Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

Hybrid Feature Selection with Metaheuristics for Improving the Accuracy of Diabetes Disease Prediction

Ida Maratul Khamidah[1,*], Suci Ramadhani[1], Aulia Khoirunnita[2]

[1]Program Studi Teknologi Rekayasa Perangkat Lunak, Politeknik Pertanian Negeri Samarinda, Samarinda, Indonesia
[2]Fakultas Teknik, Program Studi Informatika, Universitas Mulawarman, Samarinda, Indonesia
Email: [1,*]idakhamidah@politanisamarinda.ac.id, [2] suci.ramadhani.usu@gmail.com, aulia.khoirunnita@unmul.ac.id
Email Penulis Korespondensi: idakhamidah@politanisamarinda.ac.id
Submitted: **16/03/2026**; Accepted: **31/03/2026**; Published**: 31/03/2026**

Abstract–Early diagnosis of diabetes mellitus is crucial to prevent severe complications and reduce long-term healthcare costs, making accurate and efficient predictive models an important research focus in medical data analytics. However, one of the main challenges in diabetes prediction lies in the presence of irrelevant and redundant features within medical datasets, which can degrade classification accuracy, increase computational complexity, and reduce model generalizability. To address this issue, this study proposes a Hybrid Feature Selection (HFS) approach that integrates filter-based methods and meta-heuristic optimization to identify an optimal subset of features for diabetes prediction. In the proposed framework, statistical filter techniques combining Chi-square and Mutual Information are first employed to rank and reduce feature dimensionality by selecting the most relevant attributes. Subsequently, a Genetic Algorithm (GA) is applied to further optimize the feature subset by maximizing classification accuracy while minimizing the number of selected features. The effectiveness of the proposed HFS approach is evaluated using the Pima Indian Diabetes Dataset, consisting of 768 instances and 8 clinical features, and tested across multiple machine learning classifiers, including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost. Experimental results demonstrate that the proposed HFS significantly improves predictive performance compared to baseline models without feature selection. Specifically, the Random Forest classifier achieved the highest accuracy of 79.22%, compared to 74.03% in the baseline model, representing an improvement of approximately 5.2%. Additionally, notable improvements were observed in F1-score and AUC, with AUC increasing from 0.8336 to 0.8403. Beyond accuracy gains, the proposed method reduced feature dimensionality from 8 to 5 features, resulting in lower computational cost and faster model training time. These findings indicate that the hybrid integration of filter-based selection and meta-heuristic optimization provides a robust and efficient solution for feature selection in medical prediction tasks. Overall, the proposed HFS framework offers a promising approach for developing accurate, efficient, and reliable decision-support systems for early diabetes diagnosis.

Keywords: Diabetes Prediction; Feature Selection; Meta-Heuristic Optimization; Machine Learning; Hybrid Methods

1. INTRODUCTION

Diabetes mellitus is a major global health issue with a prevalence rate that continues to rise annually. As reported by the WHO, over 537 million adults worldwide were living with diabetes in 2021, a figure projected to increase to 643 million by 2030 [1]. This condition imposes a significant economic and social burden on global healthcare systems, accompanied by an increase in complications such as stroke, kidney failure, and peripheral neuropathy [2]. The high prevalence of diabetes also underscores the need for novel approaches in prevention and early detection to mitigate the risk of severe complications [3]. Consequently, research in predictive analytics for diabetes has become a strategic area within global digital health [4].

Predictive analytics based on machine learning (ML) play a crucial role in the early diagnosis of diabetes by leveraging clinical and biochemical data to detect patterns that are imperceptible to humans [5]. Predictive models enable faster medical interventions, personalized treatment, and a reduction in healthcare costs [6]. With the growth of big data in healthcare, ML algorithms such as SVM, Random Forest, and XGBoost have become increasingly efficient for detecting diabetes risk with high accuracy [7]. However, model performance remains influenced by the quality and relevance of the features used [8].

One of the main challenges in developing predictive models for diseases is the curse of dimensionality, where an increasing number of features can actually degrade model performance due to heightened redundancy and noise [9]. This leads to more complex models that are difficult to generalize to new data [10]. Therefore, feature selection becomes a crucial step in identifying the most informative subset of features, which can enhance accuracy and reduce computation time [11].

Conventional approaches to feature selection fall into three main categories: filter, wrapper, and embedded methods [12]. Filter methods are fast and simple but often overlook interactions between features, resulting in suboptimal subsets [10]. Wrapper methods yield better results but are computationally expensive due to repeated model evaluations [8]. Meanwhile, embedded methods integrate feature selection during the model training process but are often dependent on specific algorithms, making them less flexible [5].

Previous research indicates that the use of meta-heuristics, such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Whale Optimization Algorithm (WOA), can improve the accuracy of diabetes prediction models [13]. Hybrid feature selection approaches that combine filter methods and meta-heuristics have been shown to produce more stable and efficient feature subsets compared to single methods [14]. However, computational complexity and result stability remain primary challenges [4].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

Based on the findings of prior studies, it can be concluded that although meta-heuristic-based feature selection methods like GA, PSO, ACO, and WOA have successfully enhanced the accuracy of diabetes prediction models, most of these approaches still employ a single-stage feature selection strategy [10]. To date, there is limited research that systematically combines filter methods for initial feature screening with meta-heuristic algorithms as a subsequent optimization stage in the medical domain [12]. This two-stage combination is important because filter approaches can rapidly reduce the initial dimensionality, while meta-heuristics can globally explore for the best feature subset [8]. However, in the context of predicting chronic diseases such as diabetes, the implementation of hybrid feature selection that integrates both strategies remains scarce, thus leaving a significant research gap [3]. This gap provides a foundation for exploring new approaches that are more efficient, stable, and accurate for the early diagnosis of diabetes using multidimensional data [11].

