The Scientific Temper (2024) Vol. 15 (spl): 217-229

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.spl.26

https://scientifictemper.com/

RESEARCH ARTICLE

Optimizing predictive accuracy: A comparative study of feature selection strategies in the healthcare domain

M. A. Shanthi

Abstract

Feature selection is a critical preprocessing step in the development of machine learning models, particularly in the healthcare domain, where datasets often contain numerous features that may not contribute significantly to predictive performance. This study presents a comparative analysis of various feature selection techniques applied to healthcare datasets, evaluating their effectiveness in improving model accuracy and reducing computational costs. We investigate traditional filter-based methods, such as information gain and chi-square, alongside wrapper-based approaches and hybrid techniques that combine the strengths of both. Using multiple healthcare datasets encompassing diverse medical conditions, we assess the impact of these techniques on classification performance using metrics such as accuracy, precision, recall, and F1-score. Additionally, we analyze the robustness and scalability of each method in handling high-dimensional data. The findings reveal significant differences in performance, highlighting the strengths and weaknesses of each feature selection approach within the healthcare context. This comparative study provides valuable insights for researchers and practitioners, guiding them in selecting appropriate feature selection techniques to enhance predictive modeling in healthcare applications.

Keywords: Feature Selection, Filter based Feature Selection, Wrapper Approach, Optimization Technique, Clinical dataset.

Introduction

Healthcare is a fundamental pillar of societal well-being, playing a crucial role in maintaining and improving the health of individuals and communities. The significance of healthcare extends beyond the direct treatment of illness and injury; it encompasses preventive care, health education, and the promotion of healthy lifestyles, all of which contribute to a higher quality of life and increased life expectancy. In recent years, the importance of effective healthcare systems has become even more pronounced, highlighted by global challenges such as the COVID-19 pandemic, aging populations, and the rise of chronic diseases, Veena, A., & Gowrishankar, S. (2021), Salazar Reyna, R. J. (2019).

Effective healthcare systems are essential for several reasons. Firstly, they provide critical services that ensure early detection and timely treatment of diseases, thereby reducing mortality and morbidity rates. Preventive measures such as vaccinations, screenings, and public health campaigns play a vital role in mitigating the spread of infectious diseases and managing chronic conditions. Additionally, comprehensive healthcare services support mental health, maternal and child health, and geriatric care, addressing the diverse needs of the population at different life stages, Arvindhan, M., Rajeshkumar, D., & Pal, A. L. (2021).

Secondly, healthcare has a profound impact on economic stability and development. Healthy populations are more productive, with fewer workdays lost to illness and disability, leading to greater economic output and reduced healthcare costs. Investment in healthcare infrastructure and services also generates employment opportunities and drives innovation in medical research and technology, Bennett, M., Hayes, K., Kleczyk, E. J., & Mehta, R. (2022).

Furthermore, healthcare is integral to social equity and justice. Access to quality healthcare is a fundamental human right, and disparities in healthcare access and outcomes are often reflective of broader social inequalities. Ensuring that all individuals, regardless of socioeconomic status, geographic location, or cultural background, have access to

Assistant Professor, Idhaya College for Women (Affiliated to Bharathidasan University - Tiruchirappalli), Kumbakonam, Tamil Nadu, India.

***Corresponding Author:** M. A. Shanthi, Assistant Professor, Idhaya College for Women (Affiliated to Bharathidasan University - Tiruchirappalli), Kumbakonam, Tamil Nadu, India, E-Mail: drshantima75@gmail.com

How to cite this article: Shanthi, M. A. (2024). Optimizing predictive accuracy: A comparative study of feature selection strategies in the healthcare domain. The Scientific Temper, **15**(spl):217-229.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.spl.26

Source of support: Nil

Conflict of interest: None.

Received: 17/09/2024 **Accepted:** 26/09/2024 **Published**: 16/10/2024

essential health services is crucial for building inclusive and equitable societies, Nerkar, P. M., Liyakat, K. K. S., Dhaware, B. U., & Liyakat, K. S. S. (2023).

The intersection of healthcare with technology has ushered in a new era of possibilities for improving patient outcomes and operational efficiency. Advances in medical technologies, digital health solutions, and data analytics have revolutionized the way healthcare is delivered and managed. In particular, the utilization of clinical datasets has the potential to transform healthcare by enabling personalized medicine, improving diagnostic accuracy, and optimizing treatment plans. However, the complexity and high dimensionality of clinical data necessitate sophisticated analytical methods to extract meaningful insights and support decision-making processes, Kumari, J., Kumar, E., & Kumar, D. (2023).

In this context, feature selection methods have become indispensable tools for handling large clinical datasets. By identifying the most relevant and informative features, these methods enhance the interpretability and predictive performance of clinical models, facilitating better patient care and resource allocation. This study proposes a novel feature selection approach that leverages the information gain, reliefF algorithm, and whale optimization algorithm, aiming to address the challenges associated with clinical data analysis and contribute to the advancement of healthcare research and practice, Habehh, H., & Gohel, S. (2021).

Importance of Feature Selection Techniques

Feature selection is a fundamental process in the preparation of data for machine learning and data mining, particularly for high-dimensional datasets such as those encountered in clinical research. This process involves identifying and selecting the most relevant features from a dataset, which can significantly enhance the performance and efficiency of predictive models. Here are the key reasons why feature selection techniques are important, Nagarajan, S. M., Muthukumaran, V., Murugesan, R., Joseph, R. B., & Munirathanam, M. (2021), Durairaj, M., & Poornappriya, T. S. (2020):

Dimensionality Reduction

Reduces the number of features, simplifying models and making them more computationally efficient. Decreases training time and resource requirements, which is crucial when dealing with large-scale datasets, Patra, S. S., Harshvardhan, G. M., Gourisaria, M. K., Mohanty, J. R., & Choudhury, S. (2021).

Improved Model Performance

Enhances the predictive power of models by focusing on the most informative features. Eliminates irrelevant or redundant features that introduce noise, thereby improving model accuracy and robustness.

Enhanced Interpretability

Produces simpler models that are easier to understand and interpret. Facilitates clinical decision-making by providing insights into the most important factors influencing predictions.

Reduction of Overfitting

Helps in preventing overfitting by removing features that contribute to noise rather than the actual signal. Ensures that models generalize better to new, unseen data.

Cost and Resource Efficiency

Reduces the cost and effort associated with data collection and processing. Is particularly beneficial in clinical settings where some features may be expensive or difficult to measure.

Literature Review

The authors proposed a hybrid filter-wrapper approach for feature selection. An ensemble of filter methods, ReliefF and fuzzy entropy (RFE), is developed, and the union of top-n features from each method are considered. The equilibrium optimizer (EO) technique is combined with opposition-based learning (OBL), Cauchy mutation operator, and a novel search strategy to enhance its capabilities. The OBL strategy improves the diversity of the population in the search space. The Cauchy mutation operator enhances its ability to evade the local optima during the search, and the novel search strategy improves the exploration capability of the algorithm. This enhanced form of EO is integrated with eight time-varying S and V-shaped transfer functions to convert the solutions into binary form, binary enhanced equilibrium optimizer (BEE). The features from the ensemble are given as input to the binary enhanced equilibrium optimizer to extract the essential features. Fuzzy KNN based on Bonferroni mean is used as the learning algorithm, Vommi, A. M., & Battula, T. K. (2023).

