Indian Journal of Natural Sciences *www.tnsroindia.org.in ©IJONS*

Vol.15 / Issue 86 / Oct / 2024 *International Bimonthly (Print) – Open Access* *ISSN: 0976 – 0997*

RESEARCH ARTICLE

A Thrice Filtered Information Energy Optimization Based Feature Selection (TFIE-OFS) Method for Heart Disease Prediction

S. Vanaja[1]* and Hari Ganesh S[2]

[1]Research Scholar, PG & Research Department of Computer Science, H.H. The Rajah's College (Autonomous), Pudukkottai, (Affiliated to Bharathidasan University, Tiruchirappalli), Tamil Nadu, India. [2]Assistant Professor, PG & Research Department of Computer Science, H.H. The Rajah's College (Autonomous), Pudukkottai, (Affiliated to Bharathidasan University, Tiruchirappalli), Tamil Nadu, India.

Received: 21 Aug 2024 Revised: 03 Jul 2024 Accepted: 26 Oct 2024

***Address for Correspondence**
S. Vanaja
Research Scholar, PG & Research Department of Computer Science,
H.H. The Rajah's College (Autonomous), Pudukkottai,
(Affiliated to Bharathidasan University, Tiruchirappalli),
Tamil Nadu, India.

ABSTRACT

Heart disease remains a leading cause of mortality worldwide, necessitating effective classification and prediction methods to enhance early detection and intervention. This study proposes a novel Thrice Filtered Information Energy Optimization based Feature Selection (TFIE-OFS) method, which integrates Symmetrical Uncertainty, Information Gain, and Chi-Square Analysis to systematically filter and prioritize features from heart disease datasets. By employing Particle Swarm Optimization (PSO), the TFIE-OFS method optimizes feature subsets, ensuring the selection of the most informative variables while minimizing redundancy. The efficacy of the proposed method is evaluated through comprehensive experiments on benchmark heart disease datasets, where it demonstrates superior classification performance compared to existing feature selection techniques. The results indicate that TFIE-OFS significantly enhances predictive accuracy and model interpretability, providing a robust framework for heart disease classification and prediction. This innovative approach not only contributes to the field of medical data analytics but also holds potential for improving clinical decision-making in cardiology.

Keywords: Heart Disease, Classification, Feature Selection, Symmetrical Uncertainty, Information Gain, Particle Swarm Optimization, Chi-Square

INTRODUCTION

Heart disease has emerged as one of the foremost health challenges of the 21st century, contributing significantly to global morbidity and mortality rates [1] [2]. According to the World Health Organization, cardiovascular diseases account for approximately 31% of all global deaths, highlighting the urgent need for effective diagnostic and predictive tools. Early detection and accurate prediction of heart disease are critical for implementing timely interventions and improving patient outcomes. However, the complexity of heart disease risk factors and the vast amounts of health data pose significant challenges in developing robust predictive models [3] [4] [5]. Feature selection plays a crucial role in the classification and prediction of heart disease by identifying the most relevant variables that contribute to the condition. Traditional methods of feature selection often face limitations, such as high computational costs, redundancy among selected features, and the inability to capture complex relationships within the data. Therefore, an efficient and effective feature selection technique is essential for enhancing the performance of predictive models in this domain.

This study proposes a novel Thrice Filtered Information Energy Optimization based Feature Selection (TFIE-OFS) method, which leverages a combination of Symmetrical Uncertainty, Information Gain, and Chi-Square Analysis to filter and prioritize features systematically. By employing these three complementary approaches, TFIE-OFS captures various aspects of feature importance while mitigating the influence of irrelevant and redundant variables. Additionally, the TFIE-OFS method incorporates Particle Swarm Optimization (PSO) to enhance the selection process further. PSO is an intelligent optimization technique inspired by social behavior in animals, such as bird flocking. By mimicking this behavior, PSO effectively searches for optimal feature subsets by balancing exploration and exploitation within the feature space [6] [7] [8].

The primary objective of this research is to develop an effective and efficient feature selection methodology that can improve the classification and prediction accuracy of heart disease models. Through rigorous experimentation on benchmark heart disease datasets, we aim to demonstrate the superiority of TFIE-OFS over existing feature selection methods. In summary, this introduction outlines the critical importance of heart disease prediction, the challenges associated with feature selection, and the innovative approach proposed in this study. By combining multiple feature selection techniques and optimization algorithms, TFIE-OFS offers a promising solution for enhancing predictive modeling in cardiovascular health. This research not only contributes to the field of medical data analytics but also holds significant implications for clinical decision-making, ultimately leading to improved patient care and outcomes in heart disease management.