Addressing this gap, this study proposes the development of a two-stage Hybrid Feature Selection (HFS) model that combines a Chi-square-based filter method with Particle Swarm Optimization (PSO), a meta-heuristic optimization technique, to obtain the most relevant feature subset for predicting diabetes [1]. This model will be tested using the Pima Indian Diabetes dataset and compared against baseline models such as SVM, Random Forest, KNN, and XGBoost [7]. The main contribution of this research is to offer a feature selection approach that not only improves accuracy but also reduces computational complexity and enhances the stability of classification results [15]. This approach is expected to serve as an effective alternative solution for the future development of AI-based early diabetes detection systems [4].

2. RESEARCH METHODOLOGY

This research methodology systematically describes the stages and approaches employed in developing a diabetes prediction model based on Hybrid Feature Selection (HFS) with meta-heuristic algorithms. The proposed methodology encompasses the processes of medical data collection and preprocessing, the implementation of a two-stage feature selection that combines statistical filter methods and meta-heuristic optimization, as well as the development and evaluation of machine learning classification models. Each stage is designed to enhance data quality, reduce feature dimensionality complexity, and obtain an optimal feature subset capable of improving prediction accuracy and stability. Performance evaluation is conducted comprehensively using various classification metrics and cross-validation techniques, and comparisons are made with baseline models without feature selection to assess the effectiveness of the proposed approach. The research stages employed are based on Figure 1 below.

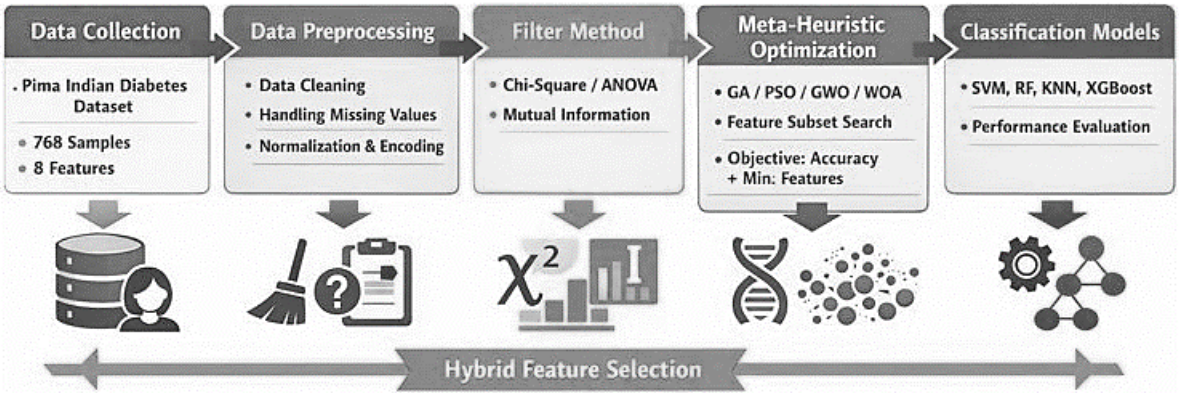


Figure 1. Stages of the Research

2.1 Data Collection

The first stage of this research involves the collection of medical data, which will serve as the foundation for developing a diabetes prediction model. The dataset utilized is a medical dataset comprising diagnostic measurements used to predict the likelihood of diabetes based on various health factors. This dataset consists of 768 records of female patients, with each record represented by 8 health attributes and one target variable (Outcome). The Outcome variable is binary, where a value of 1 indicates that the patient has been diagnosed with diabetes, and a value of 0 indicates that the patient has not been diagnosed with diabetes. This dataset is widely employed as a benchmark dataset in machine learning research within the healthcare domain, particularly for diabetes classification tasks.

Table 1. Dataset Specifications

Attribute	Description	Data Type	General Range of Values
Pregnancies	Number of times pregnant	Numeric (Integer)	0 – 17

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

Attribute	Description	Data Type	General Range of Values
Glucose	Plasma glucose concentration 2 hours after an oral glucose tolerance test	Numeric (Integer)	0 – 199
BloodPressure	Diastolic blood pressure (mm Hg)	Numeric (Integer)	0 – 122
SkinThickness	Triceps skinfold thickness (mm)	Numeric (Integer)	0 – 99
Insulin	2-hour serum insulin (μU/ml)	Numeric (Integer)	0 – 846
BMI	Body mass index (kg/m²)	Numeric (Float)	0.0 – 67.1
DiabetesPedigreeFunction	Diabetes pedigree function, indicating family history of diabetes	Numeric (Float)	0.078 – 2.42
Age	Age of the patient (years)	Numeric (Integer)	21 – 81
Outcome	Diabetes status (0 = Non-diabetic, 1 = Diabetic)	Categorical (Binary)	{0, 1}

The dataset in Table 1 was utilized to develop and evaluate the classification model for predicting diabetes onset, conduct exploratory analysis to identify patterns and correlations among health features, and test the effectiveness of the hybrid meta-heuristic-based feature selection method in improving the accuracy and efficiency of the prediction model. This dataset was adapted from data collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and has been widely employed in academic research related to medical decision support systems and machine learning-based diagnostics.

2.2 Data Preprocessing

The preprocessing stage aims to enhance data quality prior to feature selection and classification [16]. This process begins with data cleaning, which involves identifying illogical values, such as zero values in medical features (e.g., Glucose, BMI, and Insulin) that are clinically impossible [17]. These values are subsequently treated as missing values [18]. Next, missing value handling is performed, typically using imputation techniques such as class-conditional median imputation to preserve the original data distribution [17]. ollowing this, data normalization is applied to standardize the scale of numerical features and prevent dominance by particular features. Min-Max Scaling is frequently employed due to its ability to improve the performance of distance-based models such as KNN [19]. As all features are numerical, categorical encoding is unnecessary [20]. This preprocessing phase is crucial for reducing noise and enhancing model stability [21][22].

2.3 Proposed Hybrid Feature Selection

The third stage constitutes the initial phase of the Hybrid Feature Selection process. Figure 2 illustrates the Hybrid Feature Selection (HFS) Algorithm, which is detailed based on Figure 1, encompassing the Filter method and Meta-Heuristic Optimization stages.