The authors propose a new algorithm for feature selection based on a hybrid between powerful and recently emerged optimizers, namely, guided whale and dipper-throated optimizers. The proposed algorithm is evaluated using four publicly available breast cancer datasets. The evaluation results show the effectiveness of the proposed approach from the accuracy and speed perspectives. To prove the superiority of the proposed algorithm, a set of competing feature selection algorithms was incorporated into the conducted experiments. In addition, a group of statistical analysis experiments was conducted to emphasize the superiority and stability of the proposed algorithm, Atteia, G., El-kenawy, E. S. M., Samee, N. A., Jamjoom, M. M., Ibrahim, A., Abdelhamid, A. A., ... & Shams, M. Y. (2023).

The authors introduced the adaptive hybrid-mutated differential evolution (A-HMDE) method, targeting the inherent drawbacks of the differential evolution (DE)

algorithm. The A-HMDE incorporates four distinct strategies. Firstly, it integrates the mechanics of the spider wasp optimization (SWO) algorithm into DE's mutation strategies, yielding enhanced performance marked by high accuracy and swift convergence towards global optima. Secondly, adaptive mechanisms are applied to key DE parameters, amplifying the efficiency of the search process. Thirdly, an adaptive mutation operator ensures a harmonious balance between global exploration and local exploitation during optimization. Lastly, the concept of enhanced solution quality (ESQ), rooted in the RUN algorithm, guides DE to elude local optima, thus heightening the accuracy of obtained solutions, Mostafa, R. R., Khedr, A. M., Al Aghbari, Z., Afyouni, I., Kamel, I., & Ahmed, N. (2024).

The authors used three feature selection filter algorithms (FSFAs): relief filter, step disc filter, and Fisher filter algorithm and 15 classifiers using a free data mining Tanagra software having UCI Machine Learning Repository. This process is done on a medical dataset with 20 attributes and 155 instances. As a result, the error rate is obtained in terms of accuracy, which shows the performance of algorithms regarding patient survival. This work also shows the independent comparison of FSFAs with classification algorithms using continuous values and the FSFA without using classification algorithms. This paper shows that the obtained result of the classification algorithm gives promising results in terms of error rate and accuracy, Masood, F., Masood, J., Zahir, H., Driss, K., Mehmood, N., & Farooq, H. (2023).

A novel hybrid wrapper-based feature selection method is proposed to tackle these issues effectively. In order to improve the exploration ability of the particles, the Sine factor is integrated with the equilibrium optimizer (EO) technique. A bi-phase mutation (BM) scheme is integrated to enhance the exploitation phase of the EO algorithm (BM-based Hybrid EO, BMHEO). The BMHEO method is evaluated by employing four different classifiers – KNN, SVM, random forest (RF) and discriminant analysis (DA). It is observed that the random forest classifier exhibits superior performance compared to the other three classifiers. Eight S-shaped and V-shaped transfer functions are integrated to convert the solutions to binary form, Vommi, A. M., & Battula, T. K. (2023).

The authors presented a comprehensive investigation into diabetes detection models by integrating two feature selection techniques: The Akaike information criterion and genetic algorithms. These techniques are combined with six prominent classifier algorithms, including support vector machine, random forest, k-nearest neighbor, gradient boosting, extra trees, and naive Bayes. By leveraging clinical and paraclinical features, the generated models are evaluated and compared to existing approaches. The results demonstrate superior performance, surpassing accuracies of 94%. Furthermore, the use of feature selection techniques allows for working with a reduced dataset. The significance of feature selection is underscored in this study, showcasing its pivotal role in enhancing the performance of diabetes detection models, García-Domínguez, A., Galván-Tejada, C. E., Magallanes-Quintanar, R., Gamboa-Rosales, H., Curiel, I. G., Peralta-Romero, J., & Cruz, M. (2023).

The use of feature selection in gene expression studies began at the end of the 1990s with the analysis of human cancer microarray datasets. Since then, gene expression technology has been perfected, the human genome project has been completed, new microarray platforms have been created and discontinued, and RNA-seq has gradually replaced microarrays. However, most feature selection methods in the last two decades were designed, evaluated, and validated on the same datasets from the microarray technology's infancy. In this review of over 1200 publications regarding feature selection and gene expression, published between 2010 and 2020, we found that 57% of the publications used at least one outdated dataset, 23% used only outdated data, and 32% did not cite data sources, Grisci, B. I., Feltes, B. C., de Faria Poloni, J., Narloch, P. H., & Dorn, M. (2024).

This study is aimed at building a potential machine learning model to predict heart disease in the early stage, employing several feature selection techniques to identify significant features. Three different approaches were applied for feature selection, such as chi-square, ANOVA, and mutual information, and the selected feature subsets were denoted as SF1, SF2, and SF3, respectively. Then, six different machine learning models such as logistic regression (C1), support vector machine (C2), K-nearest neighbor (C3), random forest (C4), Naive Bayes (C5), and decision tree (C6) were applied to find the most optimistic model along with the best-fit feature subset, Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., ... & Moni, M. A. (2023).

The present study examines the role of feature selection methods in optimizing machine learning algorithms for predicting heart disease. The Cleveland Heart disease dataset with sixteen feature selection techniques in three categories of filter, wrapper, and evolutionary was used. Then, seven algorithms Bayes net, Naïve Bayes (BN), multivariate linear model (MLM), Support Vector Machine (SVM), logit boost, j48, and random forest were applied to identify the best models for heart disease prediction. Precision, F-measure, specificity, accuracy, sensitivity, ROC area, and PRC were measured to compare feature selection methods effect on prediction algorithms, Noroozi, Z., Orooji, A., & Erfannia, L. (2023).

Machine learning algorithms are now crucial in the medical field, especially when using medical databases to diagnose diseases. Such efficient algorithms and data processing techniques are applied to predict various diseases and offer much potential for accurate heart

disease prognosis. Therefore, this study compares the performance logistic regression, decision tree, and support vector machine (SVM) methods with and without Boruta feature selection. The Cleveland clinic heart disease dataset acquired from Kaggle, which consists of 14 features and 303 instances, was used for the investigation. It was found that the Boruta feature selection algorithm, which selects six of the most relevant features, improved the results of the algorithms, Manikandan, G., Pragadeesh, B., Manojkumar, V., Karthikeyan, A. L., Manikandan, R., & Gandomi, A. H. (2024).

The authors proposed a hybrid novel technique, CSSMO-based gene selection for cancer classification. First, we made alterations of the fitness of spider monkey optimization (SMO) with the cuckoo search algorithm (CSA) algorithm viz., CSSMO for feature selection, which helps to combine the benefit of both metaheuristic algorithms to discover a subset of genes which helps to predict a cancer disease in early stage. Further, to enhance the accuracy of the CSSMO algorithm, we choose a cleaning process, minimum redundancy maximum relevance (mRMR) to lessen the gene expression of cancer datasets. Next, these subsets of genes are classified using deep learning (DL) to identify different groups or classes related to a particular cancer disease, Mahto, R., Ahmed, S. U., Rahman, R. U., Aziz, R. M., Roy, P., Mallik, S., ... & Shah, M. A. (2023).