Background Study on Feature Selection Methods

Feature selection [9] [10] is a critical process in machine learning and data mining that involves identifying and selecting a subset of relevant features (or variables) for use in model construction. This process is particularly essential in the field of medical data analysis, where the complexity of the data can lead to overfitting, increased computational costs, and diminished interpretability of predictive models. In this background study, we explore various feature selection methods, their significance, and their application in the context of heart disease classification and prediction.

Types of Feature Selection Methods

Feature selection methods can be broadly classified into three categories: filter methods, wrapper methods, and embedded methods. Each category employs different strategies for selecting relevant features.

Filter Methods

Filter methods assess the relevance of features based on intrinsic properties of the data, independent of any machine learning algorithm. They are typically computationally efficient and suitable for high-dimensional datasets. Common filter methods include:

Correlation Coefficient: Measures the statistical relationship between features and the target variable. High correlation indicates potential relevance.

Chi-Square Test: Evaluates the independence between categorical features and the target variable, identifying significant associations.
Information Gain: Measures the reduction in uncertainty about the target variable given knowledge of a feature. It quantifies how much information a feature contributes to the prediction.

Wrapper Methods
Wrapper methods evaluate feature subsets by training a specific model and assessing its performance. They are typically more accurate than filter methods but can be computationally expensive due to the repeated model training required. Common wrapper methods include:

Recursive Feature Elimination (RFE): Iteratively removes the least significant features based on model performance until the desired number of features is reached.

Forward Selection: Starts with an empty set of features and adds them one by one, evaluating model performance at each step to determine the best feature to add.

Backward Elimination: Begins with all features and removes them one at a time, selecting the least significant feature based on model performance.

Embedded Methods
Embedded methods combine feature selection and model training into a single process. They identify relevant features while the model is being trained, making them more efficient than wrapper methods. Common embedded methods include:

Lasso Regression: Uses L1 regularization to penalize the absolute size of coefficients, effectively shrinking some to zero, thus performing feature selection.

Decision Trees and Random Forests: These algorithms naturally perform feature selection by considering the importance of features in splitting the data during tree construction.

Gradient Boosting Machines: These models can also provide feature importance scores, allowing for the selection of significant features based on their contributions to the model.

Symmetrical Uncertainty (SU) Based Feature Selection Method
Symmetrical Uncertainty (SU) [11] [12] is a feature selection method that quantifies the amount of information gained about one variable through another, balancing the measure of uncertainty in both variables. It is particularly useful in the context of categorical variables and has become a popular choice in various machine learning applications, including medical diagnostics, where the interpretation of results is crucial. SU provides a normalized measure that ranges from 0 to 1, facilitating the comparison of feature relevance across different datasets.Symmetrical Uncertainty is derived from the concept of mutual information, which measures the amount of information that knowing the value of one variable provides about another. SU is defined mathematically as follows: Symmetrical Uncertainty is derived from the concept of mutual information, which measures the amount of information that knowing the value of one variable provides about another. SU is defined mathematically as follows:

$$SU\ (\mathrm{X}, Y) = \frac{2.I(\mathrm{X}, Y)}{H(\mathrm{X}) + H(Y)}$$

Where $I(X, Y)$ is the mutual information between variables X and Y. $H(X)$ and $H(Y)$ are the entropy of variables X and Y, respectively. Mutual Information (I): This metric quantifies the reduction in uncertainty of one variable due to the knowledge of another. It is calculated as:

$$I(X, Y) = H(X) + H(Y) - H(X, Y)$$

Entropy (H): This measures the unpredictability or randomness of a variable. It is calculated using the probability distribution of the variable:

$$H(X) = -\sum_{x \in X} P(x) log P(x)$$

Information Gain (IG) Based Feature Selection Method

Information Gain (IG) [13] [14]is a widely used metric in feature selection and decision tree algorithms that measures the effectiveness of a feature in reducing uncertainty about the target variable. It quantifies the amount of information that knowing the value of a feature provides about the target outcome. IG is particularly beneficial in classification tasks, including medical diagnoses, where understanding the relationship between features and outcomes is essential for developing predictive models. Information Gain is based on the concept of entropy, which measures the unpredictability or randomness of a variable. The IG of a feature is calculated by comparing the entropy of the target variable before and after the dataset is split by that feature. Mathematically, Information Gain is defined as:

$$IG(T, A) = H(T) - H(T|A)$$

Where $IG(T, A)$is the information gain of feature A with respect to target variable T.$H(T)$is the entropy of the target variable before the split.$H(T|A)$is the conditional entropy of the target variable after the dataset is split based on feature A.

