

ARTICLE

Genetic instance-based learner algorithm for lymphography cancer diagnosis

Hayder Naser Khraibet Al-Behadili^{1†*}, and Ayad Mohammed Jabbar^{2†}¹Department of Computer Science, Shatt Al-Arab University College, Basra, Iraq²Department of Translation, College of Arts, University of Basrah, Basrah, Iraq

Abstract

Lymphography involvement is a fundamental indicator of cancer progression, making the precise classification of lymphographic findings essential for reliable diagnosis and treatment decision-making. This study presents an embedded approach for lymphography cancer detection that synergistically combines a genetic algorithm (GA) and an instance-based learning classifier. Unlike prior GA hybrids, this approach incorporates GA feature selection with K*, a relative-entropy-based approach that handles noisy, small-sample-size, class-imbalance, and redundant features in lymphography data, rather than the traditional hybrid approach that used KNN with a geometric distance metric (typically Euclidean or Manhattan). Although vital for diagnosing lymphatic diseases, lymphography poses challenges because of the data it generates. To address these challenges, the proposed approach integrates the strengths of GA to optimize the most important features. Instance-based learning (i.e., K* classifier) is used for effective pattern recognition and to classify lymph nodes as benign, normal, metastatic, malignant, or fibrotic. Extensive experimental results on a real-world dataset demonstrated that the combination significantly outperformed state-of-the-art classification algorithms and other hybrid approaches. The efficiency of the proposed approach is evaluated using the most widely used computational metrics, yielding promising results for diagnosing lymphatic diseases. This work highlights the potential of an evolutionary-based embedded classifier for classifying complex biomedical data.

[†]These authors contributed equally to this work.

***Corresponding author:**Hayder Naser Khraibet Al-Behadili
(hayderkhraibet@sa-uc.edu.iq)

Citation: Al-Behadili, H. N. K. & Jabbar, A. M. (2026). Genetic instance-based learner algorithm for lymphography cancer diagnosis. *Int J Systematic Innovation*, 10(4): 026080014.
[https://doi.org/10.6977/IJoSI.202608_10\(4\).0005](https://doi.org/10.6977/IJoSI.202608_10(4).0005)

Received: February 20, 2026**Revised:** June 20, 2026**Accepted:** July 7, 2026**Published online:** July 24, 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.

Keywords: Cancer diagnosis; Data mining; Lymphography; Metaheuristic; Parameter tuning

1. Introduction

Cancer is a broad term for a complex group of diseases characterized by the uncontrolled growth and extension of abnormal cells. It can develop in any organ or tissue when normal cell regulation is disrupted. As a result, cancer may spread to nearby tissues or invade other parts of the body via the lymphatic and blood systems (Cheng & Shi, 2025). Accurate classification of lymphatic disease is central to oncological diagnosis and treatment planning, as the lymphatic system is a principal route for metastatic disease dissemination (Lei *et al.*, 2024). In addition, lymphography helps clinicians distinguish between normal, fibrotic, metastatic, benign, and malignant states, each with distinct therapeutic and prognostic implications (Negm *et al.*, 2024). As a consequence, accurate,

reproducible classification of these states facilitates early diagnosis and has significant clinical value, motivating the development of intelligent decision-support systems. Although it holds considerable clinical importance, the lymphographic data present pose serious obstacles for automated classification.

The dataset examined in this study comprises 148 instances, each characterized by 18 predictive attributes and a 5-class outcome variable (Valeriano *et al.*, 2026). The majority of attributes are nominal, thereby increasing the complexity of data classification and extraction. The major challenge arises from the pronounced class distribution (i.e., class imbalance) in the data. Precisely, the metastases and malignant lymphoma classes comprise 81 and 61 instances, respectively (96% of the instances). In contrast, the two minority classes, fibrosis and normal, are represented by only four and two samples, respectively. Such an uneven distribution can bias classification models toward the majority classes, while the clinically important minority is ignored entirely (Yang *et al.*, 2024).

In this context, a variety of machine learning and computational approaches have been employed to support cancer visualization and diagnostic assessment. Lymphography is a diagnostic test that uses an imaging technique to visualize the lymphatic system by injecting a special contrast dye. Supervised learning (SL), specifically classification, is the most widely applied machine learning approach in clinical cancer diagnosis because it can predict known outcomes from labeled datasets. The most common classification models used in this field are k-nearest neighbors (KNNs), support vector machines (SVMs), logistic regression (LR), random forest (RFs), genetic algorithms (GAs), decision trees (DTs), naïve Bayes (NB), and artificial neural networks (ANNs). They learn from different types of data, such as clinical records, biopsy results, genomic data, and medical images, where the diagnosis is already known (e.g., medical images labeled as “presence” or “absence”).

Despite their widespread adoption, conventional SL classifiers remain particularly sensitive to the characteristics and inherent challenges exhibited by the lymphography dataset. Distance-based classification methods rely on similarity measures, such as the Manhattan and Euclidean distances, which assume meaningful numerical relationships among attribute values. This assumption is often poorly suited to datasets dominated by nominal features, since such metrics depend on interpretable numerical distances between values that categorical data do not support. Thus, the categorical type lacks inherent distance relationships or numerical ordering, thereby limiting the effectiveness in distinguishing clinically

important fibrosis from normal cases, which constitute only 4% of the instances and are therefore easily crushed.

Given these challenges, feature selection therefore plays a crucial role in medical diagnosis. It is considered a critical preprocessing step that eliminates redundant and irrelevant attributes and reduces dimensionality. Effective feature selection helps clinicians and SL models focus on key disease indicators. In this context, researchers have increasingly turned to metaheuristic optimization techniques as powerful tools for addressing feature-selection problems. A metaheuristic is a high-level, problem-solving strategy designed to find near-optimal solutions in complex optimization problems within a reasonable time. Metaheuristics can produce accurate, easy-to-interpret models that enable rapid diagnosis. Metaheuristic algorithms are usually inspired by natural phenomena, social behaviors, or physical characteristics. Famous algorithms include ant colony optimization, particle swarm optimization, simulated annealing, and GAs.