A novel multi-class-based feature extraction (MC-FE) method has been proposed for medical data classification. Genomic datasets, or gene expression-based microarray medical datasets, are categorized for cancer diagnosis. The first stage involves applying a feature extraction technique. The principal component analysis (PCA) is used to extract the features for medical data classification to detect leukemia, colon tumors, and prostate cancer. The modified particle swarm optimization (MPSO) technique is used in the second stage to pick features from high-dimensional microarray medical datasets like prostate cancer, leukemia, and colon tumors. Finally, SVM, KNN, and Naive Bayes classifiers are used to classify medical data, Razzaque, A., & Badholia, A. (2024).

A systematic literature review is conducted on five major digital databases of science and engineering. Results: The primary search included 695 articles. After removing 263 duplicated articles, 432 studies remained to be screened. Among those, 317 irrelevant papers were removed. We then excluded 77 studies according to the exclusion criteria. Finally, 38 articles were selected for this study. Conclusion: Out of 38 studies, 28 papers discussed Swarm-based algorithms, 2 papers studied genetic algorithms, and 8 papers covered algorithms in both categories. Considering the application domains, 21 of the articles focused on problems in the healthcare sector, while the rest mainly investigated issues in cybersecurity, text classification, and image processing. Hybridization with other BIAs was employed by approximately 18.5% of papers, and 13 out of 38 studies used S-shaped transfer functions. The majority of studies used supervised classification methods such as k-NN and SVM for building fitness functions, Pham, T. H., & Raahemi, B. (2023).

Information Gain-Based Feature Selection Method

Information gain (IG) is a popular feature selection method used primarily in the context of classification problems. It measures the reduction in uncertainty or entropy in the target variable due to the presence of a feature, Sharma, A., & Mishra, P. K. (2022); Ramasamy, M., & Meena Kowshalya, A. (2022).

Understanding entropy

Entropy is a measure of the unpredictability or impurity in a dataset. In the context of feature selection, it quantifies the amount of disorder or randomness in the target variable.

For a target variable Y with n possible values, the entropy $H(Y)$ is defined as:

$$H(Y) = -\sum_{i=1}^{n} P(y_i) \log_2 P(y_i) \quad (1)$$

Where $P(y_i)$ is the probability of occurrence of the i-th value of y.

Conditional Entropy

Conditional entropy quantifies the amount of entropy (uncertainty) in the target variable Y given the presence of another variable X. It is defined as:

$$H(Y|X) = -\sum_{j=1}^{m} P(x_j) \sum_{i=1}^{n} P(y_i|x_j) \log_2 P(y_i|x_j) \quad (2)$$

Where $P(x_j)$ is the probability of the j-th value of X, and $P(y_i|x_j)$ is the conditional probability of y_i given x_j.

Information Gain Calculation

Information gain (IG) is the reduction in energy of the target variable Y after observing the feature X. It measures how much knowing the feature X reduces the uncertainty about the target variable Y. The IG is calculated as:

$$IG\ (Y,X) = H(Y) - H(Y|X) \quad (3)$$

Feature Selection using Information Gain

The steps involved in selecting features using information gain are as follows:

Calculate Entropy of the Target Variable

Compute the entropy $H(Y)$ of the target variable Y using the formula mentioned above.

Calculate Conditional Entropy for each feature

For each feature X_i in the dataset, calculate the conditional entropy $H(Y|X_i)$ of the target variable given the feature.

Compute information Gain for each feature

For each feature X_i, compute the $IG(Y,X_i)$ using the formula:

$$IG(Y,X_i) = H(Y) - H(Y|X_i) \quad (4)$$

Rank features based on Information Gain

Rank the features based on their Information Gain values. Features with higher Information Gain are considered more informative and relevant for predicting the target variable.

Select Top features

Select the top k features with the highest IG values as the most relevant features for the model.

Relieff Based Feature Selection Method

The ReliefF algorithm is an extension of the original Relief algorithm and is designed to handle multi-class problems and noisy data. It is a feature weighting method that evaluates the importance of features based on their ability to distinguish between instances that are near each other, Liu, J., Zhao, L., Si, C., Guan, H., & Dong, X. (2023), Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021).

Initialization

ReliefF starts by initializing a weight vector W for all features, setting each weight to Zero:

$$W[f_i] = 0 \quad (5)$$

for each feature f_i in the dataset.

Random Sampling

ReliefF iteratively samples instances from the dataset. For each iteration, it randomly selects an instance R from the dataset.

Finding Nearest Neighbors

For the selected instance R, ReliefF identifies:

- k nearest neighbors from the same class as R (called «nearest hits»)
- k nearest neighbors from the same class as R (called «nearest hits»)

Updating Feature Weights

ReliefF updates the weights of the features based on how well they can distinguish between R and its nearest hits and misses. The update rule for the weight of a feature f is:

$$W[f] = W[f] - \frac{1}{m}\sum_{i=1}^{k}\left(\frac{|f(R) - f(H_i)|}{k}\right) + \frac{1}{m}\sum_{c \neq class(R)}\left(\frac{P(c)}{1-P(class\,(R))}\sum_{j=1}^{k}\left(\frac{|f(R)-f(M_j^c)|}{k}\right)\right) \quad (6)$$

Where $W[f]$ is the weight of feature f, m is the number of iterations, H_i is the i-th nearest hit, M_j^c is the j-th nearest miss from class c, $P(c)$ is the prior probability of class c, $f(R)$ is the value of feature f for instance R.

The update increases the weight of a feature if it helps distinguish between instances of different classes (i.e., if the difference between R and nearest misses is large) and decreases the weight if it does not help distinguish between instances of the same class (i.e., if the difference between R and nearest hits is large).

Iteration

Steps 2-4 are repeated for a predefined number of iterations or until convergence. Each iteration refines the weights, improving the ranking of features based on their ability to discriminate between instances of different classes.

Ranking and Selecting Features

After completing the iterations, the features are ranked based on their final weights. Features with higher weights are considered more important and relevant for the classification task.

ReliefF is a powerful feature selection method that evaluates feature importance based on their ability to discriminate between instances of different classes, considering local information around each instance. This method is particularly useful for handling multi-class problems and noisy data, providing a robust way to select relevant features that contribute to accurate and efficient predictive modeling.

Whale Optimization Algorithm-Based Feature Selection Method

The whale optimization algorithm (WOA) is a nature-inspired metaheuristic optimization algorithm based on the social hunting behavior of humpback whales, specifically their bubble-net feeding strategy. In feature selection, WOA can be employed to find an optimal subset of features that maximizes the performance of a predictive model. Here's a detailed explanation of how the WOA-based feature selection method works: Riyahi, M., Rafsanjani, M. K., Gupta, B. B., & Alhalabi, W. (2022), Alwateer, M., Almars, A. M., Areed, K. N., Elhosseini, M. A., Haikal, A. Y., & Badawy, M. (2021):

Stage 1: Initialization

- *Population Initialization*

Initialize a population of whales (solutions), where each whale represents a potential solution (a subset of features). The size of the population is N, and each whale's position in the search space is represented as a binary vector indicating the presence (1) or absence (0) of features.

