Science World Journal Vol. 19(No 4) 2024 https://dx.doi.org/10.4314/swj.v19i4.48
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

AN IMPROVED DIABETES MELLITUS PREDICTION MODEL THROUGH ENSEMBLE LEARNING AND GINI INDEX-BASED FEATURE SELECTION

[1]Rukkayya Yahaya Ibrahim, [1]Sahabi A. Yusuf, [1]Mohammed Abdullahi, *[2]Jeremiah Isuwa

[1]Computer Science Department, Ahmadu Bello University, Zaria - Nigeria
[2]Computer Science Department, Federal University Kashere, Gombe - Nigeria

*Corresponding Author Email Address: isuwajeremiah@fukashere.edu.ng

ABSTRACT

Diabetes Mellitus (DM) is a condition where the body cannot regulate blood sugar due to improper insulin production or use, posing a significant global health burden. Traditional detection methods rely on clinical assessments and basic lab tests, but recent technological advancements suggest that Machine Learning (ML) algorithms can predict DM more effectively and efficiently. However, current ML models face challenges like feature redundancy, irrelevancy, and dataset imbalance, which can reduce accuracy and interpretability, ultimately affecting patient outcomes. This paper aims to address these challenges by developing an enhanced ML-based DM prediction model. The proposed model leverages an ensemble soft voting classifier, integrating the Random Forest, Logistic Regression, and Naïve Bayes algorithms. Feature importance determination is facilitated by the Gini Index Random Forest (GI-RF) algorithm. Additionally, three data imbalance handling techniques random oversampling (ROS), random undersampling (RUS), and the synthetic minority oversampling technique (SMOTE) are employed to mitigate biased model development. Initially, the GI-RF algorithm identifies the top 5 most informative features from the PIMA Indians Diabetes Dataset, originally comprising 8 features. Subsequently, the dataset is subjected to each of the three imbalance handling techniques. The performance of each model variation, incorporating different imbalance handling techniques is then extensively compared. The results demonstrate that ROS notably outperforms RUS and SMOTE across multiple metrics, including accuracy, F1 score, recall, and AUC. A comparative analysis with existing studies reveals the proposed method's notable improvements across all metrics, with increases of 5% in accuracy, 8% in precision, 13% in F1 score, 18% in recall, and 4% in AUC. This demonstrates the proposed model's overall robustness and effectiveness in predictive modeling, contributing to more accurate diagnosis and treatment of DM.

Keywords: Diabetes Mellitus, Feature Importance Determination, Soft Voting Machine Learning, Imbalance Datasets, PIMA Indians Diabetes Dataset.

INTRODUCTION

The global health burden of Diabetes Mellitus (DM) is clear, as evidenced by its persistent prevalence and significant mortality rates (Mansoori *et al*., 2023). Characterized by chronic metabolic dysregulation and hyperglycemia originating from inadequate insulin production or utilization (Azeez *et al*., 2023; Chang *et al*., 2023), DM's impact extends far beyond individual health, contributing substantially to worldwide mortality figures, with over 400 million afflicted individuals and an annual death toll exceeding a million (Azeez *et al*., 2023). Alarming statistics highlight DM's deathly strength, surpassing combined mortality rates of other major diseases, including COVID-19, HIV/AIDS, cancer, and tuberculosis (Yusuf *et al*., 2023). Factors underlying DM's causal, as reported by Chang *et al*., (2023), encompass a complex interplay of genetic tendency and environmental triggers, including ethnicity, age, obesity, dietary habits, smoking, and unhealthy lifestyles.

In Nigeria specifically, DM's widespread impact is striking, affecting over 3.6 million individuals with a prevalence rate of 3.7%, according to the International Diabetes Federation (IDF) (Yusuf *et al*., 2023). Left unchecked, DM triggers a series of cardiovascular complications, accelerating atherosclerosis and augmenting the risk of morbidity and mortality (Azeez *et al*., 2023). Nonetheless, early detection holds the key to effective disease management, mitigating health risks, and averting possible complications such as cardiovascular disease, kidney disease, eye complications, and nerve damage amongst others (Azeez *et al*., 2023). Therefore, predicting an individual's susceptibility to DM has become a crucial area of research in biomedicine, carrying profound implications for health outcomes, healthcare costs, and related matters (Chang *et al*., 2023).