Among the various metaheuristic optimization methods, GAs have emerged as one of the most widely adopted approaches in solving complex optimization problems. Specifically, GAs are a class of metaheuristic search-based optimization techniques inspired by the principles of genetics and natural evolution. They mimic the natural evolutionary process of “survival of the fittest” by producing candidate solutions—called chromosomes—that evolve over generations. Exploration and exploitation are two complementary search mechanisms used in GAs. Achieving the right balance between these two terms is important for finding high-quality solutions and reaching a global optimum without premature convergence. In addition, many GA-based hybrid approaches have been introduced, with classifiers that remain critical for noisy, small sample sizes, class imbalance, and redundant features. Consequently, although such hybrid approaches may improve feature representation, they commonly fail to handle the fundamental challenges in data classification (Thambi *et al.*, 2025).

Instance-based learning is a machine learning algorithm that learns by directly comparing new data points to the most similar instances using an entropy-based distance metric rather than building a model (Raza *et al.*, 2025). Embedding two algorithms could yield a classification model more powerful than either algorithm alone. This work aims to leverage the strengths of multiple algorithms to improve overall performance. Classification accuracy and feature selection are the main challenges in lymphography-based cancer diagnosis using these metaheuristic algorithms. To provide clarity and ease the

proposed study, the remainder of the paper is structured as follows: Section 2 presents a review of related work on existing hybrid approaches for lymphatic classification, instance-based learning, and feature selection methods. In Section 3, the methods and materials of GA formulation, the K* classifier, and their integration are described. Section 4 describes the experimental setup, including the dataset, evaluation metrics, and experimental study, and discusses the results, including comparisons against hybrid approaches and state-of-the-art methods. Finally, Section 5 concludes the paper and presents future research directions.

2. Related works

Explainable SL has been examined in various studies to improve disease identification in medical assessment. SL is a fundamental machine learning technique that trains algorithms on labeled datasets. SL teaches these algorithms to map previous illness records and adjust their parameters to reduce errors over time. The algorithm identifies relationships to predict (i.e., detect) new, unseen cancer cases. Research has focused on SL algorithms. Valeriano *et al.* (2026) proposed a hybrid approach for medical consultations that leverages instance hardness to filter instances during training and utilizes a confidence-based rejection mechanism during learning, ensuring that only correct predictions are retained. Abdolrazzaghe-Nezhad and Izadpanah (2025) proposed a hybrid cancer detection approach using fuzzy logic and the fuzzy cuckoo optimization algorithm to tune the parameters and identify the optimal scales.

Feature selection is the process of identifying and selecting the most important and relevant input features from a dataset. It aims to reduce noise and computational costs whilst improving model accuracy. Some famous feature selection methods include filter, wrapper, embedded, and combined approaches. Filter approaches use statistical measures such as chi-square and correlation. Wrapper approaches select a subset of features based on the model's performance. Embedded techniques perform important features during model training (e.g., DT and regression) (Song, Wang, Qi, Wang, Li, & Song, 2025).

Zhang *et al.* (2025) introduced a structural health monitoring model based on filter feature selection. The proposed method uses the sum of squared canonical correlation coefficients between the monitored features and the prediction classes, as tested in a real-world domain and a secondary dataset. Tabasi *et al.* (2025) proposed a deep learning method with feature selection based on minimum-redundancy-maximum-relevance. Early cancer detection was achieved by analyzing gene expression and identifying

distinct molecular signatures (most informative features) across different organs. Meanwhile, a hybrid filter-genetic feature selection method was applied to a noisy, high-dimensional real-world microarray dataset; this method used Chi-squared, information gain ratio, and information gain to identify the most cancerous features (Ali & Saeed, 2023).

Another study focused on liver cancer, which is considered a heterogeneous complex disease and the leading cause of cancer death in the world and presented a stacking ensemble learning model with feature selection techniques for gene expression data (Mahmoud & Takaoka, 2025). This stacking ensemble learning model included KNN, multilayer perceptron, RF, SVM, and extreme gradient boosting. A previous work explored a gene expression framework for Alzheimer's disease and cancer diagnosis using gene expression data (Babichev *et al.*, 2025). The framework applied gene ontology analysis using a self-organizing tree algorithm and cluster-biclust analysis using an ensemble-based approach and convolutional ANN. Researchers also introduced a medical diagnostic system with three stages: outlier removal, feature selection, and chronic disease classification (Rabie *et al.*, 2025). Outliers were removed using GA, features were selected using a hybrid feature selection method, and an NB classifier was used to provide accurate diagnoses. Meanwhile, feature selection was applied to retrieve significant text from medicine packaging using the Rapid Automatic Keyword Extraction (RAKE) algorithm (Saqib *et al.*, 2025). Streamlined optical character recognition and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scoring were adopted to interpret dense text and obtain reliable medical information. In another study (Abdollahi & Nouri-Moghaddam, 2022), an ensemble-based Gas approach was applied for feature selection to predict and diagnose diabetes mellitus.

Machine learning and artificial intelligence methods have increasingly been applied across a range of oncological cancer diagnostic applications. In medical imaging, deep learning architectures have demonstrated remarkable improvements. J. Zhu *et al.* (2024) introduced an improved U-Net architecture for brain tumor segmentation from magnetic resonance imaging (MRI), enhancing feature representation to more accurately delineate malignant tissue. Similarly, Islam *et al.* (2024) developed a convolutional neural network framework for the classification of metastatic breast cancer and invasive ductal carcinoma from histopathological images, demonstrating the value of deep learning for early breast-cancer diagnosis. In addition, comparative investigations have examined the effectiveness of various machine-

learning methods for lung cancer prediction, evaluating their performance on clinical data using metrics such as precision, recall, and F1 scores (Maurya *et al.*, 2024).

These studies collectively highlight the increasing reliance on data-driven approaches for cancer diagnosis and provide a strong foundation for the research direction pursued in the present work. Nevertheless, the existing approaches differ from our work in both methodological design and data modality, whereas the above approaches have primarily focused on medical imaging data and deep learning architectures. In contrast, the present work addresses structured, tabular lymphographic data characterized by a limited sample size and pronounced class imbalance. In this setting, a GA-driven feature-selection strategy is combined with the K* instance-based classifier, providing an interpretable, computationally efficient, and data-efficient complementary alternative to the imaging-focused paradigms reported in the recent literature.

C. Zhu *et al.* (2025) introduced a wrapper approach that utilizes the red-billed blue magpie algorithm for feature selection on medical data. In their study, an elite search method, namely, collaborative hunting, was used to identify key features. Memory storage was also utilized.