- *Fitness Function*

Define a fitness function to evaluate the quality of each solution. This function typically measures the predictive accuracy of a machine-learning model using the selected features.

Stage 2: Whale Behavior Modeling

WOA mimics two main behaviors of humpback whales: the encircling prey mechanism and the bubble-net attacking method.

- *Stage 2.1: Encircling Prey*
- Whales perceive the position of the best solution (whale) found so far, updating their positions to move towards this optimal solution.
- Update the position of each whale according to the following equations:

$$\bar{D} = |\bar{C}.\bar{X}*(t) - \bar{X}(t)| \quad (7)$$

$$\bar{X}(t+1) = \bar{X}*(t) - \bar{A}.\bar{D} \quad (8)$$

Where $\bar{X}*(t)$ is the position vector of the best solution, $\bar{X}(t)$ is the position vector of the current whale, $\bar{A}$ and $\bar{C}$ are coefficient vector calculated as:

$$\bar{A} = 2\bar{a}.\bar{r} - \bar{a} \quad (9)$$

$$\bar{C} = 2.\bar{r} \quad (10)$$

Where $\bar{a}$ decreases linearly from 2 to 0 over the course of iterations, and $\bar{r}$ is a random vector in $[0,1]$.

- *Stage 2.2: Bubble – Net Attacking Model*

This method includes two strategies: shrinking encircling mechanism and spiral updating position.

- Shrinking Encircling Mechanism: This is controlled by $\bar{A}$. When $|\bar{A}| < 1$, the whales move towards the best solution.
- Spiral Updating Position: This models the helix-shaped movement of whales around their prey.

$$\bar{X}(t+1) = \bar{D}'.e^{bl}.cos(2\pi l) + \bar{X}*(t) \quad (11)$$

Where $\bar{D}' = |\bar{X}*(t) - \bar{X}(t)|$, b is a constant defining the spiral shape, and l is the random number in $[-1,1]$. The probability p is used to switch between the shrinking encircling mechanism and the spiral model. Typically, p = 0.50.

Stage 3: Exploration Phase

To enhance exploration, whales search for prey randomly based on the positions of other whales. When $|\vec{A}| \geq 1$, the whales move towards random positions in the search space, facilitating exploration.

Stage 4: Fitness Evaluation

Evaluate the fitness of each whale (solution) using the defined fitness function. This step assesses how well the selected subset of features performs in terms of model accuracy.

Stage 5: Updating Best Solution

Identify the whale with the best fitness score. Update the best-known position $\bar{X}*(t)$ if a better solution is found.

Stage 6: Iteration

Repeat steps 2-5 for a predefined number of iterations or until convergence criteria are met.

Stage 7: Selection of Optimal Feature Subset

After the iterations, the position vector of the best whale represents the optimal subset of features. Features corresponding to 1s in the binary vector are selected for the final model.

Artificial bee colony (ABC) based feature selection method

Artificial bee colony (ABC) optimization is a population-based metaheuristic inspired by the foraging behavior of honeybees. It is widely used for solving optimization problems, including feature selection. In the context of feature selection, ABC optimizes the subset of features by exploring the search space and evaluating the quality of different feature subsets based on a chosen evaluation metric, such as accuracy, F1-score, etc.

Initialization Phase

In ABC, the solution space (in feature selection, different feature subsets) is initialized randomly, and each solution is associated with a "food source."

Let, $X_{i,j}$ represent the j-th feature (binary) of the ith solution. n represent the number of features. m represents the number of bees or solutions. Each bee starts with a random solution:

$$X_i = (X_{i,1}, X_{1,2}, \ldots, X_{i,n})$$

Where $X_{i,j} \epsilon \{0,1\}$. If $X_{i,j} = 1$, the j-th feature is selected in the subset; otherwise, it is not.

Employed Bees Phase

Employed bees search for new solutions by modifying the current solutions. They create a new solution V_i by adjusting one or more dimensions (features) in the current solution using the following equation:

$$X_i = X_{i,j} + \phi_{i,j}(X_{i,j} - X_{k,j})$$

Where $\phi_{i,j}$ is a random number between [-1,1]. k is a randomly chosen solution different from i. The new solution is evaluated, and if the new solution is better, it replaces the old one.

Employed Bees Phase

Onlooker bees select solutions based on their fitness and then explore around them. The probability of selecting a solution depends on its fitness:

$$P_i = \frac{f_i}{\sum_{i=1}^{m} f_i}$$

Where f_i is the fitness value of the i-th solution. P_i is the probability of selecting the i-th solution.

Scout Bees Phase

If a solution does not improve after a certain number of iterations (limit), it is abandoned, and a new solution is randomly generated. This helps avoid local optima. A scout bee is then employed to explore new regions in the search space.

Fitness Evaluation

For feature selection, the fitness function evaluates the quality of the selected features. This is usually done using a classification algorithm (e.g., SVM, ANN, RF) and assessing its performance (accuracy, precision, etc.) on the selected feature subset. The fitness could be defined as:

$$f(X_i) = \alpha.Accuracy\ (X_i) + \beta.\frac{1}{|X_i|}$$

Where α and β are weights balancing accuracy and the number of selected features. $|X_i|$ is the number of selected features.

Termination Criteria

The algorithm repeats these phases until a stopping criterion is met, such as a maximum number of iterations or convergence of fitness values.

Result And Discussion

Dataset Description

In this research work, the three different clinical datasets are considered to evaluate the performance of the existing feature selection methods. Dermatology (https://www.kaggle.com/datasets/syslogg/dermatology-dataset), Lung Cancer (https://archive.ics.uci.edu/dataset/62/lung+cancer) and Hepatitis (https://www.kaggle.com/datasets/codebreaker619/hepatitis-data) datasets are considered in this work. Table 1 depicts the number of features in the given considered datasets.

Performance Metrics

Table 2 gives the performance metrics used in this research work, to evaluate the performance of the proposed TTO-FS methods using classification techniques, artificial neural network (ANN), random forest (RF) and support vector machine (SVM). The performance of the existing feature selection methods are evaluated with the existing feature selection techniques like information gain (IG), ReliefF (RFF), whale optimization algorithm (WOA), artificial bee colony optimization (ABO).

Performance Analysis of the Feature Selection Methods for Dermatology Dataset

Table 3 give the number of features obtained by the existing feature selection methods. From Table 3, it is clear that the IG and RFF give less number of features than the existing feature selection methods.

Table 1: Number of features in the considered datasets

Name of the dataset	*Number of features present*
Dermatology	35
Lung cancer	57
Hepatitis	20

Table 2: Performance metrics

Metrics	*Equation*
Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$
True positive rate (TPR) (Sensitivity or Recall)	$\frac{TP}{TP+FN}$
False positive rate (FPR)	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
Specificity	1- False Positive Rate (FPR)
Miss rate	1-True Positive Rate (TPR)
False discovery rate	1- Precision

Table 3: Number of features obtained by the proposed and existing feature selection methods for dermatology dataset

Feature selection techniques	*Number of features present*
Original Dataset	35
Information gain	28
ReliefF	27
Artificial bee colony	34
Whale optimization algorithm	29

Table 4: Classification accuracy (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for dermatology dataset

Feature selection methods	*Classification accuracy (in %)*		
	SVM	*RF*	*ANN*
Original dataset	43.099	46.44	48.32
IG	69.63	69.97	70.84
RFF	66.54	66.86	68.75
ABC	65.46	65.77	67.64
WOA	71.76	72.30	72.87

Table 4 gives the classification accuracy (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques. From Table 4, The original dataset showed the lowest classification accuracies, with 43.10% for SVM, 46.44% for RF, and 48.32% for ANN. Among the feature selection methods, WOA achieved the highest accuracy across all classifiers: 71.76% for SVM, 72.30% for RF, and 72.87% for ANN. IG also provided strong results, with accuracies of 69.63% for SVM, 69.97% for RF, and 70.84% for ANN. RFF method performed moderately, with accuracies of 66.54% for SVM, 66.86% for RF, and 68.75% for ANN. ABC method had slightly lower accuracies compared to RFF, achieving 65.46% for SVM, 65.77% for RF, and 67.64% for ANN.