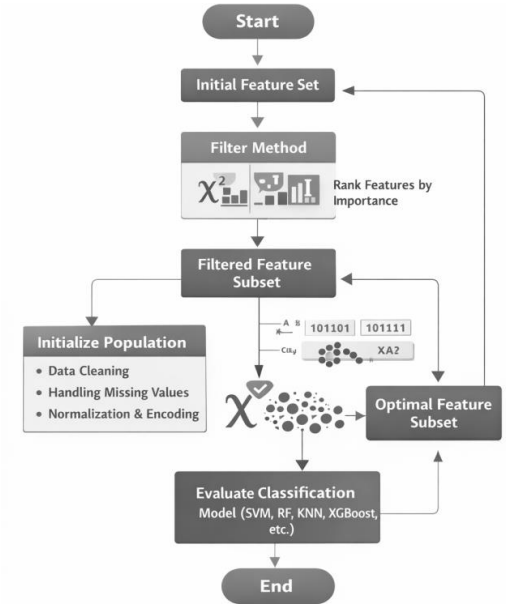


Figure 2. Hybrid Feature Selection (HFS) Algorithm

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

As illustrated in Figure 2, the first stage is the Filter method. In this stage, each feature is evaluated independently against the target variable using statistical techniques. The methods employed include Chi-square, ANOVA F-test, and Mutual Information. Chi-square measures the relationship between a feature and the target class [23], ANOVA tests the difference in means across classes[17], while Mutual Information is capable of capturing non-linear dependencies[22]. The outcome of this stage is a ranking of features based on their relevance level, after which a number of the best features are selected to proceed to the next stage. The Filter method was chosen due to its low computational complexity and its ability to rapidly reduce data dimensionality[23], thereby narrowing the search space for the subsequent meta-heuristic optimization stage.

The second stage constitutes the core of the Hybrid Feature Selection method, namely the optimization of feature combinations using meta-heuristic algorithms. In this stage, the feature subset resulting from the filter method is utilized as the initial population or search space. The meta-heuristic algorithms employed include the Genetic Algorithm (GA), where each chromosome is represented as a binary vector, with 1 indicating a selected feature and 0 indicating a feature not selected. The Fitness Function, defined in Equation 1 as the objective function, combines two primary goals: maximizing classification accuracy and minimizing the number of features. This encapsulates the essence of Hybrid Feature Selection: achieving high accuracy with low complexity. The GA evolution process begins with Selection, choosing the best individuals based on fitness probability, followed by Crossover, which involves the exchange of genes between solutions, and Mutation, which explores new solutions. In addition to GA, other meta-heuristic algorithms utilized are Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), and Whale Optimization Algorithm (WOA) [24].

$$F = \alpha * (1 - Accuracy) + \beta * \frac{|S|}{|T|} \quad (1)$$

where |S| is number of selected features, |T| total features, and α, β are weighting factors.

2.4 Classification Models

The next stage is the classification process using the optimal feature subset generated by the Hybrid Feature Selection method. Several machine learning algorithms are employed to evaluate the performance of the method, namely Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), XGBoost, and Logistic Regression. The use of various models aims to assess the consistency and generalizability of the feature selection method across different algorithms.

Support Vector Machine (SVM) is effective in handling high-dimensional data[17], K-Nearest Neighbor (KNN) is simple and effective on standardized datasets[19], Logistic Regression is used as an interpretable baseline [20], while Random Forest and XGBoost represent high-performance ensemble models. Random Forest (RF) excels in handling non-linearity and noise[24]. XGBoost is recognized for its high performance on medical tabular data[17]. All models are trained and tested using the same feature subset to ensure a fair and objective comparison of the results.

2.5 Performance Evaluation

The final stage is the evaluation of the diabetes prediction model's performance. The evaluation was conducted using several standard metrics: Accuracy, Precision, Recall, F1-score[25][26][27], and the Area Under the Curve (AUC-ROC)[28][29] as presented in Figure 3. The use of multiple metrics is crucial in the medical domain to ensure that the model is not only accurate but also capable of correctly detecting diabetic patients.

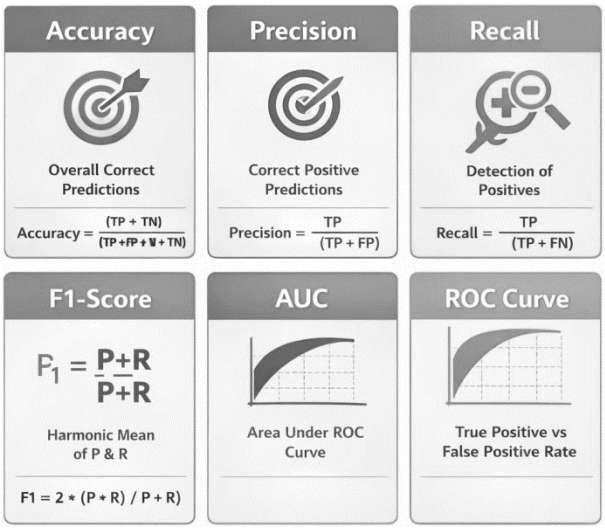


Figure 3. Model Performance Evaluation

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

Based on Figure 3, to enhance the reliability of the results, a 10-fold cross-validation technique was employed[30][31], In this method, the dataset is partitioned into ten subsets, and the training and testing processes are conducted iteratively. The evaluation results were utilized to compare the model's performance before and after the implementation of Hybrid Feature Selection, as well as to assess the effectiveness of the proposed approach.

3. RESULTS AND DISCUSSION

This section presents a comprehensive analysis of the performance of the diabetes disease prediction model developed using a metaheuristic-based Hybrid Feature Selection approach. The experimental results are systematically presented through the evaluation of various classification models, both with and without feature selection, to assess the impact of implementing the proposed method. The test results are analyzed using evaluation metrics relevant to the medical domain, such as accuracy, precision, recall, F1-score, and AUC, thereby providing a thorough overview of the model's capability to accurately detect diabetic patients. The subsequent discussion focuses on comparing model performance, interpreting the effect of feature reduction on improving accuracy and computational efficiency, and analyzing the stability of the classification results. Furthermore, the findings of this study are linked to the results of previous research to identify the advantages, limitations, and practical implications of the Hybrid Feature Selection approach in developing machine learning-based medical decision support systems.