Traditionally, the detection of DM relied heavily on clinical assessments, patient-reported symptoms, and basic laboratory tests (Mushtaq *et al*., 2022). Physicians assessed patients for classic signs associated with diabetes, including polyuria, polydipsia, unexplained weight loss, fatigue, and blurred vision (Mushtaq *et al*., 2022). Urinalysis for glucose detection was a common practice, although with limitations in accuracy, particularly in detecting early-stage diabetes (Yusuf *et al*., 2023). Clinicians also checked for diabetes-related problems like eye damage, slow wound healing, and frequent infections to confirm the diagnosis (Yusuf *et al*., 2023).

However, with technological advancements, the utilization of Machine Learning (ML) algorithms for DM prediction has gained prominence (Tan *et al*., 2023). ML algorithms such as Support Vector Machines (SVM), Random Forest (RF), Neural Networks (NN), Naïve Bayes (NB), and K-Nearest Neighbor (KNN) leverage large datasets comprising demographic profiles, medical histories, and diagnostic results to develop predictive models for identifying individuals at risk of developing diabetes or experiencing associated complications (Dritsas and Trigka, 2022; Mustofa *et al*., 2023; Ramadhan Nur Ghaniaviyanto *et al*., 2021).

Although ML algorithms trained on large medical datasets have profound predictive abilities, using all features, including irrelevant and redundant ones, increases computational complexity, overfitting, and reduces model interpretability. Hence, the importance of the Feature Importance Measure (FIM) in these datasets before applying ML algorithms cannot be overstated (Isuwa *et al*., 2023). Effective FIM aids in pinpointing the most impactful variables contributing to DM prediction, thereby enhancing algorithm accuracy and efficiency (Laila *et al*., 2022; Ramadhan *et al*., 2021). Additionally, the abundance of features may exacerbate data imbalance issues, a common occurrence in medical datasets due to the rareness of certain health conditions, ethical constraints, and the design of clinical trials among others, complicating the identification of meaningful patterns or relationships, particularly within the minority class (Elseddawy *et al*., 2022; Liu *et al*., 2023; Sadeghi *et al*., 2022).

Examples of FIM methods applied include filter methods utilizing statistical measures for individual feature evaluation, such as correlation analysis (Deng *et al*., 2022), Relief (Jun Dou *et al*., 2022), and Chi-Square (Isuwa *et al*., 2023). The wrapper method assesses a subset of features using a learning algorithm i.e., Sequential Forward and Backward Searches (Huda and Banka, 2020), Particle Swarm Optimization (Kennedy and Eberhart, 1995), and Genetic Algorithms (Holland, 1984). Finally, the embedded method that evaluates the relevance of features directly within the model training process i.e., Lasso (L1 regularization) and Ridge (L2 regularization) regression (Sahu *et al*., 2018; Wang *et al*., 2024), and the Gini-Index Random Forest (GI-RF) (Algehyne *et al*., 2022). Consequently, numerous studies have tackled the challenges of feature importance and data imbalance in DM datasets. For instance, Dritsas and Trigka, (2022) employed various ML algorithms, including RF, SVM, and NB, to predict the likelihood of DM disease. They utilized four filter methods i.e., Pearson correlation, Gain Ratio, Naïve Bayes, and Gini Index RandomForest to determine feature importance in the widely used PIMA Indians Diabetes datasets and applied the Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance. Similarly, Kumari *et al*., (2021) utilized a soft voting classifier, integrating RF, Logistic Regression (LR), and NB for DM prediction. Despite surpassing base classifiers and previous studies, their model exhibited notably low accuracy, precision, F1 score, recall, and AUC. These shortcomings originated from overlooking critical factors such as appropriate data preprocessing, including feature importance determination, and neglecting the dataset's severe imbalance. Despite the dataset containing only eight variables, potential redundant or irrelevant features could impair the model's predictive capacity. Moreover, the substantial class imbalance, with the majority class nearly twice the size of the minority class in the employed PIMA Indians Diabetes dataset, risked biasing the model against minority instances, resulting in inaccurate predictions.

Therefore, this paper aims to address the limitations of prior research by utilizing the soft voting classifier proposed by Kumari *et al*., (2021) and integrating the FIM stage utilizing the GI-RF algorithm to identify the most crucial features from the PIMA Indians Diabetes dataset. This approach seeks to enhance model interpretability and mitigate overfitting. Additionally, to combat challenges associated with biased models, inaccurate metrics, and limited generalization, the study proposes to rectify dataset imbalance through experimentation with three widely used techniques: random undersampling (RUS), oversampling (ROS), and the SMOTE method. These efforts are directed towards enhancing the overall performance and reliability of DM disease prediction models. Specifically, we undertake the following:

1. Implement techniques to handle data imbalance to mitigate biased model development challenges.
2. Determine feature importance using an appropriate algorithm to select only significant features crucial for the model's overall performance.
3. Assess the performance of the proposed model using the PIMA Diabetes Mellitus benchmark dataset, comparing it with studies from the literature, focusing on classification metrics including accuracy, precision, F1 score, recall, and AUC.