Table 5 gives the true positive rate (in %) obtained by the proposed and existing feature selection methods using ANN, RF and SVM classification techniques. From Table 5, The Original Dataset produced the lowest True Positive Rates (TPR) across all classifiers, with 52.61% for SVM, 52.94% for RF, and 52.80% for ANN. The WOA method showed the highest TPR for RF (76.37%) and competitive rates for SVM (75.37%) and ANN (70.54%). IG exhibited strong TPR results for SVM (76.07%), RF (74.59%), and ANN (71.35%), making it another high-performing method. RFF) delivered moderate TPR results, with 69.18% for SVM, 67.68% for RF, and 65.45% for ANN. ABC method showed slightly lower TPR values compared to RFF, with 64.34% for SVM, 66.57% for RF, and 68.29% for ANN.

Table 6 gives the false positive rate (in %) obtained by the Existing Feature Selection methods using ANN, RF and SVM classification techniques. From Table 6, The Original Dataset recorded the highest False Positive Rates, with 67.17% for

SVM, 61.08% for RF, and 56.83% for ANN, indicating weaker performance in minimizing false positives. IG also performed well, yielding lower FPR values of 35.62% for SVM, 34.77% for RF, and 29.73% for ANN. RFF method delivered moderate FPRs, with 46.53% for SVM, 45.66% for RF, and 40.82% for ANN. ABC method showed slightly higher FPRs compared to RFF, with 47.42% for SVM, 46.75% for RF, and 41.71% for ANN.

Table 7 gives the precision (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques. From the Table 7, The Original Dataset showed the lowest precision across all classifiers, with 45.81% for SVM, 49.01% for RF, and 51.72% for ANN. WOA achieved the highest precision rates across all classifiers, with 71.97% for SVM, 71.45% for RF, and 78.97% for ANN, demonstrating superior performance. IG also showed strong precision results, with 68.79% for SVM, 68.81% for RF, and 73.60% for ANN. RFF method delivered moderate precision values, achieving 59.68% for SVM, 59.72% for RF, and 62.51% for ANN. ABC method had slightly lower precision rates compared to RFF, with 58.57% for SVM, 58.61% for RF, and 61.43% for ANN.

Table 8 gives the miss rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques. From Table 8, The original dataset recorded the highest miss rates across all classifiers, with 47.39% for SVM, 47.06% for RF, and 47.20% for ANN, indicating the poorest performance in minimizing missed detections. WOA achieved the lowest miss rates for SVM (24.63%) and RF (23.63%), while it performed moderately for ANN (29.46%). IG also delivered strong results, with miss rates of 23.93% for SVM, 25.41% for RF, and 28.65% for ANN. RFF method produced moderate miss rates, with 32.82% for SVM, 36.52% for RF, and 39.76% for ANN. ABC method showed slightly higher miss rates compared to RFF, with 33.91% for SVM, 37.61% for RF, and 40.85% for ANN.

Table 9 gives the specificity (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques. From Table 9, The original dataset showed the lowest specificity across all classifiers, with 32.83% for SVM, 38.92% for RF, and 43.17% for ANN, indicating weaker performance in correctly identifying negative cases. WOA achieved the highest specificity across all classifiers, with 67.82% for SVM, 67.20% for RF, and 75.78% for ANN, demonstrating strong performance. IG also performed well, with specificity values of 64.38% for SVM, 65.23% for RF, and 70.27% for ANN. The RFF method

Table 5: True Positive Rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for dermatology dataset

Feature selection methods	*True positive rate (in %)*		
	SVM	*RF*	*ANN*
Original dataset	52.61	52.94	52.80
IG	76.07	74.59	71.35
RFF	69.18	67.68	65.45
ABC	64.34	66.57	68.29
WOA	75.37	76.37	70.54

Table 6: False positive rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for dermatology dataset

Feature selection methods	*False Positive Rate (in %)*		
	SVM	*RF*	*ANN*
Original dataset	67.17	61.08	56.83
IG	35.62	34.77	29.73
RFF	46.53	45.66	40.82
ABC	47.42	46.75	41.71
WOA	32.18	32.8	24.22

Table 7: Precision (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for dermatology dataset

Feature selection methods	*Precision (in %)*		
	SVM	*RF*	*ANN*
Original dataset	45.81	49.01	51.72
IG	68.79	68.81	73.60
RFF	59.68	59.72	62.51
ABC	58.57	58.61	61.43
WOA	71.97	71.45	78.97

Table 8: Miss rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for dermatology dataset

Feature selection methods	*Miss rate (in %)*		
	SVM	*RF*	*ANN*
Original dataset	47.39	47.06	47.2
IG	23.93	25.41	28.65
RFF	32.82	36.52	39.76
ABC	33.91	37.61	40.85
WOA	24.63	23.63	29.46

Table 9: Specificity (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for dermatology dataset

Feature selection methods	*Specificity (in %)*		
	SVM	*RF*	*ANN*
Original dataset	32.83	38.92	43.17
IG	64.38	65.23	70.27
RFF	53.49	54.32	59.38
ABC	52.38	53.21	58.24
WOA	67.82	67.2	75.78

delivered moderate specificity, with 53.49% for SVM, 54.32% for RF, and 59.38% for ANN. ABC method showed slightly lower specificity compared to RFF, achieving 52.38% for SVM, 53.21% for RF, and 58.24% for ANN.

Performance Analysis of the Feature Selection Methods for Lung Cancer Dataset

Table 10 give the number of features obtained by the Proposed TTO-FS method and existing feature selection methods. From Table 10, it is clear that the WOA gives a smaller number of features than the existing feature selection methods.

Table 11 gives the classification accuracy (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for lung cancer dataset. From Table 12, The original dataset showed the lowest classification accuracy across all classifiers, with 43.97% for SVM, 44.98% for RF, and 48.32% for ANN. WOA achieved the highest accuracy across all classifiers, with 71.67% for SVM, 71.47% for RF, and 72.59% for ANN, making it the best-performing feature selection method. IG also demonstrated strong performance, with accuracies of 69.34% for SVM, 70.94% for RF, and 70.84% for ANN. RFF method showed moderate accuracy, achieving 58.43% for SVM, 59.85% for RF, and 59.73% for ANN. ABC method had slightly lower accuracies compared to RFF, with 57.34% for SVM, 58.74% for RF, and 58.64% for ANN.