A crucial initial stage in preprocessing is identifying the presence of missing values (a) in each feature. In the diabetes dataset used, several medical attributes are conceptually impossible to have a value of zero (e.g., Glucose, Insulin, BMI); therefore, zero values are considered as representing missing data.

	0
Pregnancies	0
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0
Outcome	0

(a)

	0
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

(b)

Figure 4. Number of Missing Values: (a) before imputation and (b) after imputation

Based on Figure 4 (a), it can be observed that the features Insulin (374 missing values) and SkinThickness (227 missing values) have the highest number of missing values, followed by BloodPressure (35), BMI (11), and Glucose (5). Meanwhile, the features Pregnancies, DiabetesPedigreeFunction, Age, and Outcome contain no missing values. This condition indicates the need for careful handling of missing values, as their proportion is quite significant and could affect model stability if ignored.

To address the issue of missing values identified in the previous stage, an imputation process was performed using the median value for each feature. The median method was chosen because it is more robust to outliers and suitable for medical data, which often has a non-normal distribution. Figure 4 (b) shows that all features now have zero missing values, indicating that the imputation process was successfully completed. With the missing values resolved, the dataset has become more consistent and is ready for use in the normalization, feature selection, and modeling stages without the risk of bias due to incomplete data.

The next stage is to analyze the value range of each feature before normalization. This analysis is important for understanding the differences in scale among features, which can have a significant impact on distance-based machine learning algorithms and optimization.

	min	max
Pregnancies	0.000	17.00
Glucose	44.000	199.00
BloodPressure	24.000	122.00
SkinThickness	7.000	99.00
Insulin	14.000	846.00
BMI	18.200	67.10
DiabetesPedigreeFunction	0.078	2.42
Age	21.000	81.00

(a)

	min	max
Pregnancies	0.0	1.0
Glucose	0.0	1.0
BloodPressure	0.0	1.0
SkinThickness	0.0	1.0
Insulin	0.0	1.0
BMI	0.0	1.0
DiabetesPedigreeFunction	0.0	1.0
Age	0.0	1.0

(b)

Figure 5. Feature Value Ranges: (a) before normalization and (b) after normalization

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

Based on Figure 5 (a), a considerably extreme difference in scale among features is observable; for instance, Glucose (44–199), Insulin (14–846), and BMI (18.2–67.1). This disparity in scale has the potential to cause features with larger values to dominate the model's learning process, thereby rendering normalization an essential step. To address this imbalance in scale among features, the Min-Max Normalization technique was applied, mapping all feature values into the range [0, 1]. This technique is commonly employed in machine learning as it preserves the proportions of the original data. In Figure 5 (b), it is evident that all features have been successfully normalized, exhibiting a minimum value of 0 and a maximum of 1. This outcome ensures that each feature contributes equally to the feature selection and classification processes, thereby enhancing the stability and convergence of the employed machine learning and meta-heuristic algorithms.

Upon the completion of preprocessing, a correlation analysis among features was conducted using a heatmap to understand the linear relationships between variables as well as their relationships with the target variable (Outcome). This analysis aids in identifying potential feature redundancy.

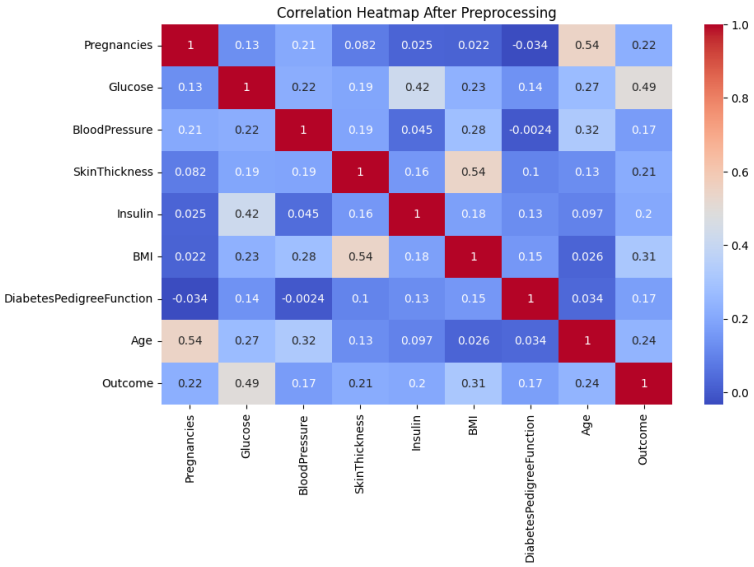


Figure 6. Correlation Heatmap After Preprocessing

Based on Figure 5, the Glucose feature exhibits the highest correlation with the Outcome (≈0.49), followed by BMI (≈0.31) and Age (≈0.24), indicating the significant role of these features in diabetes prediction. Furthermore, high correlations are observed between SkinThickness and BMI (≈0.54) as well as between Pregnancies and Age (≈0.54), suggesting information redundancy. These findings underscore the urgency of implementing Hybrid Feature Selection to eliminate redundant features while retaining the most informative ones.

Upon the completion of all preprocessing stages, the subsequent step is to analyze the class distribution of the target variable, Outcome (Figure 7). This analysis is crucial for determining whether the dataset is balanced or imbalanced, as class imbalance can adversely affect the performance of classification models, particularly on evaluation metrics such as accuracy, recall, and AUC.

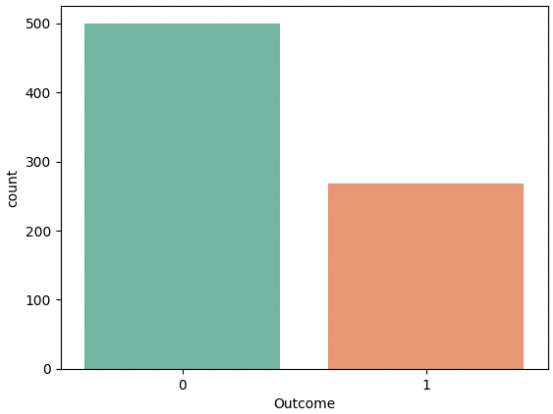


Figure 7. Distribution of Outcome Variable Classes (Diabetes vs. Non-Diabetes)

Based on Figure 6, the class distribution shows that there are 500 instances (≈65%) with label 0 (non-diabetes) and 268 instances (≈35%) with label 1 (diabetes). These results indicate a moderate class imbalance, where the non-diabetes class dominates the dataset. This condition has the potential to cause the model to be biased

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

towards predicting the majority class if not addressed properly. Therefore, the use of comprehensive evaluation metrics such as Precision, Recall, F1-score, and AUC, along with the implementation of cross-validation, becomes crucial to ensure that the proposed Hybrid Feature Selection model can recognize both classes fairly and not only optimize predictions for the majority class.