Diabetes prediction remains a major area of research. Machine learning models have been proposed for its diagnosis. A previous study applied particle swarm optimization as a wrapper for feature selection, then used RF, NB, and DT to classify and predict diabetes (Abdollahi & Aref, 2024).

Other studies combined filter and wrapper feature selection approaches. The ReliefF method was proposed as a preprocessing step for dimensional reduction (Yavuz *et al.*, 2025). A wrapper approach that stochastically selects one of two optimization algorithms, namely, JAYA and sine-cosine algorithms, was then used.) proposed a hybrid filter-and-wrapper feature selection approach for medical applications. They combined the binary cheetah optimizer algorithm as the wrapper approach and fuzzy joint mutual information as the filter method. Song, Wang, Qi, Wang, Song, and Shang-Guan (2025) introduced a hybrid approach to address the imbalance in cancer gene expression data. This approach involved two distinct stages: the lasso regression algorithm was employed to establish a threshold as a filter approach to identify a subset of features with a weight order. In the second stage, backward and forward sequential selection were used to determine the subset weight. The two sets of features were then combined and subjected to another round of selection using the two implemented strategies.

Embedded methods directly integrate selection into

the training process of SL models. It learns the machine learning model and selected features directly. Amini and Hu (2021) examined embedded methods and applied a hybrid, two-layer feature selection approach. The first layer included GA to search for an optimal subset of features. The second layer employed the elastic net method to further eliminate irrelevant and redundant features. Similarly, another study explored optimal subset selection using hybrid GAs with embedded regularization methods (Liu *et al.*, 2018). The GA was used for global search, whilst embedded regularization methods were applied for local refinement.

The disadvantage of filter approaches is that they independently evaluate each feature, ignoring its contribution relative to the group. Thus, these features may perform poorly on their own but exhibit strong performance when used as part of a subset of features. The performance of wrapper approaches in classification accuracy is superior to that of filter methods, but the computational cost remains the main obstacle in feature selection processing. Additionally, combined methods have the following drawbacks: they involve filter measures, independently examine features, do not work well all the time, and always require examining features as a whole. Embedded feature selection methods play a crucial role by integrating a learning model and automatic feature selection into one process. In these methods, features are selected during the learning process and directly coded in the construction vector as extra dimensions. Therefore, the use of embedded methods for feature selection has become popular over time.

This section provides a comprehensive review of feature selection methods used in the medical domain in real-world settings. Common SL and feature selection methods are explained, with a focus on the medical domain. In the next section, the preparation of medical data and the building of the embedded approach are described. Additional explanation tools are given.

3. Methods and materials

This section outlines the roadmap for this study and its three sequential stages: building the model, describing the explanatory tools, and preparing the medical data. In this section, the proposed genetic instance-based learner algorithm is explained by integrating the two aspects. In the first stage, the main objective of this integration is to identify the best feature subset and improve classification accuracy. In the proposed genetic instance-based learner algorithm, the main components are as follows: population initialization, fitness function calculation, selection, crossover, mutation, population list update, stopping

criteria check, classification model construction, and model quality calculation (Figure 1).

The process starts by creating the initial population by introducing a set of individuals (i.e., a subset of features). Each individual has a set of genes (data attributes). In the feature selection stage, two dynamic array-lists are created. The first one is a list of probabilities for each subset of feature values $\mathbb{L}_n$, denoted as probability subset: FeaturesList ($\mathbb{L}_f$) = { $p(\mathbb{L}_f1)$, $p(\mathbb{L}_f2)$, $p(\mathbb{L}_f3)$, ..., $p(\mathbb{L}_fn)$ }, which detects the probability value of each subset of features. The second dynamic array-list collects the quality of each subset value, indicated as QualityList ($\mathbb{L}_q$) = { $q(\mathbb{L}_q1)$, $q(\mathbb{L}_q2)$, ..., $q(\mathbb{L}_qn)$ }. When each value $\mathbb{L}_qn$ is used, its determined quality $q(\mathbb{L}_qn)$ is calculated.

The algorithm creates an initial population of size $10 \times$ feature numbers (Table 1). The initial population has a huge influence on the quality of the selected features. Unlike conventional integrated approaches that rely on geometric distance metrics (e.g., Euclidean or Manhattan), the proposed approach hybridizes GA feature selection with the K* classifier, which incorporates a built-in relative-entropy-based approach for handling incomplete, noisy, and mixed-type lymphography data. Thus, the relative entropic measure is adopted as a heuristic function (Equation 1) for the initial population of the proposed algorithm. This function measures the difference between two probability distributions using entropy and information theory for each feature and the class outcome.

$$D(P||Q) = \sum_{i=1}^n P_i \log \left(\frac{p_i}{q_i} \right) \quad (1)$$

Table 1. Parameters of a genetic-instance-based learner

Parameter	Value
Population size	$10 \times$ feature numbers
Maximum number of generations	300
Gene length	Number of features
Crossover rate	0.8
Mutation rate	$1/N$, where N = attributes number

where both p_i and q_i denote the probabilities assigned by P and Q to outcome i , respectively. Within the proposed initialization heuristic, the two distributions are defined over the class variable Y , which takes the four class values of the diagnostic lymphography dataset. Therefore, Equation 2 calculates the informative feature X about class Y in the multi-class classification:

$$Y \in \{0, 1, 2, 3\} \quad (2)$$

where normal = 0, metastases = 1, malignant lymph = 2, and fibrosis = 3.

For nominal candidate feature X that takes possible values $x \in V(X)$, let P correspond to the class distribution conditioned on a specific feature value x , whereas Q indicates the true class distribution in a given feature:

$$P = P(Y | X = x) \quad (3)$$

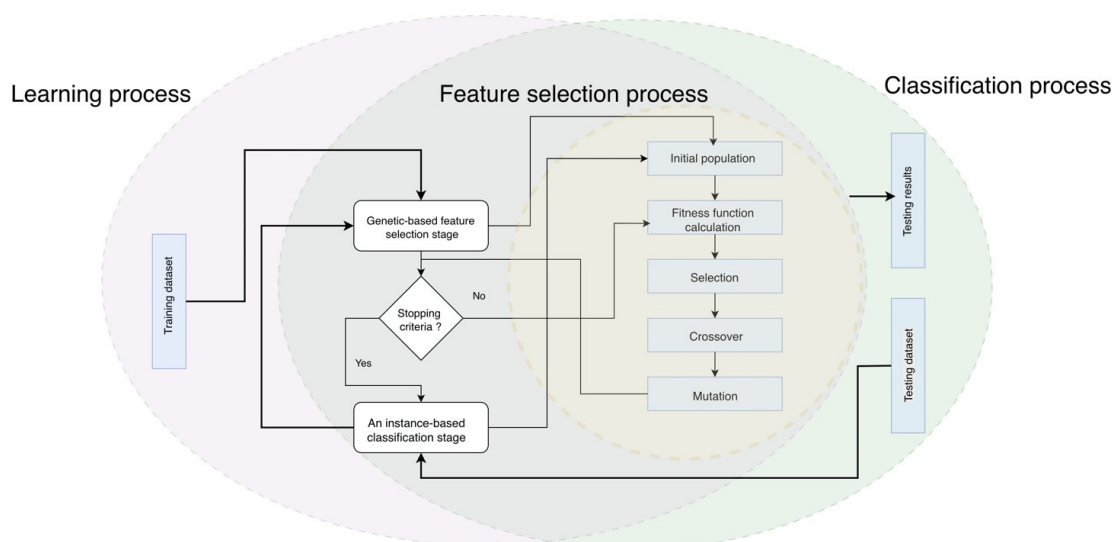


Figure 1. Proposed model architecture for a genetic-instance-based learner algorithm

$$Q = P(Y) \quad (4)$$

The Kullback–Leibler divergence $D(P(Y | X = x) \parallel P(Y))$ measures how much observation of a specific feature value $X = x$ alters the class distribution compared with the overall class. To assess a single relevance score for a feature (X), these per-value divergences are aggregated through a weighted average over the feature's values, weighted by their frequency:

$$\text{Rel}(X) = \sum_{\{x \in V(X)\}} P(X = x) \cdot D(P(Y | X = x) \parallel P(Y)) \quad (5)$$

Therefore:

$$\text{Rel}(X) = \sum_{\{x \in V(X)\}} P(X = x) \cdot \sum_{\{y \in Y\}} P(Y = y | X = x) \cdot \log \left(\frac{P(Y = y | X = x)}{P(Y = y)} \right) \quad (6)$$

A feature with a high $\text{Rel}(X)$ indicates that it provides more information about the diagnostic class variable. Therefore, it increases the probability of inclusion in the randomly generated initial population (i.e., chromosomes). The genetic search is subsequently driven by the built-in fitness function, which evaluates each candidate feature subset using the K* classifier.

The tournament procedure randomly selects k parents from the population. k is set as the number of features (e.g., assuming a dataset with 10 features, $k = 10$). The fitness quality of each chromosome is evaluated using the pure-accuracy equation (Equation 7), which is defined as the classification accuracy obtained by the K* classifier when restricted to the selected feature subset:

$$\text{Fitness}(S) = \text{Accuracy}_{K^*}(S) \quad (7)$$

where $\text{Accuracy}_{K^*}(S)$ represents the classification accuracy of the K* classifier, (S) denotes the subset of features encoded by the chromosome, and parent selection

was conducted using binary tournament selection with a tournament size of ($k = 2$). This configuration was adopted deliberately to maintain low selection pressure, thereby promoting genetic diversity within the population and reducing the likelihood of premature convergence. Such considerations are particularly important given the small, class-imbalanced dataset, where fitness estimates are inherently noisy. Furthermore, binary tournament selection relies solely on relative rankings of fitness values rather than on their absolute values, making it robust and reliable even in the narrow dynamic range of accuracy-based fitness. This selection is iteratively applied to choose the next parent, improving the population over generations. The crossover rate parameter determines whether the selected two chromosomes undergo crossover. In this work, the crossover rate parameter is distributed within $[0.7, \dots, 0.8]$, corresponding to the values explored during the parameter tuning phase. In the final experiments, a fixed crossover rate of 0.8 was used, which yielded superior overall performance. A two-point crossover uses a segment from both parents to exchange and create offspring. The two positions randomly select two distinct cut positions $c1$ and $c2$, where $1 \leq C1 \leq C2 \leq L$.

The values of the outer segment of these two points are exchanged for the two chromosomes (Figure 2).

The mutation operator prevents early convergence and maintains population diversity by introducing new genetic material. The mutation rate used in this study is calculated according to the attribute number N :

$$\text{Mutation rate} = \frac{1}{N} \quad (8)$$

Some genes may be mutated during this stage. Therefore, the fitness values of the two new chromosomes are compared with those of the original chromosomes. Evaluation and replacement are demonstrated by the feature's relevance to the learning algorithm itself. If the

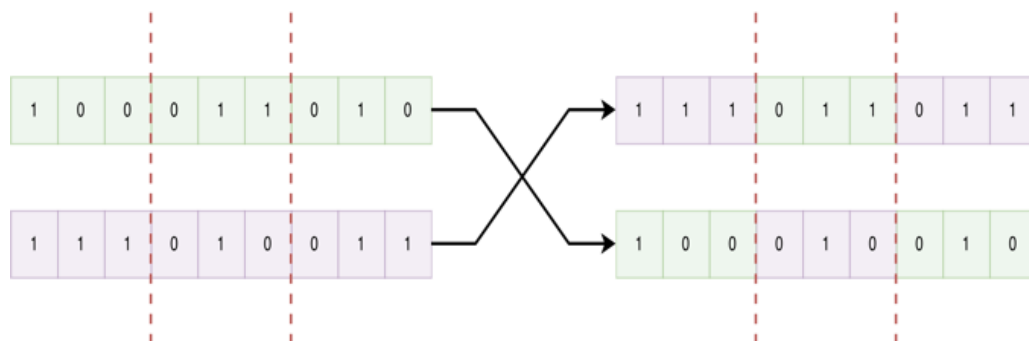


Figure 2. Two-point crossover in the proposed algorithm

new chromosomes are of higher quality than the originals, they replace the originals. The crossover and mutation stages are repeated until the termination condition is met, i.e., either the maximum number of generations using crossover and mutation operators is reached or no further improvement is declared.

In the second stage, the training phase of the instance-based classifier model is built; all instances are memorized. In the classification phase, each new instance is compared with the closest instances in the training set using the entropy-based distance function (Equation 1). The class is then determined based on similarity to the closest instance. Figure 3 provides pseudocode for the proposed genetic instance-based learner algorithm.