Entropy is calculated using the formula:

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$$

Where$P(x)$is the probability of occurrence of value x.

Chi-Square Analysis Based Feature Selection Method

Chi-Square Analysis [15] [16]is a statistical method used to determine the independence between categorical variables. In the context of feature selection, the Chi-Square test assesses the relationship between each feature and the target variable to identify significant predictors. This method is particularly valuable in classification tasks, especially in medical datasets where categorical variables are prevalent. By evaluating how well each feature correlates with the target outcome, Chi-Square Analysis helps improve model performance and interpretability. The Chi-Square test compares the observed frequencies of occurrences in a contingency table with the expected frequencies under the assumption of independence. The Chi-Square statistic is calculated using the following formula:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$x^2$is the Chi-Square statistic, $O_i$is the observed frequency for category i, E_i is the expected frequency for the category i.

The expected frequency is calculated as:

$$E_i = \frac{(row\ total \times column\ total)}{grand\ total}$$

The resulting Chi-Square statistic indicates how much the observed counts deviate from the expected counts. A higher Chi-Square value suggests a stronger association between the feature and the target variable.

Particle Swarm Optimization Algorithm

Particle Swarm Optimization (PSO)[17] [18] is a nature-inspired optimization algorithm developed by James Kennedy and Russell Eberhart in 1995. It is based on the social behavior of birds and fish, where individuals

(particles) in a swarm collaborate to find optimal solutions within a search space. PSO is particularly effective for solving complex optimization problems, including those in machine learning, engineering design, and parameter tuning. Its simplicity, ease of implementation, and ability to converge to global optima make it a popular choice in various applications.
In PSO, each particle represents a potential solution in the search space and has two primary characteristics:

Position: The current location of the particle, representing a possible solution to the optimization problem.

Velocity: The rate of change of the particle's position, determining how the particle moves through the search space. Particles adjust their positions based on their own experience and that of their neighbors, balancing exploration (searching new areas) and exploitation (refining existing solutions).The PSO algorithm consists of the following key steps:

Initialization
Define the optimization problem and its objective function.
Initialize a swarm of particles with random positions and velocities within the defined search space.
Set parameters such as the number of particles, maximum iterations, and coefficients for cognitive and social components.

Fitness Evaluation
Evaluate the fitness of each particle by calculating the objective function's value at its current position.

Update Personal and Global Bests
For each particle, compare its fitness with its personal best (the best position it has encountered so far) and update it if the current position is better.
Determine the global best (the best position encountered by any particle in the swarm) based on fitness evaluations.

Update Velocity and Position
Adjust each particle's velocity using the following formula:$v_i = \omega . v_i + C_1 . r_1 . (p_i - x_i) + C_2 . r_2 . (g - x_i)$where $v_i$is the particle's current velocity.$\omega$is the inertia weight that controls exploration versus exploitation.C_1 and $C_2$are acceleration coefficients (typically set between 1.5 and 2). r_1 and $r_2$are random numbers between 0 and 1. $p_i$is the personal best position of particle i. g is the global best position. Update the particle's position using: $x_i = x_i + v_i$

Iteration
Repeat steps 2 to 4 until a stopping criterion is met (e.g., maximum iterations or a satisfactory fitness level).

Output
Return the global best position and its corresponding fitness value as the optimal solution.

A Thrice Filtered Information Energy Optimization Based Feature Selection (TFIE-OFS) Method
The following are the step-by-step procedure for the proposed TFIE-OFS method.

Step 1: Data Preprocessing
Data Collection: Gather the dataset for the prediction or classification task (e.g., heart disease dataset).

Handle Missing Values: Impute missing values using techniques such as mean, median, or mode, or remove entries with significant missing data.

Feature Scaling & Encoding : Normalize numerical features using scaling techniques (e.g., Min-Max Scaling or Z-Score normalization). Convert categorical features into numerical format using label encoding or one-hot encoding.

Split the Dataset: Divide the dataset into training and testing sets. The training set will be used for feature selection and model building, while the test set will be used to evaluate performance.

Step 2: Apply Three Filtering Methods: The feature selection begins by applying three filters independently, each of which evaluates the relevance of the features with respect to the target variable.