3.1 Results

After the data underwent preprocessing and normalization stages, a filter-based feature selection process was conducted to measure the relevance level of each feature to the target variable, Outcome (Table 2). In this study, a Hybrid Filter approach was employed, combining Chi-Square and Mutual Information. This enables the capture of both statistical relationships and non-linear information dependencies between features and the class. The results from both methods were then averaged in the form of a ranking to obtain a more stable feature evaluation that does not depend on a single criterion.

Table 2. Feature Rankings Based on Hybrid Filter

Feature	Score / Rank
Glucose	1.0
Age	2.5
BMI	3.0
Pregnancies	3.5
Insulin	5.5
DiabetesPedigreeFunction	6.5
SkinThickness	6.5
BloodPressure	7.5

Based on Table 2, the Glucose feature occupies the highest rank (score 1.0), indicating that blood glucose level is the most dominant indicator in distinguishing between diabetic and non-diabetic patients. This finding is consistent with clinical knowledge and the results of the previous correlation analysis. The Age, BMI, and Pregnancies features are also ranked highly, signifying the significant contribution of age, obesity, and pregnancy history to the risk of diabetes. Conversely, the BloodPressure, SkinThickness, and DiabetesPedigreeFunction features have lower rankings, suggesting their relatively smaller individual relevance or potential information redundancy with other features. The results of this ranking were subsequently used as the initial input for the meta-heuristic optimization stage, where algorithms such as the Genetic Algorithm (GA) are tasked with seeking the optimal feature subset combination from these selected features, considering the balance between classification accuracy and the number of features, as can be seen in the following Table 3.

Table 3. Log of Best Fitness per Generation in the Genetic Algorithm

Generation	Best Fitness
Generation 1/20	0.7109
Generation 2/20	0.7154
Generation 3/20	0.7109
Generation 4/20	0.7245
Generation 5/20	0.7264
Generation 6/20	0.7305
Generation 7/20	0.7305
Generation 8/20	0.7305
Generation 9/20	0.7305
Generation 10/20	0.7264
Generation 11/20	0.7285
Generation 12/20	0.7285
Generation 13/20	0.7285
Generation 14/20	0.7136
Generation 15/20	0.7136
Generation 16/20	0.7089
Generation 17/20	0.7089
Generation 18/20	0.7136
Generation 19/20	0.7136
Generation 20/20	0.7188

Based on Table 3, it can be observed that the Best Fitness value experienced a significant increase in the early generations, reaching a maximum value of 0.7305 in the 6th generation. This indicates that the GA successfully found a solution with better classification performance compared to previous generations. After reaching this point, the fitness value tended to stabilize until the 9th generation, suggesting the occurrence of

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

solution convergence. Fluctuations in the fitness value in subsequent generations reflect the exploration mechanism generated by the mutation operation, which aims to avoid local optima traps. Although a slight decrease occurred in the final generation, the fitness value remained within a competitive range, demonstrating the stability of the optimization process.

Based on this optimization process, the GA produced the best feature subset consisting of Glucose, BMI, Pregnancies, Insulin, and DiabetesPedigreeFunction. This subset reflects a combination of features that not only possess high individual relevance but also provide optimal collective contribution to the classification model's performance. Following the acquisition of the optimal feature subset through the Hybrid Feature Selection method based on Filter and Genetic Algorithm, the subsequent stage involved evaluating the performance of several machine learning algorithms commonly used in diabetes disease prediction. This evaluation aimed to measure the extent to which the selected feature subset could enhance the classification model's capability in distinguishing between diabetic and non-diabetic patients. Performance measurement was conducted using the metrics (Accuracy, Precision, Recall, and F1-Score) presented in Table 4, and the AUC presented in Figure 8, thereby providing a comprehensive overview of the model's quality, particularly on datasets with imbalanced class distribution.

Table 4. Performance of Machine Learning Models with Hybrid Feature Selection

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	0.7922	0.7018	0.7273	0.7143	0.8403
SVM	0.7727	0.7174	0.6000	0.6535	0.8522
XGBoost	0.7338	0.6250	0.6364	0.6306	0.7892
KNN	0.7208	0.6111	0.6000	0.6055	0.7730

Based on Table 4, the Random Forest model demonstrates the best overall performance, achieving the highest accuracy (≈79.22%) along with a good balance between precision, recall, and F1-score. This indicates its capability in handling complex and non-linear medical data.

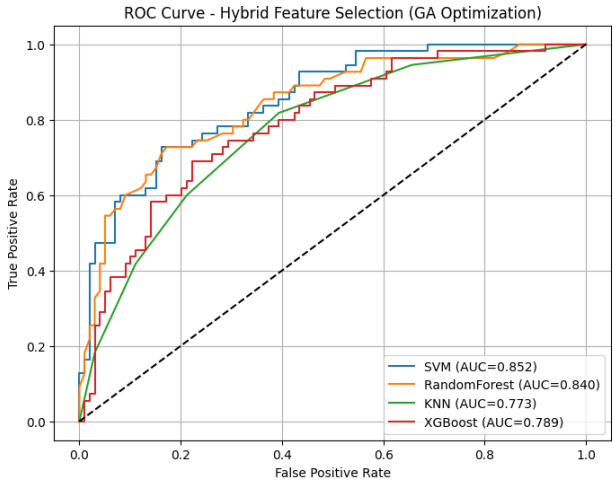


Figure 8. ROC and AUC

The SVM model achieved the highest AUC value (≈0.85), as illustrated in Figure 8 above, indicating its excellent discriminative ability in separating diabetic and non-diabetic classes, despite its relatively lower recall value (Table 4). Meanwhile, XGBoost and KNN demonstrated lower performance compared to Random Forest and SVM, suggesting that these two models were less than optimal in utilizing the feature subset resulting from GA optimization on this dataset. Overall, these results indicate that Hybrid Feature Selection is capable of enhancing the effectiveness of classification models, particularly in ensemble and margin-based learning algorithms, while maintaining a balance between predictive performance and model complexity.