4. Experimental study

An experimental evaluation was conducted using real-world benchmark datasets (lymphography) from the University of California, Irvine (UCI). This dataset contains four possible final diagnostic classes: metastases, fibrosis, malignant lymphatic involvement, and normal cases. Additionally, the performance of the proposed hybrid algorithm was tested on other medical benchmarking datasets. All experiments were performed using state-of-the-art baseline classification algorithms commonly used in lymphography diagnosis.

4.1. Dataset

The lymphography dataset consisted of medical patient

information used to investigate cancer research, provided by the University Medical Centre and the Oncology Institute, and associated with the National Cancer Institute and other oncology organizations. It consisted of 18 attributes (i.e. features) and 148 cases with 4 possible final diagnostic classes. This study aims to identify the main attributes (factors) expected to be highly related to the occult development of cancer. Table 2 provides details of the dataset, including attribute type and attribute value.

4.2. Evaluation procedure

In this stage, a 10-fold cross-validation procedure was employed. The dataset used in this procedure was partitioned into 10 equal-sized subsets. In each iteration, nine subsets were used for learning, and the remaining subset was employed for testing. This process ran 10 times, each time utilizing different subsets for learning and testing, to ensure that all subsets were used in both stages. Finally, the performances for the 10 tests were averaged and reported using evaluation metrics selected from conventional classification literature. These benchmarks included the classification accuracy, true positive (TP) rate, false positive (FP) rate, *F*-measure, precision, recall, Kappa statistic, mean absolute error, root-mean-squared error, relative absolute error, and root-relative-squared error:

- (i) Classification accuracy: a machine learning criterion that measures the proportion of correct predictions by representing the overall probability of the correct classification.

```

1  Input: Dataset D with all training instances, m features, population size, and maximum number of generations
2  Output: classification model & Best feature subset
3  Population Initialization
4  Repeat
5  Evaluation Initial Population
6  Select parent chromosomes based on tournament procedure
7  Crossover: apply two-point crossover
8  Mutation: randomly flip bits
9  Fitness function calculation:
10 For each offspring chromosome:
11 Select subset of features indicated in the chromosome
12 Classify using IBL
13 Compute fitness value
14 Replacement
15 Until (GenerationsIndex  $\geq$  Maximum number of generations) OR (ImprovementTest  $\geq$  ImprovementIndex)
16 Select Best Chromosome
17 For each training instance:
18 Train IBL model using global best subset of features
19 Select the k instances with the smallest using entropy-based distance
20 Determine class based on similarity of closest instance
21 Return the IBL model and the best subset of feature.
22 End

```

Figure 3. Genetic-instance-based learner pseudocode

- (ii) TP rate: also known as the sensitivity metric; it is defined as a measurement of the correctly predicted positive cases to the total number of actual positive cases.
- (iii) FP rate: also known as the false alarm metric; it is defined as a measurement of the cases that are incorrectly classified as positive to the total number of actual negative cases.
- (iv) *F*-measure: a popular machine learning metric used to measure the harmonic mean of precision and recall.
- (v) Recall: a machine learning performance criterion that measures the proportion of TP instances a model correctly predicts.
- (vi) Precision: a machine learning performance criterion that measures the proportion of TP instances to all predicted positives.
- (vii) Kappa statistic: also known as an interrater reliability metric; it is defined as a measurement of a model's performance compared with random chance.
- (viii) Mean absolute error: also known as a regression metric; it is defined as a measurement of the average of the absolute differences between actual and predicted values.
- (ix) Root-mean-squared error: also known as the standard deviation of residuals metric; it is defined as the square root of the average of squared differences between actual and predicted values.
- (x) Root-relative-squared error: also known as unit-independent performance criteria; it indicates how much better a model behaves compared with a simple classifier that simply predicts the average of the actual values.
- (xi) Relative absolute error: it is a popular machine learning metric used to measure the performance of a classification model by partitioning the overall absolute error by the absolute difference between the model's mean and the actual values.

4.3. Results and discussion

The embedded method directly integrates feature selection into the training process by incorporating three GA-based approaches with different lazy-learning KNN models. Each of these models has distinct characteristics and components (e.g., GA–KNN and GA–instance-based k [IBk]).

The GA–KNN method combines the proposed GA-based feature selection with the well-known KNN classifier and uses the Euclidean distance to measure similarity. GA–K* is our proposed algorithm that uses GA for feature subset selection and KNN with a relative-entropy-based similarity measure. GA–IBk can determine

the weights for each neighbor based on their distance from the test instance. The influence of the closest neighbors is then used in classification.

Table 2. Characteristics of lymphography data

No.	Feature name	Attribute value	Attribute type
1	Lymphatics	4	Nominal
2	Block of Affere	2	Nominal
3	Block of Lymph. C	2	Nominal
4	Block of Lymph. S	2	Nominal
5	By Pass	2	Nominal
6	Extravasates	2	Nominal
7	Regeneration	2	Nominal
8	Early Uptake in	2	Nominal
9	Lym.nodes dimin	[0–3]	Numerical
10	Ym.nodes enlar	[1–4]	Numerical
11	Changes in lym	3	Nominal
12	Defect in node	4	Nominal
13	Changes in node	4	Nominal
14	Changes in stru	8	Nominal
15	Special forms	3	Nominal
16	Dislocation of	2	Nominal
17	Exclusion of no	2	Nominal
18	No. of nodes in	[0–80]	Numerical
19	Class	4	Nominal

The efficacy of the proposed models for lymphography-based cancer diagnosis was assessed using three different lazy-learning models (Table 3). These benchmarks included classification accuracy, TP rate, FP rate, *F*-measure, precision, and recall. The best results were achieved by the GA–K* model, which uses an entropy-based distance function across all evaluation metrics.

Table 4 presents a comparison of the proposed classifier with the most relevant state-of-the-art classification algorithms (overall performance): SVM, NB, RF, DT, and KNN. GA–K* outperformed the other classifiers in classification accuracy, TP rate, FP rate, *F*-measure, precision, and recall.

An additional experiment was performed on the proposed classifier to provide deeper insights into prediction error measures and regression (Table 5). The following evaluation metrics were adopted to demonstrate similarity to target values and assess robustness to class

imbalance: Kappa statistic, mean absolute error, root-mean-squared error, relative absolute error, and root-relative-squared error. A high Kappa statistic indicates a strong classification agreement between actual and predicted class values. By contrast, low values of the other four metrics indicate similarity to the true values and thereby indicate good model performance. A comparison with other classifiers indicated that the proposed classifier achieved superior performance in all evaluation metrics.