Table 12 gives the true positive rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for the lung cancer dataset. From Table 12, The original dataset recorded the lowest TPR across all classifiers, with 51.26% for SVM, 47.68% for RF, and 52.76% for ANN. WOA achieved the highest TPR for SVM (82.30%) and performed well with RF (74.90%) and ANN (71.19%), making it the top performer for SVM. IG also delivered strong TPR results, achieving 73.05% for SVM, 75.50% for RF, and 74.45% for ANN. RFF method showed moderate TPR values, with 62.16% for SVM, 64.41% for RF, and 63.34% for ANN. ABC method had slightly lower TPR compared to RFF, with 61.27% for SVM, 63.32% for RF, and 62.25% for ANN.

Table 13 gives the false positive rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for lung cancer dataset, from Table 13. The original dataset produced the highest False Positive Rates (FPR) across all classifiers, with 63.80% for SVM, 57.67% for RF, and 56.58% for ANN, indicating the poorest performance in minimizing false positives. WOA achieved the lowest FPR across all classifiers, with 31.91% for SVM, 32.31% for RF, and 25.60% for ANN, demonstrating its effectiveness in reducing false positives. IG also performed well, yielding FPR values of 35.31% for SVM, 33.75% for RF, and 32.87% for ANN. The RFF method delivered moderate FPRs, achieving 44.42% for SVM, 42.84% for RF, and 43.78% for ANN. ABC method had slightly higher FPRs compared to RFF, with 45.53% for SVM, 43.75% for RF, and 44.69% for ANN.

Table 14 gives the precision (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for the lung cancer dataset. From Table 14, The original dataset showed the lowest

Table 10: Number of Features obtained by the Proposed and Existing Feature Selection methods for Lung Cancer Dataset

Feature selection techniques	*Number of features present*
Original dataset	57
Information gain	48
ReliefF	46
Artificial bee colony	51
Whale optimization algorithm	45

Table 11: Classification accuracy (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for lung cancer dataset

Feature selection methods	*Classification accuracy (in %)*		
	SVM	*RF*	*ANN*
Original Dataset	43.97	44.98	48.32
IG	69.34	70.94	70.84
RFF	58.43	59.85	59.73
ABC	57.34	58.74	58.64
WOA	71.67	71.47	72.59

Table 12: True positive rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for lung cancer dataset

Feature selection methods	*True positive rate (in %)*		
	SVM	*RF*	*ANN*
Original Dataset	51.26	47.68	52.76
IG	73.05	75.50	74.45
RFF	62.16	64.41	63.34
ABC	61.27	63.32	62.25
WOA	82.3	74.90	71.19

Table 13: False positive rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for Lung Cancer dataset

Feature selection methods	*False positive rate (in %)*		
	SVM	*RF*	*ANN*
Original dataset	63.8	57.67	56.58
IG	35.31	33.75	32.87
RFF	44.42	42.84	43.78
ABC	45.53	43.75	44.69
WOA	31.91	32.31	25.60

precision across all classifiers, with 46.11% for SVM, 52.34% for RF, and 50.84% for ANN, indicating poor performance in correctly identifying positive cases. WOA achieved the highest precision across all classifiers, with 72.76% for SVM, 71.92% for RF, and 78.18% for ANN, demonstrating superior performance. IG also performed well, delivering precision values of 69.21% for SVM, 69.77% for RF, and 70.04% for ANN. RFF method produced moderate precision results, with 58.32% for SVM, 58.68% for RF, and 61.13% for ANN. ABC method had slightly lower precision compared to RFF, with 57.43% for SVM, 57.79% for RF, and 60.24% for ANN.

Table 15 gives the miss rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for the lung cancer dataset. From Table 15, The original dataset showed the highest Miss Rates across all classifiers, with 48.74% for SVM, 52.32% for RF, and 47.24% for ANN, indicating a higher rate of missed positive cases. WOA achieved the lowest miss rate for SVM (17.7%) and performed moderately for RF (25.1%) and ANN (28.81%), showing significant improvement in minimizing missed detections. IG also delivered strong results, reducing the miss rate to 29.65% for SVM, 24.5% for RF, and 25.55% for ANN. RFF method showed moderate Miss Rates, with 38.54% for SVM, 35.56% for RF, and 36.67% for ANN. ABC method had slightly higher Miss Rates compared to RFF, with 39.45% for SVM, 36.67% for RF, and 37.78% for ANN.

Table 16 gives the specificity (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for the lung cancer dataset.

Table 14: Precision (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for lung cancer dataset

Feature selection methods	*Precision (in %)*		
	SVM	*RF*	*ANN*
Original dataset	46.11	52.34	50.84
IG	69.21	69.77	70.04
RFF	58.32	58.68	61.13
ABC	57.43	57.79	60.24
WOA	72.76	71.92	78.18

Table 15: Miss rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for lung cancer dataset

Feature selection methods	*Miss Rate (in %)*		
	SVM	*RF*	*ANN*
Original dataset	48.74	52.32	47.24
IG	29.65	24.5	25.55
RFF	38.54	35.56	36.67
ABC	39.45	36.67	37.78
WOA	17.7	25.1	28.81

From Table 16, The original dataset recorded the lowest specificity across all classifiers, with 36.2% for SVM, 42.33% for RF, and 43.42% for ANN, indicating poor performance in correctly identifying negative cases. WOA achieved the highest specificity across all classifiers, with 68.09% for SVM, 67.69% for RF, and 74.4% for ANN, making it the best-performing feature selection method. IG also performed well, with specificity values of 64.91% for SVM, 66.25% for RF, and 67.13% for ANN. RFF method provided moderate specificity results, achieving 55.82% for SVM, 55.34% for RF, and 56.24% for ANN. ABC method showed slightly lower specificity compared to RFF, with 54.71% for SVM, 54.45% for RF, and 55.35% for ANN.

Performance Analysis of the Proposed TTO-FS Method for Hepatitis Dataset

Table 17 give the number of features obtained by the existing feature selection methods. From, Table 17, it is clear that the IG and WOA methods gives less number of features than the existing feature selection methods.

Table 18 gives the classification accuracy (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for the Hepatitis dataset, from Table 18. The original dataset exhibited the lowest classification accuracy across all classifiers, with 44.93% for SVM, 45.81% for RF, and 50.16% for ANN, indicating limited effectiveness in predicting outcomes. WOA achieved the highest accuracy for both SVM and ANN at 74.04%, and 72.20% for RF, demonstrating its superior performance among the feature selection methods. IG also produced strong results, with accuracy rates of 68.81% for SVM,

Table 16: Specificity (in %) obtained by the Existing Feature Selection methods using ANN, RF and SVM classification techniques for Lung Cancer dataset

Feature selection Methods	*Specificity (in %)*		
	SVM	*RF*	*ANN*
Original Dataset	36.2	42.33	43.42
IG	64.91	66.25	67.13
RFF	55.82	55.34	56.24
ABC	54.71	54.45	55.35
WOA	68.09	67.69	74.4

Table 17: Number of features obtained by the proposed and existing feature selection methods for hepatitis dataset

Feature selection techniques	*Number of features present*
Original dataset	20
Information gain	14
ReliefF	15
Artificial bee colony	18
Whale optimization algorithm	14

67.11% for RF, and 66.19% for ANN, indicating effective feature selection. RFF method showed moderate accuracy, achieving 57.92% for SVM, 58.22% for RF, and 55.28% for ANN. ABC method yielded slightly lower accuracy rates compared to RFF, with 56.81% for SVM, 57.32% for RF, and 54.19% for ANN.