The subsequent stage comprehensively evaluates the impact of applying a Filter-based and Genetic Algorithm Hybrid Feature Selection (Hybrid GA) on the performance of machine learning models compared to baseline models that utilize all features without selection. This comparison was conducted to ensure that the observed performance improvement is not merely a result of model complexity, but rather a consequence of selecting a more relevant and informative feature subset, as presented in Table 5. The evaluation was performed using the metrics of Accuracy, Precision, Recall, F1-Score, and AUC, thereby enabling a fair and comprehensive analysis across various aspects of classification performance.

Table 5. Comparison of Baseline vs Hybrid Feature Selection Performance

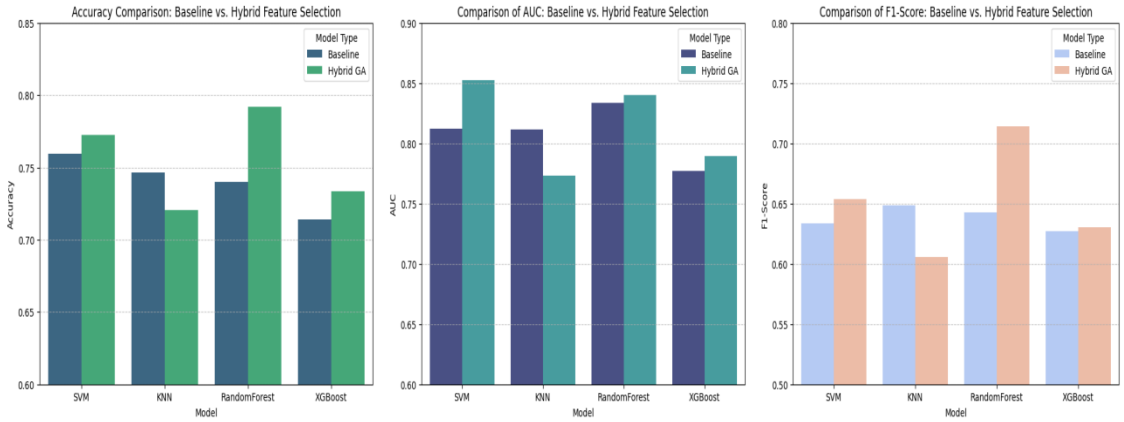
Model	Type	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	Hybrid GA	0.7922	0.7018	0.7273	0.7143	0.8403

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

Model	Type	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	Baseline	0.7403	0.6316	0.6545	0.6429	0.8336
SVM	Hybrid GA	0.7727	0.7174	0.6000	0.6535	0.8522
SVM	Baseline	0.7597	0.6957	0.5818	0.6337	0.8119
XGBoost	Hybrid GA	0.7338	0.6250	0.6364	0.6306	0.7892
XGBoost	Baseline	0.7143	0.5873	0.6727	0.6271	0.7774
KNN	Hybrid GA	0.7208	0.6111	0.6000	0.6055	0.7730
KNN	Baseline	0.7468	0.6429	0.6545	0.6486	0.8118

Based on Table 5, the application of Hybrid Feature Selection consistently improves the performance of the Random Forest, SVM, and XGBoost models, as evidenced by the increase in accuracy, precision, F1-score, and AUC compared to the baseline models. The most significant improvement is observed in the Random Forest model, where accuracy increases by approximately 5.2%, accompanied by clear enhancements in recall and F1-score. This indicates that optimal feature selection assists the model in recognizing complex patterns in medical data more effectively.

Conversely, in the KNN model, the baseline performance is slightly superior compared to the Hybrid GA version. This suggests that KNN, as a distance-based algorithm, tends to be sensitive to dimensionality reduction and may lose certain information when the number of features is reduced. This finding confirms that the effectiveness of feature selection is model-dependent and does not always yield uniform improvements across all machine learning algorithms. A clearer comparison of the results in Table 5 can be seen in the bar chart in Figure 9 below.



(a) (b) (c)

Figure 9. Comparison of: (a) Accuracy, (b) AUC, and (c) F1-Score

The bar charts presented in Figure 9 provide a visual comparison between the Baseline and Hybrid GA models across each algorithm. In the accuracy chart, it is evident that the Hybrid GA bars are higher for Random Forest, SVM, and XGBoost, indicating improved performance following feature selection. The AUC chart reveals a more pronounced contrast, particularly for SVM, where Hybrid GA yields a better ROC curve in distinguishing between positive and negative classes. Meanwhile, the F1-score chart demonstrates that Hybrid GA enhances the balance between precision and recall in most models, especially in Random Forest, making it more suitable for medical applications that demand high accuracy and sensitivity.

3.2 Discussion

This study investigates the effectiveness of a Hybrid Feature Selection (HFS) approach that integrates filter-based methods and Genetic Algorithm (GA) optimization for improving diabetes prediction using machine learning models. The experimental results demonstrate that appropriate preprocessing, feature reduction, and optimization significantly influence model performance, particularly in medical datasets characterized by noise, redundancy, and class imbalance.

The preprocessing stage, including missing value imputation and Min–Max normalization, ensured data consistency and comparability across features with heterogeneous scales. This step was crucial, as several clinical attributes such as Insulin and SkinThickness contained a high proportion of missing values, which could otherwise degrade model stability and predictive reliability. Correlation analysis further revealed redundancy among certain features, highlighting the necessity of feature selection prior to classification.