As shown in Table 6, the strong overall classification performance achieved by GA–K* is not attributable to class imbalance. Despite the limited representation of the fibrosis and normal classes (4 and 2 instances, respectively), both were recognized without error (recall = 1.000), and the two majority classes were classified with comparably high recall and precision. Consequently, this observation is further supported by the balanced macro-averaged *F*-measure (0.993). Notably, this confirms that the proposed method maintained robust predictive performance across both majority and minority classes. The results therefore reflect an important characteristic of a clinical decision-support system: rare conditions are often the most consequential; nevertheless, the minority-class estimates were derived

from a small number of instances and should be interpreted with due consideration and appropriate caution.

Further experiments were performed using common medical benchmark datasets with varying attribute counts, label values, and instance counts. Table 7 demonstrates that the proposed classifier outperformed RF, SVM, and KNN across all datasets. Furthermore, the proposed hybrid classifier outperformed NB in seven datasets and DT in five datasets. Considering the overall average ranking of classification performance, the proposed classifier achieved the highest result. This performance is consistent with the expected advantage of using genetic feature selection with the instance-based learner: it allows selecting the most relevant subset of features and consequently improves classification results.

On the lymphography data, the GA–K* approach demonstrated superiority across all four diagnostic categories, surpassing both the other alternative GA-based lazy-learner hybrids (GA–KNN and GA–IBk) and the conventional classifiers tested. The performance advantages can lie in the complementary interaction between two components that support one another. The GA effectively removes attributes that are redundant or weakly

Table 3. Classifier performance for the lymphography domain

Classifiers	Class	<i>F</i> -measure	Precision	Recall	TP rate	FP rate	Accuracy (%)
Proposed classifier (GA–K*)	Normal	1.000	1.000	1.000	1.000	0.000	98.6486
	Metastases	0.988	0.976	1.000	1.000	0.030	
	Malign_lymph	0.983	1.000	0.967	0.967	0.000	
	Fibrosis	1.000	1.000	1.000	1.000	0.000	
	Average	0.986	0.987	0.986	0.986	0.016	
GA–IBk	Normal	1.000	1.000	1.000	1.000	0.000	90.9091
	Metastases	0.900	0.900	0.900	0.900	0.083	
	Malign_lymph	0.900	0.900	0.900	0.900	0.083	
	Fibrosis	1.000	1.000	1.000	1.000	0.000	
	Average	0.909	0.909	0.909	0.909	0.076	
GA–KNN	Normal	1.000	1.000	1.000	1.000	0.000	96.6667
	Metastases	0.974	0.950	1.000	1.000	0.091	
	Malign_lymph	0.933	1.000	0.875	0.875	0.000	
	Fibrosis	1.000	1.000	1.000	1.000	0.000	
	Average	0.966	0.968	0.967	0.967	0.058	

Abbreviations: FP: False positive; GA: Genetic algorithm; IbK: Instance-based k; KNN: k-nearest neighbor; TP: True positive.

Table 4. Classifier performance for the lymphography domain

Classifiers	Class	F-measure	Precision	Recall	TP rate	FP rate	Accuracy (%)
GA–K*	Normal	1.000	1.000	1.000	1.000	0.000	98.6486
	Metastases	0.988	0.976	1.000	1.000	0.030	
	Malign_lymph	0.983	1.000	0.967	0.967	0.000	
	Fibrosis	1.000	1.000	1.000	1.000	0.000	
	Average	0.986	0.987	0.986	0.986	0.016	
KNN	Normal	1.000	1.000	1.000	1.000	0.000	93.3333
	Metastases	0.947	0.947	0.947	0.947	0.091	
	Malign_lymph	0.875	0.875	0.875	0.875	0.045	
	Fibrosis	1.000	1.000	1.000	1.000	0.000	
	Average	0.933	0.933	0.933	0.933	0.070	
DT	Normal	1.000	1.000	1.000	1.000	0.000	70.2703
	Metastases	0.777	0.792	0.763	0.763	0.258	
	Malign_lymph	0.690	0.678	0.702	0.702	0.224	
	Fibrosis	0.286	0.250	0.333	0.333	0.022	
	Average	0.735	0.738	0.732	0.732	0.236	
RF	Normal	0.667	1.000	0.500	0.500	0.000	84.4595
	Metastases	0.874	0.849	0.901	0.901	0.194	
	Malign_lymph	0.807	0.828	0.787	0.787	0.115	
	Fibrosis	0.857	1.000	0.750	0.750	0.000	
	Average	0.843	0.846	0.845	0.845	0.154	
NB	Normal	0.800	0.667	1.000	1.000	0.007	85.8108
	Metastases	0.873	0.857	0.889	0.889	0.179	
	Malign_lymph	0.831	0.860	0.803	0.803	0.092	
	Fibrosis	1.000	1.000	1.000	1.000	0.000	
	Average	0.858	0.859	0.858	0.858	0.136	
SVM	Normal	1.000	1.000	1.000	1.000	0.000	95.4545
	Metastases	0.957	1.000	0.917	0.917	0.000	
	Malign_lymph	0.941	0.889	1.000	1.000	0.071	
	Fibrosis	1.000	1.000	1.000	1.000	0.000	
	Average	0.955	0.960	0.955	0.955	0.026	

Abbreviations: DT: Decision tree; FP: False positive; GA: Genetic algorithm; KNN: k-nearest neighbor; NB: naïve Bayes; RF: Random forest; SVM: Support vector machine; TP: True positive.

Table 5. Classifier performance for the lymphography domain

Classifiers	Kappa statistic	Mean absolute error	Root-mean-squared error	Relative absolute error	Root-relative-squared error
GA–K*	0.9744	0.0181	0.0887	6.7497	24.3474
KNN	0.8723	0.0844	0.2061	30.5324	54.6439
DT	0.4921	0.1338	0.3658	52.4000	103.3506
RF	0.7010	0.1469	0.2540	54.7753	69.7500
NB	0.7321	0.0937	0.2409	34.9332	66.1412
SVM	0.9209	0.2538	0.3178	91.7615	84.1956

Abbreviations: DT: Decision tree; GA: Genetic algorithm; KNN: k-nearest neighbor; NB: naïve Bayes; RF: Random forest; SVM: Support vector machine.

