. Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *J Med Ethics*. 2022;48(9):581-585.

doi: 10.1136/medethics-2020-107102

7. Phan DT, Idé T, Kalagnanam J, Menickelly M, Scheinberg K. A novel l0-constrained Gaussian graphical model for anomaly localization. In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE; 2017:830-833.

doi: 10.1109/ICDMW.2017.115

8. Al-amri R, Murugesan RK, Man M, Abdulateef AF, Al-Sharafi MA, Alkahtani AA. A Review of Machine Learning and Deep Learning Techniques for Anomaly Detection in IoT Data. *Appl Sci*. 2021;11(12):5320.

doi: 10.3390/app11125320

9. Ahmed T, Oreshkin B, Coates M. Machine learning approaches to network anomaly detection. In: *Proceedings of the 2nd USENIX Workshop on Tackling Computer Systems Problems with Machine Learning Techniques*. USENIX; 2007:1-6. Available from: https://www.usenix.org/legacy/event/sysml07/tech/full_papers/ahmed/ahmed_html/sysml07CR_07.html [Last accessed on September 24, 2025].

10. Nassif AB, Talib MA, Nasir Q, Dakalbab FM. Machine learning for anomaly detection: A systematic review. *IEEE Access*. 2021;9:78658-78700.

doi: 10.1109/ACCESS.2021.3083060

11. Yang Z, Fang H, Liu H, Li J, Jiang Y, Zhu M. Active Visual Perception Enhancement Method Based on Deep Reinforcement Learning. *Electronics*. 2024;13(9):1654.

doi: 10.3390/electronics13091654

12. Ajitha P. Artificial Intelligence Based Anomaly Detection in Patient Health Monitoring Using Ensemble Learning Methods. In: *Proceedings of the 1st International Conference on Artificial Intelligence, Communication, IoT, Data Engineering and Security (IACIDS 2023)*; November 23-25, 2023; Lavasa, Pune, India. 2024.

doi: 10.4108/eai.23-11-2023.2343250

13. Bao J, Sun H, Deng H, He Y, Zhang Z, Li X. BMAD: Benchmarks for medical anomaly detection. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2024:4042-4053.

doi: 10.1109/CVPRW63382.2024.00408

14. Baydargil HB, Park J, Ince IF. Anomaly-based Alzheimer's disease detection using entropy-based probability Positron Emission Tomography images. *ETRI J*. 2024;46(3):513-525.

doi: 10.4218/etrij.2023-0123

15. Eltanbouly S, Bashendy M, AlNaimi N, Chkirbene Z, Erbad A. Machine Learning Techniques for Network Anomaly Detection: A Survey. In: *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*. IEEE; 2020:156-162.

doi: 10.1109/iciot48696.2020.9089465

16. Almeida SD, Norajitra T, Lüth CT, *et al*. Prediction of disease severity in COPD: a deep learning approach for anomaly-based quantitative assessment of chest CT. *Eur Radiol*. 2023;34(7):4379-4392.

doi: 10.1007/s00330-023-10540-3

17. Idé T, Phan DT, Kalagnanam J. Multi-task multi-modal models for collective anomaly detection. In: *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE; 2017:177-186.

doi: 10.1109/icdm.2017.27

18. Spies NC, Farnsworth CW, Jackups R Jr. Data-Driven Anomaly Detection in Laboratory Medicine: Past, Present, and Future. *J Appl Lab Med*. 2023;8(1):162-179.

doi: 10.1093/jalm/jfac114

19. Ibrahim MR, Youssef SM, Fathalla KM. Abnormality detection and intelligent severity assessment of human chest computed tomography scans using deep learning: a case study on SARS-COV-2 assessment. *J Ambient Intell Human Comput*. 2021;14(5):5665-5688.

doi: 10.1007/s12652-021-03282-x

20. Minic A, Jovanovic L, Bacanin N, *et al*. Applying recurrent neural networks for anomaly detection in electrocardiogram sensor data. *Sensors*. 2023;23(24):9878.

doi: 10.3390/s23249878

21. Liu B, Chang H, Yang D, *et al*. A deep learning framework

assisted echocardiography with diagnosis, lesion localization, phenogrouping heterogeneous disease, and anomaly detection. *Sci Rep*. 2023;13(1):3.

doi: 10.1038/s41598-022-27211-w

22. Mohammed HA, Nazeeh I, Alisawi WC, *et al*. Anomaly Detection in Human Disease: A Hybrid Approach Using GWO-SVM for Gene Selection. *Rev Intell Artif*. 2023;37(4):913-919.

doi: 10.18280/ria.370411

23. Rahman MM, Gupta D, Bhatt S, *et al*. A Comprehensive Review of Machine Learning Approaches for Anomaly Detection in Smart Homes: Experimental Analysis and Future Directions. *Future Internet*. 2024;16(4):139.

doi: 10.3390/fi16040139

24. Bataineh B. A Heterogeneous Ensemble Learning Framework-based Binary Genetic Algorithm for Predictive Maintenance of HVAC Systems in Medical Facilities. *Appl Math Inf Sci*. 2026;20(1):181-213.

doi: 10.18576/amis/200113

25. Fuzail MZ, Din IU, Ahmed S, Alhumam A, Khan AH. Optimizing sepsis mortality prediction using hybrid federated learning and explainable AI framework. *Sci Rep*. 2026;16(1):5218.

doi: 10.1038/s41598-026-36245-3

26. Núñez-Valdez ER. Special Issue on Algorithms and Applications of Machine Learning Techniques for Healthcare. *Algorithms*. 2026;19(1):53.