Table 19 gives the true positive rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for Hepatitis dataset. From the Table 19, the original dataset recorded the lowest TPR across all classifiers, with 55.11% for SVM, 49.44% for RF, and 54.26% for ANN, indicating a lower ability to correctly identify positive cases. WOA achieved the highest TPR, with 84.55% for SVM, 80.57% for RF, and 81.74% for ANN, showcasing its effectiveness in enhancing the identification of true positives. IG also showed strong performance, with TPR values of 67.35% for SVM, 69.42% for RF, and 70.57% for ANN, indicating improved positive case identification. RFF method provided moderate TPR results, achieving 56.43% for SVM, 58.31% for RF, and 60.46% for ANN. ABC method had slightly lower TPRs compared to RFF, with 55.65% for SVM, 57.53% for RF, and 59.68% for ANN.

Table 20 gives the false positive rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for the hepatitis dataset. From Table 20, the original dataset exhibited the highest false positive rates across all classifiers, with 64.74% for SVM, 59.04% for RF, and 54.40% for ANN, indicating a higher incidence of incorrectly identified positive cases. WOA) achieved the lowest FPR across all classifiers, with 35.76% for SVM, 36.21% for RF, and 34.56% for ANN, demonstrating its effectiveness in minimizing false positives. IG also showed strong results, with FPRs of 28.79% for SVM, 35.32% for RF, and 38.08% for ANN, indicating effective feature selection. RFF method produced moderate FPR values, achieving 37.88% for SVM, 36.43% for RF, and 39.19% for ANN. ABC method had slightly higher FPRs compared to RFF, with 38.06% for SVM, 37.65% for RF, and 40.32% for ANN.

Table 21 gives the precision (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for the Hepatitis dataset. From Table 21, the original dataset recorded the lowest precision across all classifiers, with 44.75% for SVM, 52.80% for RF, and 52.67% for ANN, indicating a lower accuracy in correctly identifying positive cases. IG method demonstrated strong precision, achieving 74.80% for SVM, 67.58% for RF, and 64.97% for ANN, showcasing its effectiveness in enhancing precision. WOA also performed well, with precision values of 68.81% for SVM, 69.09% for RF, and 72.55% for ANN, indicating a high level of accuracy in positive case identification. The RFF method showed moderate precision results, with 63.91% for SVM, 57.47% for RF, and 53.86% for ANN. ABC method had slightly lower precision compared to RFF, achieving 61.13% for SVM, 56.69% for RF, and 52.08% for ANN.

Table 22 gives the miss rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for Hepatitis dataset. From Table 22, the original dataset exhibited the highest Miss Rates across

Table 18: Classification accuracy (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for hepatitis dataset

Feature selection methods	*Classification accuracy (in %)*		
	SVM	*RF*	*ANN*
Original Dataset	44.93	45.81	50.16
IG	68.81	67.11	66.19
RFF	57.92	58.22	55.28
ABC	56.81	57.32	54.19
WOA	74.04	72.20	74.04

Table 19: True positive rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for hepatitis dataset

Feature selection methods	*True positive rate (in %)*		
	SVM	*RF*	*ANN*
Original dataset	55.11	49.44	54.26
IG	67.35	69.42	70.57
RFF	56.43	58.31	60.46
ABC	55.65	57.53	59.68
WOA	84.55	80.57	81.74

Table 20: False positive rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for Hepatitis dataset

Feature selection methods	*False positive rate (in %)*		
	SVM	*RF*	*ANN*
Original dataset	64.74	59.04	54.40
IG	28.79	35.32	38.08
RFF	37.88	36.43	39.19
ABC	38.06	37.65	40.32
WOA	35.76	36.21	34.56

Table 21: Precision (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for the Hepatitis dataset

Feature selection methods	*Precision (in %)*		
	SVM	*RF*	*ANN*
Original dataset	44.75	52.80	52.67
IG	74.80	67.58	64.97
RFF	63.91	57.47	53.86
ABC	61.13	56.69	52.08
WOA	68.81	69.09	72.55

Table 22: Miss rate (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for Hepatitis dataset

Feature selection methods	*Miss rate (in %)*		
	SVM	*RF*	*ANN*
Original dataset	44.89	50.56	45.74
IG	32.65	30.58	29.43
RFF	41.57	41.69	39.42
ABC	42.79	42.81	40.64
WOA	15.45	19.43	18.26

Table 23: Specificity (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for Hepatitis dataset

Feature selection methods	*Specificity (in %)*		
	SVM	*RF*	*ANN*
Original dataset	35.26	40.96	45.6
IG	71.21	64.68	61.92
RFF	60.12	55.79	50.81
ABC	59.35	56.91	49.05
WOA	64.24	63.79	65.44

all classifiers, with 44.89% for SVM, 50.56% for RF, and 45.74% for ANN, indicating a substantial number of missed positive cases. WOA achieved the lowest miss rate across all classifiers, with 15.45% for SVM, 19.43% for RF, and 18.26% for ANN, demonstrating its effectiveness in correctly identifying positive cases. IG also showed strong performance, with miss rates of 32.65% for SVM, 30.58% for RF, and 29.43% for ANN, indicating significant improvement in positive case identification. RFF method produced moderate miss rates, achieving 41.57% for SVM, 41.69% for RF, and 39.42% for ANN. ABC method had similar results to RFF, with Miss Rates of 42.79% for SVM, 42.81% for RF, and 40.64% for ANN.

Table 23 gives the specificity (in %) obtained by the existing feature selection methods using ANN, RF and SVM classification techniques for Hepatitis dataset. From Table 23, the original dataset exhibited the lowest specificity across all classifiers, with 35.26% for SVM, 40.96% for RF, and 45.60% for ANN, indicating limited effectiveness in correctly identifying negative cases. IG method achieved the highest specificity, with values of 71.21% for SVM, 64.68% for RF, and 61.92% for ANN, demonstrating its capability to enhance the identification of true negatives. WOA also performed well, yielding specificity values of 64.24% for SVM, 63.79% for RF, and 65.44% for ANN, indicating its effectiveness in identifying negative cases. RFF method produced moderate specificity results, achieving 60.12% for SVM, 55.79% for RF, and 50.81% for ANN. ABC method had slightly lower specificity compared to RFF, with 59.35% for SVM, 56.91% for RF, and 49.05% for ANN.

Conclusion

The results and discussions presented throughout this analysis demonstrate the significant impact of feature selection methods on the classification performance of various machine learning techniques, specifically in the context of the Hepatitis dataset. The findings indicate that feature selection plays a crucial role in improving the effectiveness of classifiers by enhancing metrics such as classification accuracy, true positive rate, precision, and specificity while reducing miss rate and false positive rate.

Among the feature selection methods evaluated, the whale optimization algorithm (WOA) emerged as the most effective approach across multiple classification metrics, achieving high accuracy and low miss rates and false positive rates. This suggests that WOA not only facilitates better identification of positive cases but also minimizes incorrect classifications. Similarly, information gain (IG) also demonstrated strong performance, significantly improving True Positive Rate and Precision while maintaining reasonable specificity levels.