Table 6. Per-class performance of GA–K* for the lymphography domain

Class	Support	Precision	Recall	F-measure	ROC area	PRC area
Normal	2	1.000	1.000	1.000	1.000	1.000
Metastases	81	0.976	1.000	0.988	0.999	1.000
Malignant lymph	61	1.000	0.967	0.983	0.999	0.999
Fibrosis	4	1.000	1.000	1.000	1.000	1.000
Weighted avg.	148	0.987	0.986	0.986	0.999	0.999
Macro avg.	-	0.994	0.992	0.993	–	–

Abbreviations: GA: Genetic algorithm; PRC: Precision–recall curve; ROC: Receiver operating characteristic curve.

Table 7. Classifier performance for other medical domains

Dataset	Metric	GA–K*	KNN	DT	RF	NB	SVM
HeartStatlog	Accuracy (%)	85.1852	79.6296	76.6667	81.4815	83.7037	84.0741
	Rank	1	5	6	4	3	2
Hepatitis	Accuracy (%)	90.3226	80.6452	83.8710	85.1613	84.5161	85.1613
	Rank	1	6	5	2.5	4	2.5
WisconsinBreastCancer	Accuracy (%)	97.1429	95.1359	94.5637	96.5665	95.9943	96.9957
	Rank	1	5	6	3	4	2
Breast Cancer	Accuracy (%)	75.8741	72.3776	75.9245	69.5804	71.6783	69.5804
	Rank	2	3	1	5.5	4	5.5
Pima Indians	Accuracy (%)	81.8182	70.1823	73.8281	75.7813	76.3021	77.3438
	Rank	1	6	5	4	3	2
Dermatology	Accuracy (%)	97.2973	94.5355	93.9891	95.9016	97.3678	95.3552
	Rank	2	5	6	3	1	4
Hypothyroid	Accuracy (%)	99.3057	91.5164	99.5758	99.3107	95.281	93.6108
	Rank	3	6	1	2	4	5
Lung Cancer	Accuracy (%)	81.2500	68.7500	78.1250	68.7500	78.1250	65.6250
	Rank	1	4.5	2.5	4.5	2.5	6
Average rank		1.50	5.06	4.06	3.56	3.19	3.63

Abbreviations: DT: Decision tree; GA: Genetic algorithm; KNN: k-nearest neighbor; NB: naïve Bayes; RF: Random forest; SVM: Support vector machine.

informative, and, subsequently, K* classifies them using an entropy-based distance, which is particularly well suited to categorical, locally structured features, and performs far better than the Manhattan or Euclidean metrics used by the KNN variants. The performance pattern observed across other domains (Table 7) is less uniform: GA–K* achieved the second-highest classification on the Breast Cancer dataset (75.87%) and ranked third on the Hypothyroid dataset (99.31%) among the six classifiers. These results are interpreted through the principles of the “No Free Lunch theorem,” which states that no single learning classifier can be expected to outperform all alternatives across every application domain. The Hypothyroid dataset is highly imbalanced and padded with low-variance binary attributes, while the Breast Cancer dataset has only a small number of features with missing values. Under such conditions, feature selection is inherently reduced as fewer informative feature interactions are available for exploitation. Nevertheless, the proposed method remained competitive across all datasets. The performance further indicates that GA–K* is particularly effective on moderate-sized, categorical problems of the kind lymphography diagnosis. These results should be weighed against the small size and skew of the dataset; since fitness is driven by overall accuracy, class-balanced future work should incorporate evaluation and testing of the proposed approach on larger lymphographic cohorts, which are the natural next steps for establishing clinical applicability.

5. Conclusion

This research presented a lymphography cancer diagnosis method that collaboratively embeds GA within the learning process of an instance-based learning classifier. The proposed approach effectively enhanced feature relevance, enabling the classifier to focus on the most discriminative patterns in a lymphographic dataset. The integration of the evolutionary algorithm (i.e., GA) and instance-based classification optimized detection accuracy whilst reducing noise and redundancy in this medical domain. Overall, the assessment demonstrated that the proposed embedded approach outperformed standalone approaches and other embedded classifiers across all evaluation metrics for diagnosing lymphography cancer and other medical datasets. For future research directions, this work may be extended by testing with additional high-dimensional lymphography and medical data, incorporating hybrid evolutionary approaches, performing parameter adaptation, and optimizing the method’s generalizability to clinical applicability. Another future work will be to explore a multi-objective formulation that jointly maximizes classification performance and minimizes the number of selected features, using a class-balanced metric to replace

the pure-accuracy fitness function.

Acknowledgments

The authors would like to express their sincere gratitude to Shatt Al-Arab University for providing research facilities and laboratory resources.

Funding

Shatt Al-Arab University funded this research.

Conflict of interest

The authors declare that there are no conflicts of interest regarding this research.

Author contributions

Conceptualization: Hayder Naser Khraibet Al-Behadili

Formal analysis: Hayder Naser Khraibet Al-Behadili

Investigation: Hayder Naser Khraibet Al-Behadili

Methodology: Ayad Mohammed Jabbar

Writing–original draft: Hayder Naser Khraibet Al-Behadili, Ayad Mohammed Jabbar

Writing–review & editing: Hayder Naser Khraibet Al-Behadili, Ayad Mohammed Jabbar

Availability of data

The secondary dataset used in this study is publicly available from UCI (<https://archive.ics.uci.edu/dataset/63/lymphography>).

References

- Abdollahi, J., & Aref, S. (2024). Early prediction of diabetes using feature selection and machine learning algorithms. *SN Computer Science*, 5(2), 217.
<https://doi.org/10.1007/s42979-023-02545-y>
- Abdollahi, J., & Nouri-Moghaddam, B. (2022). Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. *Iran Journal of Computer Science*, 5(3), 205–220.
<https://doi.org/10.1007/s42044-022-00100-1>
- Abdolrazzaghi-Nezhad, M., & Izadpanah, S. (2025). A new hybrid fuzzy bio-inspired classifier for cancer detection using cuckoo optimization and hyper-planes. *Data Technologies and Applications*, 59(3), 416–451.
<https://doi.org/10.1108/DTA-06-2024-0647>
- Ali, W., & Saeed, F. (2023). Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data. *Processes*, 11(2), 562.
<https://doi.org/10.3390/pr11020562>
- Amini, F., & Hu, G. (2021). A two-layer feature selection method using genetic algorithm and elastic net. *Expert Systems with*

Applications, 166, 114072.