Step 2.1: Symmetrical Uncertainty (SU) Filter
Calculate Symmetrical Uncertainty: Compute the SU score for each feature by measuring the correlation between the feature and the target variable.
Rank Features: Rank the features based on their SU scores.

Select Top Features: Select the top k features with the highest SU scores. These are considered the most relevant to the target variable.

Step 2.2: Information Gain (IG) Filter
Calculate Information Gain: Compute the IG for each feature by quantifying the reduction in entropy when the feature is known.
Rank Features: Rank the features based on their IG scores, where higher IG values indicate greater importance.

Select Top Features: Select the top kkk features with the highest IG scores.

Step 2.3: Chi-SquareFilter
Perform Chi-Square Test: For each feature, apply the Chi-Square test to determine its level of independence from the target variable.

Rank Features: Rank the features based on their Chi-Square values. Higher values suggest a stronger dependency between the feature and the target.

Select Top Features: Select the top k features with the highest Chi-Square values.

Step 3: Combine Results of the Three Filters
Step 3.1: **Intersection of Feature Sets:** Combine the results of the three filters (SU, IG, and Chi-Square) by taking the intersection of the top-ranked features from each method.
This ensures that only the most relevant features, as agreed upon by all three methods, are retained for further analysis.
The combined feature set is smaller and more focused on high-impact predictors.

Step 4: Particle Swarm Optimization (PSO) for Feature Selection
Once the filtered feature set is determined, PSO is used to further optimize the selection of features.

Step 4.1: Initialize PSO Algorithm
Swarm Initialization: Initialize a population (swarm) of particles. Each particle represents a potential subset of the selected features from Step 3.

Position & Velocity: Initialize the position and velocity of each particle randomly. The position represents a feature subset.

Step 4.2: Define Fitness Function

Classification Performance: The fitness function evaluates how well the selected feature subset performs in classification. Use a machine learning model (e.g., Support Vector Machine, Decision Tree) to calculate performance metrics such as accuracy or F1-score.

Fitness Value: The fitness value for each particle is the classification accuracy (or any relevant performance metric) of the model using the particle's feature subset.

Step 4.3: Update Particle Velocity & Position

For each particle, update the velocity using the above mentionedformula.

Step 4.4: Evaluate Fitness of Each Particle

After updating the position, evaluate the fitness (classification performance) of the new feature subset for each particle.

Update Personal Best (p_i): If a particle's current fitness is better than its personal best, update its personal best.
Update Global Best (g): If any particle's current fitness is better than the global best, update the global best.

Step4.5: Iterate Until Convergence

Repeat the velocity and position updates, and re-evaluate the fitness of particles for a predefined number of iterations or until convergence (when no further improvement is observed in the global best).

Step 5: Select Optimal Feature Subset

Step 5.1: Identify the Best Feature Subset

Once PSO converges, the global best position (i.e., the best-performing feature subset) is selected as the optimal feature set.

Step 5.2: Train Final Model

Use the optimal feature subset to train the final classification model (e.g., on a Decision Tree, SVM, or any chosen model).

RESULT AND DISCUSSION

Performance Metrics

The performance of the proposed Feature Selection method is evaluated with their existing feature selection methods like Genetic Algorithm, Artificial Bee Colony (ABC) Optimization, Whale Optimization Algorithm (WOA), and Cultural Algorithm (CA) using classification techniques like Artificial Neural Network (ANN), Support Vector Machine (SVM) and Random Forest (RF). The dataset used in this research work is considered from the Kaggle Repository [19] Table 1 depicts the performance metrics used to evaluate the performance of the proposed and existing feature selection methods. Table 2 depicts the number of features obtained by the Proposed and existing feature selection methods. From the table 2, it is clear that the proposed TFIE-OFS method gives less number of features than the existing feature selection methods. Table 2 compares the number of features selected by the existing feature selection methods (Symmetrical Uncertainty (SU), Information Gain (IG), Chi-Square (CS)) and the proposed Thrice Filtered Information Energy Optimization-based Feature Selection (TFIE-OFS) method. The proposed TFIE-OFS method selects a more refined set of features, emphasizing its efficiency in identifying the most relevant attributes for classification. The SU method selects a total of 9 features, including key attributes like "chol," "cp," and "exang.". The IG method also selects 7 features, but with variations such as the absence of "age" and "trestbps.". The CS method selects 7 features but includes a different set, emphasizing "old peak" and "trestbps.". The proposed TFIE-OFS method selects 6 features, streamlining the selection to the most critical attributes such as "cp," "ca," and "thal."