The hybrid filter stage, combining Chi-square and Mutual Information, successfully identified the most relevant features related to diabetes outcomes. Glucose consistently ranked as the most informative attribute, followed by BMI, Age, and Pregnancies, aligning with established clinical knowledge. These results confirm that the filter stage effectively reduced dimensionality while preserving essential discriminatory information.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

Subsequently, the Genetic Algorithm refined the feature subset by optimizing a fitness function that balances classification accuracy and feature count. The GA showed rapid convergence in early generations and produced a compact subset of features that achieved stable performance. This indicates that the proposed HFS approach can efficiently explore the feature space and avoid suboptimal local solutions.

Model evaluation results indicate that Random Forest achieved the best overall performance after applying HFS, benefiting from its ensemble structure and ability to model non-linear relationships. SVM demonstrated strong discriminative capability as reflected by its high AUC, while XGBoost and KNN showed moderate improvements. Comparative analysis with baseline models confirmed that HFS consistently enhanced performance for most classifiers, although KNN exhibited sensitivity to dimensionality reduction due to its distance-based nature.

Overall, these findings demonstrate that the proposed Hybrid Feature Selection framework effectively improves prediction accuracy, reduces model complexity, and enhances robustness. The approach is particularly suitable for medical decision-support systems, where interpretability, efficiency, and reliable classification are critical.

Despite promising results, this study has several limitations that provide opportunities for future research. First, the experiments were conducted using a single benchmark dataset, which may limit the generalizability of the findings to other populations or real-world clinical settings. Future studies should validate the proposed approach on larger and more diverse medical datasets, including multi-center and longitudinal data. Second, this research focused primarily on Genetic Algorithm as the meta-heuristic optimizer. Although GA demonstrated effective performance, other optimization techniques such as Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), or Whale Optimization Algorithm (WOA) may offer different trade-offs between convergence speed and solution quality. Comparative studies involving multiple meta-heuristics could further strengthen the robustness of the framework. Third, class imbalance was addressed indirectly through evaluation metrics rather than explicit resampling or cost-sensitive learning strategies. Incorporating techniques such as SMOTE, adaptive fitness functions, or imbalance-aware classifiers could further improve sensitivity to minority classes, which is critical in medical diagnosis.

Finally, future work may explore the integration of explainable AI (XAI) methods to enhance the interpretability of selected features and model predictions. This would increase clinical trust and facilitate adoption in real-world healthcare environments.

4. CONCLUSION

This study demonstrates that the proposed Hybrid Feature Selection (HFS) approach based on meta-heuristic optimization effectively enhances the performance of diabetes prediction models. By integrating filter-based feature ranking and Genetic Algorithm optimization, the proposed method successfully reduced feature dimensionality from 8 to 5 key attributes while simultaneously improving classification performance. Experimental results show that the Random Forest classifier achieved an accuracy improvement from 74.03% (baseline) to 79.22% (Hybrid GA), representing an increase of approximately 5.2%, along with improvements in F1-score and AUC. These findings confirm that eliminating irrelevant and redundant features not only increases predictive accuracy but also improves computational efficiency and model robustness, which are critical factors in medical decision-support systems. Despite these promising results, further research is required to validate the generalizability of the proposed approach. Future work will focus on evaluating the HFS framework on larger and more diverse real-world datasets, exploring its integration with deep learning models for automated feature representation, and extending the optimization process toward multi-objective optimization to simultaneously balance accuracy, interpretability, and computational cost.

REFERENCES

[1] Sirmayanti, Pulung Hendro PRASTYO, Mahyati, and Farhan RAHMAN, “A systematic literature review of diabetes prediction using metaheuristic algorithm-based feature selection: Algorithms and challenges method,” *Appl. Comput. Sci.*, vol. 21, no. 1, pp. 126–142, 2025, doi: 10.35784/acs_6849.

[2] A. Singh, N. Prakash, and A. Jain, “Meta-Heuristic Optimization for the Multi-Classification of Chronic Disease: A Review With Machine Learning Perspectives,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 15, no. 3, p. e70030, 2025, doi: 10.1002/widm.70030.

[3] S. Malik *et al.*, “Hybrid metaheuristic optimization for detecting and diagnosing noncommunicable diseases,” *Sci. Rep.*, vol. 15, no. 1, p. 7816, 2025, doi: 10.1038/s41598-025-91136-3.

[4] E. H. Houssein, E. Saber, A. A. Ali, and Y. M. Wazery, “Integrating metaheuristics and artificial intelligence for healthcare: basics, challenging and future directions,” *Artif. Intell. Rev.*, vol. 57, no. 8, p. 205, 2024, doi: 10.1007/s10462-024-10822-2.

[5] A. Dyoub and I. Letteri, “Dataset optimization for chronic disease prediction with bio-inspired feature selection,” *arXiv Prepr. arXiv2401.05380*, 2023, doi: 10.48550/arxiv.2401.05380.

[6] N. Tasnim, S. Al Mamun, M. Shahidul Islam, M. S. Kaiser, and M. Mahmud, “Explainable mortality prediction model for congestive heart failure with nature-based feature selection method,” *Appl. Sci.*, vol. 13, no. 10, p. 6138, 2023, doi: 10.3390/app13106138.

[7] A. Salhi, R. Alshamrani, A. Althbiti, A. Ismail, M. Abd-ElRahman, and B. M. Hassan, “Optimizing high dimensional data classification with a hybrid AI driven feature selection framework and machine learning schema,” *Sci. Rep.*, vol. 15, no. 1,