<https://doi.org/10.1016/j.eswa.2020.114072>

Babichev, S., Liakh, I., & Skvor, J. (2025). Integrating data mining, deep learning, and gene ontology analysis for gene expression-based disease diagnosis systems. *IEEE Access*, 13, 21265–21278.

<https://doi.org/10.1109/ACCESS.2025.3535999>

Cheng, C. H., & Shi, S. (2025). Artificial intelligence in cancer: Applications, challenges, and future perspectives. *Molecular Cancer*, 24(1), 274.

<https://doi.org/10.1186/s12943-025-02450-3>

Islam, T., Hoque, E., Ullah, M., Islam, T., Akter Nishu, N., & Islam, R. (2024). CNN-based deep learning approach for classification of invasive ductal and metastasis types of breast carcinoma. *Cancer Medicine*, 13(16), e70069.

<https://doi.org/10.1002/cam4.70069>

Lei, P., Fraser, C., Jones, D., Ubellacker, J. M., & Padera, T. P. (2024). Lymphatic system regulation of anti-cancer immunity and metastasis. *Frontiers in Immunology*, 15, 1449291.

<https://doi.org/10.3389/fimmu.2024.1449291>

Liu, X., Liang, Y., Sai, W., Yang, Z. Y., & Ye, H. S. (2018). A hybrid genetic algorithm with wrapper-embedded approaches for feature selection. *IEEE Access*, 6(1), 22863–22874.

<https://doi.org/10.1109/ACCESS.2018.2818682>

Mahmoud, A., & Takaoka, E. (2025). An enhanced machine learning approach with stacking ensemble learner for accurate liver cancer diagnosis using feature selection and gene expression data. *Healthcare Analytics*, 7, 100373.

<https://doi.org/10.1016/j.health.2024.100373>

Maurya, S. P., Sisodia, P. S., Mishra, R., & Pratap, D. (2024). Performance of machine learning algorithms for lung cancer prediction: A comparative approach. *Scientific Reports*, 14(1), 18562.

<https://doi.org/10.1038/s41598-024-58345-8>

Negm, A. S., Collins, J. D., Bendel, E. C., Takahashi, E., Koepsel, E. M. K., Gehling, K. J., ... & Thompson, S. M. (2024). MR lymphangiography in lymphatic disorders: Clinical applications, institutional experience, and practice development. *Radiographics*, 44(2), e230075.

<https://doi.org/10.1148/rg.230075>

Rabie, A. H., Aldawsari, M., Saleh, A. I., Saraya, M., & Rashad, M. (2025). HFSA: Hybrid feature selection approach to improve medical diagnostic system. *PeerJ Computer Science*, 11(1), e2764.

<https://doi.org/10.7717/peerj-cs.2764>

Raza, R., Wang, G., Wong, K. W., Laga, H., & Fisichella, M. (2025). ITL-LIME: Instance-based transfer learning for enhancing local explanations in low-resource data settings. *Proceedings*

of the 34th ACM International Conference on Information and Knowledge Management, November 2025, 2482–2492.

<https://doi.org/10.1145/3746252.3761183>

Saqib, S. M., Muhammad, O., Mazhar, T., Iqbal, M., Guizani, S., & Hamam, H. (2025). Medicine image classification using deep learning: Highlighting the MedNetMoBiL hybrid model. *Discover Computing*, 28(1), 103.

<https://doi.org/10.1007/s10791-025-09619-w>

Song, Y. W., Wang, J. S., Qi, Y. L., Wang, Y. C., Li, S., & Song, H. M. (2025). PFPSS: A doublelayer parallel embedded feature selection method for cancer gene expression data. *Journal of Big Data*, 12(1), 136.

<https://doi.org/10.1186/s40537-025-01144-3>

Song, Y. W., Wang, J. S., Qi, Y. L., Wang, Y. C., Song, H. M., & Shang-Guan, Y. P. (2025). Serial filter-wrapper feature selection method with elite guided mutation strategy on cancer gene expression data. *Artificial Intelligence Review*, 58(4), 111.

<https://doi.org/10.1007/s10462-024-11029-1>

Tabasi, S., Ahanian, I., Amiri, N., Dabanloo, N. J., & Barghamadi, H. (2025). Deep learning based cancer classification using selected gene expression data. *Signal Processing and Renewable Energy*, 9(3), 1–10.

<https://doi.org/10.57647/j.spre.2025.090318>

Thambi, S. V., G, I., Kammari, K. S., & Sahay, A. (2025). Hybrid metaheuristic optimisation for lung cancer image classification: Leveraging MOEA, PSO, and ACO algorithms. *Procedia Computer Science*, 258(1), 3781–3793.

<https://doi.org/10.1016/j.procs.2025.04.633>

Valeriano, M. G., Marzag, D. K., Montelongo, A., Roberto, C., Kiffer, V., Katz, N., & Lorena, A. C. (2026). Filtering instances and rejecting predictions to obtain reliable models in healthcare. *Machine Learning*, 115(1), 15.

<https://doi.org/10.1007/s10994-025-06941-8>

Yang, Y., Khorshidi, H. A., & Aickelin, U. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: Insights for medical problems. *Frontiers in Digital Health*, 6, 1430245.

<https://doi.org/10.3389/fdgth.2024.1430245>

Yavuz, G., Moghanjoughi, M. K., Dumlu, H., & Çakir, H. İ. (2025). A feature selection method combining filter and wrapper approaches for medical dataset classification. *Vietnam Journal of Computer Science*, 12(4), 375–393.

<https://doi.org/10.1142/S2196888824500246>

Zhang, S., Wang, T., Worden, K., Sun, L., & Cross, E. J. (2025). Canonical-correlation-based fast feature selection for structural health monitoring. *Mechanical Systems and Signal Processing*, 223(1), 111895.

<https://doi.org/10.1016/j.ymssp.2024.111895>

Zhu, C., Wang, Z., Peng, Y., & Xiao, W. (2025). An improved Red-billed blue magpie feature selection algorithm for medical data processing. *PLoS ONE*, 20(5), e0324866.

<https://doi.org/10.1371/journal.pone.0324866>

Zhu, J., Zhang, R., & Zhang, H. (2024). An MRI brain tumor segmentation method based on improved U-Net. *Mathematical Biosciences and Engineering*, 21(1), 778–791.

<https://doi.org/10.3934/mbe.2024033>