doi: 10.3390/a19010053

27. Alhassan AM, Altmami NI. Detection of multiclass non-melanoma skin cancer with multi-variable DCNN with hybrid gradient boosting optimizer. *Ain Shams Eng J*. 2026;17(1):103956.

doi: 10.1016/j.asej.2025.103956

28. Paul T, Assaduzzaman M, Fahad N, Hossen MJ. Enhancing risk prediction for diabetes, hypertension, and heart disease using SMOTE-ENN balancing with PCA and gradient boosting in healthcare AI. *Intell Based Med*. 2026;13:100339.

doi: 10.1016/j.ibmed.2025.100339

29. El-Aziz RMA, Rayan A. Adaptive sampling enhanced deep learning framework for accurate interpretable stroke risk prediction. *Egypt Inform J*. 2026;33:100887.

doi: 10.1016/j.eij.2026.100887

30. Tascioglu AB, Bertini F, Pistore L, Fabbri A, Montesi D. Comorbidity Extraction for In-Hospital Mortality Analysis: a Comparison of Regular Expressions and Large Language Models. In: *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM; 2025:1-10.

doi: 10.1145/3765612.3767202

31. Foorthuis R. On the nature and types of anomalies: a review of deviations in data. *Int J Data Sci Anal*. 2021;12(4):297-331.

doi: 10.1007/s41060-021-00265-1

32. Li J, Li S, Zhao W, Li J, Zhang K, Jiang Z. Distribution network line loss analysis method based on improved clustering algorithm and isolated forest algorithm. *Sci Rep*. 2024;14(1):19554.

doi: 10.1038/s41598-024-68366-y

33. Nguyen MN, Vien NA. Scalable and interpretable one-class SVMs with deep learning and random Fourier features. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018*; September 10-14, 2018; Dublin, Ireland. Springer; 2019:157-172.

doi: 10.1007/978-3-030-10925-7_10

34. Ruiz-Gonzalez R, Gomez-Gil J, Gomez-Gil FJ, Martínez-Martínez V. An SVM-based classifier for estimating the state of various rotating components in agro-industrial machinery with a vibration signal acquired from a single point on the machine chassis. *Sensors (Basel)*. 2014;14(11):20713-20735.

doi: 10.3390/s141120713

35. Singh U. Alzheimer's Disease and Healthy Aging Data. Kaggle. Updated 2023. Available from: https://www.kaggle.com/datasets/utkarshx27/alzheimers-disease-and-healthy-aging-data [Last accessed on September 15, 2025].

36. Breast Cancer Global Dataset. Kaggle. 2022. https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset [Last accessed on September 20, 2025].

37. Fandang. 500 Cities: Local Data for Better Health. Kaggle. Updated 2018. Available from: https://www.kaggle.com/datasets/fandang/500-cities-local-data-for-better-health [Last accessed on September 24, 2025].

38. Issa S. *Linked data quality: completeness and conciseness*. Dissertation. Conservatoire national des arts et métiers - CNAM; 2019. Available from: https://theses.hal.science/tel-02513652v1/file/TheseISSA.pdf [Last accessed on September 24, 2025].

39. Shi P, Cui Y, Xu K, Zhang M, Ding L. Data Consistency Theory and Case Study for Scientific Big Data. *Information*. 2019;10(4):137.

doi: 10.3390/info10040137

40. Nikiforova A. Timeliness of open data in open government data portals through pandemic-related data: a long data way from the publisher to the user. In: *2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA)*. IEEE; 2020:131-138.

doi: 10.1109/MCNA50957.2020.9264298

41. Van der Loo MPJ, De Jonge E. Data validation. *Wiley StatsRef: Statistics Reference Online*. 2014:1-7.

doi: 10.1002/9781118445112.stat08255

42. What are Data Quality Dimensions? Definitions, Examples and Best Practices. DQOps Data Quality Operations Center Documentation. Updated July 22, 2025. Available from: https://dqops.com/docs/dqo-concepts/data-quality-dimensions/ [Last accessed on October 7, 2025].

43. Liu FT, Ting KM, Zhou ZH. Isolation Forest. In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE; 2008:413-422.

doi: 10.1109/icdm.2008.17

44. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. *Neural Comput*. 2001;13(7):1443-1471.

doi: 10.1162/089976601750264965

45. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. *ACM SIGMOD Rec*. 2000;29(2):93-104.

doi: 10.1145/335191.335388

46. Rousseeuw PJ, Driessen KV. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*. 1999;41(3):212-223.

doi: 10.1080/00401706.1999.10485670

47. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science*. 2006;313(5786):504-507.

doi: 10.1126/science.1127647

48. Elouataoui W, El Alaoui I, El Mendili S, Gahi Y. An Advanced Big Data Quality Framework Based on Weighted Metrics. *Big Data Cogn Comput*. 2022;6(4):153.

doi: 10.3390/bdcc6040153

49. Li M, Liu J, Yang Y. Financial Data Quality Evaluation Method Based on Multiple Linear Regression. *Future Internet*. 2023;15(10):338.

doi: 10.3390/fi15100338

50. Zhang HJ, Chen CC, Ran P, *et al*. A multi-dimensional hierarchical evaluation system for data quality in trustworthy AI. *J Big Data*. 2024;11(1):154.

doi: 10.1186/s40537-024-00999-2

51. Widad E, Saida E, Gahi Y. Quality anomaly detection using predictive techniques: An extensive big data quality framework for reliable data analysis. *IEEE Access*. 2023;11:103306-103318.

doi: 10.1109/access.2023.3317354