Table 3 depicts the classification accuracy (in %) obtained by the Heart Disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets. Table 3 presents the classification accuracy (in %) of various classification techniques (ANN, RF, SVM) applied to the heart disease dataset after using different feature selection methods: Genetic Algorithm (GA), Artificial Bee Colony (ABC), Whale Optimization Algorithm (WOA), Crow Search Algorithm (CA), and the proposed TFIE-OFS method. The original dataset without feature selection provides the lowest classification accuracy across all classifiers: 48.32% (ANN), 43.97% (RF), and 42.86% (SVM), indicating the necessity of feature selection to improve performance. Among the existing feature selection methods, CA yields the highest accuracy: 72.59% (ANN), 71.67% (RF), and 69.78% (SVM).The proposed TFIE-OFS method achieves the highest classification accuracy across all classifiers, with 94.91% (ANN), 93.46% (RF), and 85.69% (SVM), showing a substantial improvement over the other methods.GA achieves respectable results but still lags behind TFIE-OFS, with accuracies of 70.84% (ANN), 69.34% (RF), and 67.43% (SVM).ABC and WOA produce similar but slightly lower accuracies, ranging between 55-59% for all classifiers.

Table 4 depicts the True Positive Rate (in %) obtained by the Heart Disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets. Table 4 presents the True Positive Rate (in %) of various classification techniques (ANN, RF, SVM) applied to the heart disease dataset after using different feature selection methods: Genetic Algorithm (GA), Artificial Bee Colony (ABC), Whale Optimization Algorithm (WOA), Crow Search Algorithm (CA), and the proposed TFIE-OFS method. The original dataset without feature selection results in the lowest True Positive Rate (TPR), with 52.76% (ANN), 51.26% (RF), and 46.35% (SVM), highlighting the need for feature selection to improve detection rates. Among the existing methods, CA achieves the highest TPR, with 83.19% (ANN), 82.3% (RF), and 69.24% (SVM).The proposed TFIE-OFS method significantly outperforms all other methods, achieving TPRs of 95.51% (ANN), 92.42% (RF), and 79.96% (SVM), indicating superior performance in correctly identifying positive cases.GA shows competitive results with TPRs of 74.45% (ANN), 73.05% (RF), and 71.16% (SVM), but still lags behind TFIE-OFS.ABC and WOA exhibit similar performance, with TPRs ranging from 59-63% across all classifiers, which is higher than the original dataset but lower than GA and CA.

Table 5 depicts the False Positive Rate (in %) obtained by the Heart Disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets. Table 5 presents the False Positive Rate (FPR in %) of various classification techniques (ANN, RF, SVM) applied to the heart disease dataset after using different feature selection methods: Genetic Algorithm (GA), Artificial Bee Colony (ABC), Whale Optimization Algorithm (WOA), Crow Search Algorithm (CA), and the proposed TFIE-OFS method. The original dataset shows the highest False Positive Rate across all classifiers, with 56.58% (ANN), 63.8% (RF), and 64.32% (SVM), highlighting its poor ability to minimize false positives without feature selection. Among the existing methods, CA achieves the lowest FPR, with 25.60% (ANN), 31.91% (RF), and 33.42% (SVM).The proposed TFIE-OFS method significantly outperforms all other methods, reducing the FPR to 5.72% (ANN), 5.36% (RF), and 13.36% (SVM), indicating a dramatic reduction in false positives and enhancing the accuracy of classification.GA performs moderately, reducing FPR to 32.87%(ANN), 35.31% (RF), and 36.22% (SVM), which is significantly better than the original dataset but higher than CA and TFIE-OFS.ABC and WOA have similar FPR values, ranging from 43-48%, which is better than the original dataset but worse than GA and CA.

Table 6 depicts the Precision (in %) obtained by the heart disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets. Table 6 presents the Precision (in %) obtained by various classification techniques (ANN, RF, SVM) applied to the heart disease dataset after using different feature selection methods: Genetic Algorithm (GA), Artificial Bee Colony (ABC), Whale Optimization Algorithm (WOA), Crow Search Algorithm (CA), and the proposed TFIE-OFS method. The original dataset yields the lowest precision across all classifiers: 51.60% (ANN), 47.71% (RF), and 46.53% (SVM), indicating high levels of false positives without feature selection. Among the existing feature selection methods, CA achieves the highest precision, with 80.17% (ANN), 73.20% (RF), and 72.83% (SVM).The proposed TFIE-OFS method dramatically improves precision across all classifiers, with the highest values: 95.53% (ANN), 94.32% (RF), and 82.65% (SVM), demonstrating its superior performance in accurately identifying positive instances.GA also delivers competitive precision results, with 73.52%

(ANN), **70.60%** (RF), and **69.81%** (SVM), though it still falls short of CA and TFIE-OFS.**ABC** and **WOA** perform similarly, with precision values in the **60-63%** range, which is an improvement over the original dataset but not as effective as GA, CA, or TFIE-OFS.