In contrast, the original dataset consistently showed the poorest performance across all evaluated metrics, highlighting the necessity of employing robust feature selection techniques to enhance model performance. Other methods, such as random forest feature (RFF) and artificial bee colony (ABC), while showing moderate effectiveness, did not match the performance levels achieved by WOA and IG.

Overall, the findings underscore the importance of carefully selecting appropriate feature selection methods in machine learning workflows. By leveraging advanced feature selection techniques, practitioners can significantly enhance the predictive performance of classifiers, ultimately leading to more accurate and reliable decision-making in applications such as disease diagnosis, including Hepatitis. Future work should explore further optimizations and additional datasets to validate the robustness of these findings and to investigate the applicability of these feature selection methods across different domains.

References

Alwateer, M., Almars, A. M., Areed, K. N., Elhosseini, M. A., Haikal, A. Y., & Badawy, M. (2021). Ambient healthcare approach with hybrid whale optimization algorithm and Naïve Bayes classifier. *Sensors*, 21(13), 4579.

Arvindhan, M., Rajeshkumar, D., & Pal, A. L. (2021). A review of challenges and opportunities in machine learning for healthcare. *Exploratory Data Analytics for Healthcare*, 67-84.

Atteia, G., El-kenawy, E. S. M., Samee, N. A., Jamjoom, M. M., Ibrahim, A., Abdelhamid, A. A., ... & Shams, M. Y. (2023). Adaptive dynamic dipper throated optimization for feature selection in medical data. *Computers, Materials & Continua*, 75(1), 1883-1900.

Bennett, M., Hayes, K., Kleczyk, E. J., & Mehta, R. (2022). Similarities and differences between machine learning and traditional advanced statistical modeling in healthcare analytics. *arXiv*

preprint arXiv:2201.02469.

Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., ... & Moni, M. A. (2023). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. *BioMed Research International*, 2023(1), 6864343.

Durairaj, M., & Poornappriya, T. S. (2020). Why feature selection in data mining is prominent? A survey. *In Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications: AISGSC 2019* (pp. 949-963). Springer International Publishing.

García-Domínguez, A., Galván-Tejada, C. E., Magallanes-Quintanar, R., Gamboa-Rosales, H., Curiel, I. G., Peralta-Romero, J., & Cruz, M. (2023). Diabetes Detection Models in Mexican Patients by Combining Machine Learning Algorithms and Feature Selection Techniques for Clinical and Paraclinical Attributes: A Comparative Evaluation. *Journal of Diabetes Research*, 2023(1), 9713905.

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, 9, 19304-19326.

Grisci, B. I., Feltes, B. C., de Faria Poloni, J., Narloch, P. H., & Dorn, M. (2024). The use of gene expression datasets in feature selection research: 20 years of inherent bias?. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2), e1523.

Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current genomics*, 22(4), 291.

https://archive.ics.uci.edu/dataset/62/lung+cancer

https://www.kaggle.com/datasets/codebreaker619/hepatitis-data

https://www.kaggle.com/datasets/syslogg/dermatology-dataset

Kumari, J., Kumar, E., & Kumar, D. (2023). A structured analysis to study the role of machine learning and deep learning in the healthcare sector with big data analytics. *Archives of Computational Methods in Engineering*, 30(6), 3673-3701.

Liu, J., Zhao, L., Si, C., Guan, H., & Dong, X. (2023). Improved ReliefF-based feature selection algorithm for cancer histology. *Biomedical Signal Processing and Control*, 85, 104980.

Mahto, R., Ahmed, S. U., Rahman, R. U., Aziz, R. M., Roy, P., Mallik, S., ... & Shah, M. A. (2023). A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection. *BMC bioinformatics*, 24(1), 479.

Manikandan, G., Pragadeesh, B., Manojkumar, V., Karthikeyan, A. L., Manikandan, R., & Gandomi, A. H. (2024). Classification models combined with Boruta feature selection for heart disease prediction. *Informatics in Medicine Unlocked*, 44, 101442.

Masood, F., Masood, J., Zahir, H., Driss, K., Mehmood, N., & Farooq, H. (2023). Novel approach to evaluate classification algorithms and feature selection filter algorithms using medical data. *Journal of Computational and Cognitive Engineering*, 2(1), 57-67.

Mostafa, R. R., Khedr, A. M., Al Aghbari, Z., Afyouni, I., Kamel, I., & Ahmed, N. (2024). An adaptive hybrid mutated differential evolution feature selection method for low and high-dimensional medical datasets. *Knowledge-Based Systems*, 283, 111218.

Nagarajan, S. M., Muthukumaran, V., Murugesan, R., Joseph, R. B., & Munirathanam, M. (2021). Feature selection model for healthcare analysis and classification using classifier ensemble technique. *International Journal of System Assurance Engineering and Management*, 1-12.

Nerkar, P. M., Liyakat, K. K. S., Dhaware, B. U., & Liyakat, K. S. S. (2023). Predictive Data Analytics Framework Based on Heart Healthcare System (HHS) Using Machine Learning. *Journal of Advanced Zoology*, 44, 3673-3686.

Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Scientific Reports*, 13(1), 22588.

Patra, S. S., Harshvardhan, G. M., Gourisaria, M. K., Mohanty, J. R., & Choudhury, S. (2021). Emerging healthcare problems in high-dimensional data and dimension reduction. *Advanced Prognostic Predictive Modelling in Healthcare Data Analytics*, 25-49.

Pham, T. H., & Raahemi, B. (2023). Bio-inspired feature selection algorithms with their applications: a systematic literature review. *IEEE Access*, 11, 43733-43758.

Ramasamy, M., & Meena Kowshalya, A. (2022). Information gain based feature selection for improved textual sentiment analysis. *Wireless Personal Communications*, 125(2), 1203-1219.

Razzaque, A., & Badholia, A. (2024). PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification. *Measurement: Sensors*, 31, 100945.

Riyahi, M., Rafsanjani, M. K., Gupta, B. B., & Alhalabi, W. (2022). Multiobjective whale optimization algorithm-based feature selection for intelligent systems. *International Journal of Intelligent Systems*, 37(11), 9037-9054.

Salazar Reyna, R. J. (2019). Systematic Literature Review of Data Science, Data Analytics and Machine Learning Applied into Healthcare Engineering Systems.

Sharma, A., & Mishra, P. K. (2022). Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. *International Journal of Information Technology*, 14(4), 1949-1960.

Veena, A., & Gowrishankar, S. (2021). Healthcare analytics: Overcoming the barriers to health information using machine learning algorithms. *In Image Processing and Capsule Networks: ICIPCN 2020* (pp. 484-496). Springer International Publishing.

Vommi, A. M., & Battula, T. K. (2023). A binary Bi-phase mutation-based hybrid Equilibrium Optimizer for feature selection in medical datasets classification. *Computers and Electrical Engineering*, 105, 108553.

Vommi, A. M., & Battula, T. K. (2023). A hybrid filter-wrapper feature selection using Fuzzy KNN based on Bonferroni mean for medical datasets classification: A COVID-19 case study. *Expert Systems with Applications*, 218, 119612.