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2704−2714
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9541

p. 35038, 2025, doi: 10.1038/s41598-025-08699-4.
[8] J. Piri, P. Mohapatra, R. Dey, B. Acharya, V. C. Gerogiannis, and A. Kanavos, "Literature review on hybrid evolutionary approaches for feature selection," *Algorithms*, vol. 16, no. 3, p. 167, 2023, doi: 10.3390/a16030167.
[9] H. Alirezapour, N. Mansouri, and B. Mohammad Hasani Zade, "A comprehensive survey on feature selection with grasshopper optimization algorithm," *Neural Process. Lett.*, vol. 56, no. 1, p. 28, 2024, doi: 10.1007/s11063-024-11514-2.
[10] M. A. S. Ali, P. P. Fathimathul Rajeena, and D. S. Abd Elminaam, "An Efficient Heap Based Optimizer Algorithm for Feature Selection," *Mathematics*, vol. 10, no. 14, p. 2396, 2022, doi: 10.3390/math10142396.
[11] M. H. Nadimi-Shahraki, Z. Asghari Varzaneh, H. Zamani, and S. Mirjalili, "Binary starling murmuration optimizer algorithm to select effective features from medical data," *Appl. Sci.*, vol. 13, no. 1, p. 564, 2022, doi: 10.3390/app13010564.
[12] S. A. Al-Shalif *et al.*, "A systematic literature review on meta-heuristic based feature selection techniques for text classification," *PeerJ Comput. Sci.*, vol. 10, p. e2084, 2024, doi: 10.7717/peerj-cs.2084.
[13] K. H. Abdulkareem, M. A. Mohammed, Z. A. A. Alyasseri, D. Z. Khutar, and O. A. Alomari, "WOA-COVID-19: Whale Optimization Algorithm for Selection of Multi-Examination Features based on COVID-19 Infections," *Mesopotamian J. Comput. Sci.*, vol. 2025, pp. 172–185, 2025, doi: 10.58496/MJCSC/2025/010.
[14] N. Mohd Ali, R. Besar, and N. A. Ab. Aziz, "Hybrid feature selection of breast cancer gene expression microarray data based on metaheuristic methods: A comprehensive review," *Symmetry (Basel).*, vol. 14, no. 10, p. 1955, 2022, doi: 10.3390/sym14101955.
[15] Q. A. Z. Jabbar, "Hybrid Feature Selection Using Secretary Bird Optimization and Decision Tree Classifier," *J. La Multiapp*, vol. 6, no. 3, pp. 631–645, 2025, doi: 10.37899/journallamultiapp.v6i3.2196.
[16] M. S. Salih, R. K. Ibrahim, S. R. Zeebaree, D. Asaad, L. M. Zebari, and N. M. Abdulkareem, "Diabetic prediction based on machine learning using PIMA Indian dataset," *Commun. Appl. Nonlinear Anal.*, vol. 31, no. 5s, pp. 138–156, 2024, doi: 10.52783/cana.v31.1008.
[17] A. A. Ali, G. R. Galal, and H. S. Hassan, "Diabetes Prediction on Pima Indians Dataset Using Machine Learning Techniques," *Int. J. Environ. Sci.*, vol. 11, no. 7, 2025, doi: 10.64252/3a8wqx36.
[18] S. R. Mishra and S. Dash, "Machine Learning Based Diabetes Prediction Using the PIMA Indian Dataset," in *2024 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 2024, pp. 1–6. doi: 10.1109/SCOPES64467.2024.10991027.
[19] Y. D. Pratama and A. Salam, "Comparison of Data Normalization Techniques on KNN Classification Performance for Pima Indians Diabetes Dataset," *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 693–706, 2025, doi: https://doi.org/10.30871/jaic.v9i3.9353.
[20] Y. Guan, C. J. Tsai, and S. Zhang, "Research on Diabetes Prediction Model of Pima Indian Females," in *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science*, 2023, pp. 294–303. doi: 10.1145/3644116.3644168.
[21] P. Verma and A. Khatoon, "Data Mining Applications in Healthcare: A Comparative Analysis of Classification Techniques for Diabetes Diagnosis Using the PIMA Indian Diabetes Dataset," in *2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM)*, 2024, pp. 1–5. doi: 10.1109/ICIPTM59628.2024.10563296.
[22] S. Malakar, S. D. Roy, S. Das, S. Sen, J. D. Velasquez, and R. Sarkar, "Computer based diagnosis of some chronic diseases: a medical journey of the last two decades," *Arch. Comput. Methods Eng.*, vol. 29, no. 7, p. 5525, 2022, doi: 10.1007/s11831-022-09776-x.
[23] A. Abu-Shareha, M. M. Abualhaj, M. A. Alsharaiah, A. Al-Saaidah, and A. Achuthan, "Diabetes Prediction Through Classification Using Pima Dataset: Survey and Evaluation," *J. Soft Comput. Data Min.*, vol. 6, no. 1, pp. 1–20, 2025, doi: 10.30880/jscdm.2025.06.01.001.
[24] G. Pradhan *et al.*, "Optimized forest framework with a binary multineighborhood artificial bee colony for enhanced diabetes mellitus detection," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 194, 2024, doi: 10.1007/s44196-024-00598-2.
[25] G. M. Foody, "Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient," *PLoS One*, vol. 18, no. 10, p. e0291908, 2023, doi: 10.1371/journal.pone.0291908.
[26] A. Humphrey *et al.*, "Machine-learning classification of astronomical sources: estimating F1-score in the absence of ground truth," *Mon. Not. R. Astron. Soc. Lett.*, vol. 517, no. 1, pp. L116–L120, 2022, doi: 10.1093/mnrasl/slac120.
[27] W. Jia, Y. Qin, and C. Zhao, "Rapid detection of adulterated lamb meat using near infrared and electronic nose: A F1-score-MRE data fusion approach," *Food Chem.*, vol. 439, p. 138123, 2024, doi: 10.1016/j.foodchem.2023.138123.
[28] A. M. Carrington *et al.*, "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 329–341, 2022, doi: 10.1109/TPAMI.2022.3145392.
[29] P. Adeodato and S. Melo, "Kolmogorov-Smirnov and ROC curve metrics for binary classification performance assessment are equivalent," in *International Conference on Pattern Recognition (ICPR)*, 2022, pp. 1194–1199. doi: 10.1109/ICPR56361.2022.9956449.
[30] S. M. Malakouti, M. B. Menhaj, and A. A. Suratgar, "The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction," *Clean. Eng. Technol.*, vol. 15, p. 100664, 2023, doi: 10.1016/j.clet.2023.100664.
[31] S. M. Malakouti, "Improving the prediction of wind speed and power production of SCADA system with ensemble method and 10-fold cross-validation," *Case Stud. Chem. Environ. Eng.*, vol. 8, p. 100351, 2023, doi: 10.1016/j.cscee.2023.100351.