Table 7 depicts the Specificity (in %) obtained by the heart disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets. Table 7 shows the Specificity (in %) of various classification techniques (ANN, RF, SVM) applied to the heart disease dataset after different feature selection methods: Genetic Algorithm (GA), Artificial Bee Colony (ABC), Whale Optimization Algorithm (WOA), Crow Search Algorithm (CA), and the proposed TFIE-OFS method. The original dataset exhibits the lowest Specificity across all classifiers: 43.42% (ANN), 36.2% (RF), and 35.68% (SVM), indicating poor performance in identifying true negatives without feature selection. Among the existing methods, CA provides the highest Specificity, with 74.4% (ANN), 68.09% (RF), and 66.58% (SVM), reflecting a notable improvement. The proposed TFIE-OFS method achieves the highest Specificity across all classifiers: 94.28% (ANN), 92.64% (RF), and 86.64% (SVM), showcasing its superior ability to correctly identify negative cases.GA also performs well, yielding Specificity values of 67.13% (ANN), 64.69% (RF), and 63.78% (SVM), but remains less effective than CA and TFIE-OFS.ABC and WOA show moderate improvements, with Specificity values in the 51-56% range, indicating some effectiveness but falling behind GA, CA, and TFIE-OFS.

Table 8 depicts the Miss Rate (in %) obtained by the Heart Disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets. Table 8 presents the Miss Rate (in %) obtained by various classification techniques (ANN, RF, SVM) when applied to the heart disease dataset processed using different feature selection methods: Genetic Algorithm (GA), Artificial Bee Colony (ABC), Whale Optimization Algorithm (WOA), Crow Search Algorithm (CA), and the proposed TFIE-OFS method. The original dataset produces the highest Miss Rates: 47.24% (ANN), 48.74% (RF), and 53.65% (SVM), indicating a high rate of misclassification without feature selection.GA reduces the Miss Rate significantly to 25.55% (ANN), 26.95% (RF), and 28.84% (SVM), though it is still not as effective as CA or the proposed TFIE-OFS method.ABC and WOA have moderate Miss Rates, ranging from 36.66% to 40.87%, showing a better performance than the original dataset but are less effective than GA, CA, or TFIE-OFS.CA achieves a substantial reduction in Miss Rate, especially for ANN and RF classifiers, with values of 16.81% (ANN) and 17.7% (RF). However, for SVM, the Miss Rate is relatively higher at 30.76%.The proposed TFIE-OFS method achieves the lowest Miss Rates: 4.49% (ANN), 7.58% (RF), and 20.04% (SVM), demonstrating its superior ability to minimize classification errors across all classifiers.

Table 9 depicts the False Discovery Rate (in %) obtained by the heart disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets. Table 9 presents the False Discovery Rate (FDR in %) obtained by applying various classification techniques (ANN, RF, SVM) to the heart disease dataset processed using different feature selection methods: Genetic Algorithm (GA), Artificial Bee Colony (ABC), Whale Optimization Algorithm (WOA), Crow Search Algorithm (CA), and the proposed TFIE-OFS method. The original dataset exhibits high False Discovery Rates, with values of 48.4% (ANN), 52.29% (RF), and 53.47% (SVM), indicating a substantial number of false positives when no feature selection is applied. The GA method leads to a noticeable reduction in FDR to 26.48% (ANN), 29.4% (RF), and 30.19% (SVM), demonstrating some effectiveness in improving classification accuracy compared to the original dataset.ABC and WOA achieve moderate reductions in FDR, with values ranging from 36.24% to 43.99%, reflecting a better performance than the original dataset but still higher than GA and CA. The CA method significantly lowers the FDR, achieving 19.83% (ANN), 26.8% (RF), and 27.17% (SVM), indicating its strong potential in minimizing false discoveries. The proposed TFIE-OFS method shows remarkable results, yielding the lowest False Discovery Rates of 4.47% (ANN), 5.68% (RF), and 17.35% (SVM), showcasing its effectiveness in correctly identifying true positives and significantly reducing false positives.

CONCLUSION

The Thrice Filtered Information Energy Optimization Based Feature Selection (TFIE-OFS) method represents a significant advancement in feature selection techniques for classification and prediction tasks, particularly in high-dimensional datasets such as those used in heart disease diagnosis. By integrating three robust filtering methods—Symmetrical Uncertainty, Information Gain, and Chi-Square Analysis—TFIE-OFS effectively narrows down the feature space to include only the most relevant predictors. This multi-filtering approach ensures a comprehensive assessment of feature importance, leading to a more refined selection process. Furthermore, the incorporation of Particle Swarm Optimization (PSO) enhances the feature selection process by optimizing the subset of features based on their contribution to model performance. The results from the various evaluations of the Thrice Filtered Information Energy Optimization-based Feature Selection (TFIE-OFS) method indicate its significant efficacy in improving the performance of heart disease classification tasks. The proposed TFIE-OFS method consistently outperformed existing feature selection techniques across multiple metrics, demonstrating its ability to enhance classification accuracy, reduce false discovery rates, minimize miss rates, and maintain high specificity.

REFERENCES

1. Katarya, Rahul, and Sunit Kumar Meena. "Machine learning techniques for heart disease prediction: a comparative study and analysis." *Health and Technology* 11.1 (2021): 87-97.
2. Shanti, M. A., and K. Saravanan. "Knowledge data map—A framework for the field of data mining and knowledge discovery." *International Journal of Computer Engineering & Technology* 8.5 (2017): 67-77.
3. Shanti, M. A., and K. Saravanan. "An Effect of Data Mining Techniques in Public Healthcare-A Case Study." *International Journal of Civil Engineering and Technology* 9.9 (2018): 115-122.
4. Shanti, M. A. "A Study to Analyse the Quality of Work Life with Special Reference to Private Sector Bank Employees in Kumbakonam Town of Thanjavur District." *Our Heritage*, vol. 68, 2020.
5. Shanti, M. A. "Earthquake Prediction Using SVM Based Time Predictable Technique." *International Journal of Computer Sciences and Engineering (IJCSE)*, vol. 7, no. 4, 2019, pp. 2347–2693.
6. Kasthuri, S., and A. Nisha Jebaseeli. "Review analysis of Twitter sentimental data." *Bioscience Biotechnology Research Communications (BBRC),(UGC CARE Journal-Web of Science), Special Issue* 13.6 (2020): 209-214.
7. Kasthuri, S., and A. Nisha Jebaseeli. "Social network analysis in data processing." *Adalya Journal,(UGC CARE-B Journal–Web of Science), Impact Factor* 5 (2020): 260-263.
8. Ambika, G. "Advanced Human Activity Recognition: Leveraging Adaptive Neural Networks and Diverse Machine Learning Algorithms on IoT Data." *Fuzzy Systems and Soft Computing*, vol. 19, no. 02(V), 2024, pp. 11–17.
9. Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems* 37.5 (2019): 6743-6760.
10. Durairaj, M., and T. S. Poornappriya. "Why feature selection in data mining is prominent? A survey." *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications: AISGSC 2019*. Springer International Publishing, 2020.
11. Ambika, G. "Processing Over Encrypted Query Data in Internet of Things (IoTs): CryptDBs, MONOMI and SDB." *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 6, Issue-8, 2018, p. 8.
12. Dheepak, T. "Enhancing the Cloud Security with ECC based Key Generation Technique." *Annals of the Romanian Society for Cell Biology* 25.2 (2021): 3874-3891.
13. Malathi, T., and T. Dheepak. "Enhanced Regression Method for Weather Forecasting." *The Scientific Temper*, vol. 15, special issue, 16 Oct. 2024, pp. 146–149. *The Scientific Temper*.
14. Suresh, T., T. Dheepak, and R. Kayalvizhi. "Optimizing QOS in Mobile Ad Hoc Networks Through Advanced Routing Protocols Under Wormhole Attack Scenarios." *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 11, 2023, pp. 584–594.

15. Suresh, T., T. Dheepak, and K. Saraswathi. "Enhancing Sentiment Analysis for Autistic Children: A Hybrid Approach Using SBERT and Ensemble Learning." *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 9, 2023, pp. 4649–4656.
16. Dheepak, T. "Optimizing Routing Protocols in Mobile Adhoc Networks Using Firefly Optimization Algorithm." *Webology*, vol. 18, no. 5, 2021.
17. Dheepak, T. "An Enhanced Access Control Mechanism for Mobile Cloud Computing." *Design Engineering* (2021): 10805-10814.
18. Dheepak, T. "Trust Based Cluster Selection for Intrusion Detection in Mobile Ad Hoc Networks." *Technology*, vol. 11, no. 10, 2020, pp. 421–430.
19. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Table 1: Performance Metrics used in this research work

Metrics	Equation
Accuracy	TP+TN/TP+FN+TN+FP
True Positive Rate (TPR) (Sensitivity or Recall)	TP/TP+FN
False Positive Rate (FPR)	FP/FP+TN
Precision	TP/TP+FP
True Negative Rate (Specificity)	1- False Positive Rate (FPR)
Miss Rate	1-True Positive Rate (TPR)
False Discovery Rate	1- Precision

Table 2: Number of Features obtained by the Existing and Proposed Feature Selection Methods

Feature Index	Number of Features selected by existing feature selection methods and proposed TFIE-OFS method			
	SU	IG	CS	TFIE-OFS
1	chol	cp	age	cp
2	cp	age	old peak	ca
3	exang	thal	trestbps	thal
4	ca	ca	cp	slope
5	slope	trestbps	ca	exang
6	old peak	exang	thal	chol
7	sex	slope	exang	
8	age		slope	
9	trestbps			

Table 3: Classification Accuracy (in %)obtained by the heart disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets

Feature Selection Methods	Classification Accuracy (in %) by Classification Techniques		
	ANN	RF	SVM
Original dataset	48.32	43.97	42.86
GA	70.84	69.34	67.43
ABC	59.73	58.43	56.32
WOA	58.64	57.34	55.43
CA	72.59	71.67	69.78
Proposed TFIE- OFS method	94.91	93.46	85.69

Table 4: True Positive Rate (in %)obtained by the heart disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets

Feature Selection Methods	True Positive Rate (in %) by Classification Techniques		
	ANN	RF	SVM
Original dataset	52.76	51.26	46.35
GA	74.45	73.05	71.16
ABC	63.34	62.16	60.24
WOA	62.25	61.27	59.13
CA	83.19	82.3	69.24
Proposed TFIE- OFS method	95.51	92.42	79.96

Table 5: False Positive Rate (in %)obtained by the heart disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets

Feature Selection Methods	False Positive Rate (in %) by Classification Techniques		
	ANN	RF	SVM
Original dataset	56.58	63.8	64.32
GA	32.87	35.31	36.22
ABC	43.78	44.42	47.35
WOA	44.69	45.53	48.43
CA	25.60	31.91	33.42
Proposed TFIE- OFS method	5.72	5.36	13.36

Table 6: Precision (in %)obtained by the heart disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets

Feature Selection Methods	Precision(in %) by Classification Techniques		
	ANN	RF	SVM
Original dataset	51.60	47.71	46.53
GA	73.52	70.60	69.81
ABC	62.80	61.57	56.01
WOA	63.76	60.86	57.54
CA	80.17	73.20	72.83
Proposed TFIE- OFS method	95.53	94.32	82.65

Table 7: Specificity(in %)obtained by the heart disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets

Feature Selection Methods	Specificity (in %) by Classification Techniques		
	ANN	RF	SVM
Original dataset	43.42	36.2	35.68
GA	67.13	64.69	63.78
ABC	56.22	55.58	52.65
WOA	55.31	54.47	51.57
CA	74.4	68.09	66.58
Proposed TFIE- OFS method	94.28	92.64	86.64

Table 8: Miss Rate (in %)obtained by the Heart Disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets

Feature Selection Methods	Miss Rate (in %) by Classification Techniques		
	ANN	RF	SVM
Original dataset	47.24	48.74	53.65
GA	25.55	26.95	28.84
ABC	36.66	37.84	39.76
WOA	37.75	38.73	40.87
CA	16.81	17.7	30.76
Proposed TFIE- OFS method	4.49	7.58	20.04

Table 9: False Discovery Rate (in %)obtained by the heart disease dataset using original dataset, GA, ABC, WOA, CA and proposed TFIE-OFS methods processed datasets

Feature Selection Methods	False Discovery Rate (in %) by Classification Techniques		
	ANN	RF	SVM
Original dataset	48.4	52.29	53.47
GA	26.48	29.4	30.19
ABC	37.2	38.43	43.99
WOA	36.24	39.14	42.46
CA	19.83	26.8	27.17
Proposed TFIE- OFS method	4.47	5.68	17.35