A. Feature Importance Measure (FIM)

FIM refers to a technique used in ML to determine the significance or relevance of different features or variables in a dataset for predicting a target outcome (Algehyne *et al*., 2022). It helps identify which features have the most influence on the model's predictions and can assist in FS, where only the most important features are retained for model training (Algehyne *et al*., 2022). Various methods are employed to measure feature importance, including statistical approaches like correlation analysis, as well as algorithm-specific techniques such as decision tree-based methods like Gini impurity or information gain (Isuwa *et al*., 2023). Understanding feature importance is crucial for building accurate and interpretable ML models, especially in complex datasets with numerous variables. Three techniques, namely filter, wrapper, and ensemble methods, are utilized for measuring feature importance.

i. Filter method of feature importance measure

This method utilizes statistical techniques like correlation analysis (Deng *et al*., 2022), relief (Jun Dou *et al*., 2022), mutual information (Song *et al*., 2021), and chi-square (Isuwa *et al*., 2023) to assess and rank individual features according to their importance. Each feature is assigned a score indicating its level of importance, with higher scores indicating greater significance (Isuwa *et al*., 2023). These techniques are commonly recognized for their effectiveness in managing large datasets (Alrefai and Ibrahim, 2022).

ii. Wrapper method of feature importance measure

In contrast to the filter method, the wrapper method assesses a subset of features using a learning algorithm (Bandyopadhyay *et al*., 2023; Chen *et al*., 2023). Wrapper algorithms can be categorized as heuristic or metaheuristic. Heuristic methods involve intelligent guesses during subset selection, which may not always yield optimal results (Isuwa *et al*., 2022). Examples include Sequential Forward and Backward searches. On the other hand, metaheuristics aim to improve the performance of heuristic methods (Isuwa *et al*., 2022). Inspired by natural phenomena, they are considered intelligent algorithms. Examples include the Particle Swarm Algorithm (Kennedy and Eberhart, 1995), the Genetic Algorithm (Holland, 1984), and Ant Colony Optimization (Gao, 2020), among others.

iii. Embedded method of feature importance measure

The embedded method of FIM is a technique used in ML to evaluate the relevance of features directly within the model training process (Algehyne *et al*., 2022). Unlike filter methods, which assess features independently of the learning algorithm, embedded methods integrate feature selection into the model's training procedure. This approach enables the model to determine the most important features while simultaneously optimizing its performance on the given task (Algehyne *et al*., 2022).

One common example of an embedded method is regularization techniques, such as Lasso (L1 regularization) and Ridge (L2 regularization) regression (Sahu *et al*., 2018; Wang *et al*., 2024).

https://dx.doi.org/10.4314/swj.v19i4.48

These methods penalize the coefficients of less important features during model training, effectively reducing their impact on the final predictions (Sahu *et al.*, 2018). By doing so, the model automatically selects the most informative features while mitigating the risk of overfitting. Another example of an embedded method is the Gini Index Random Forest (Dritsas and Trigka, 2022). These algorithms inherently assess feature importance by evaluating how much each feature contributes to reducing impurity or increasing information gain at each node of the tree (Dritsas and Trigka, 2022). Features with higher importance scores are more likely to be selected for splitting nodes, while less important features are pruned away during the tree-building process.

iv. Gini Index Random Forest embedded method of feature importance measure

The GI-RF is a bagging ensemble that uses binary decision trees as its underlying classifiers. (Algehyne *et al*., 2022). In this method, feature importance is determined based on the Gini impurity measure, which is a criterion used in decision trees to evaluate the purity of a split (Algehyne *et al*., 2022). The Gini impurity measures the degree of misclassification at a particular node in a decision tree. A lower Gini impurity indicates that a node contains predominantly samples from a single class, making it purer (Ramadhan, Adiwijaya, Romadhony, *et al*., 2021). In the context of RF, the Gini impurity is used to evaluate the importance of each feature when making decisions at each node of the individual trees within the forest (Algehyne *et al*., 2022). The feature importance is calculated based on how much the Gini impurity decreases when a particular feature is used for splitting the data at each node (Algehyne *et al*., 2022). The GI-RF method aggregates the feature importance scores from all the individual trees in the forest to determine the overall importance of each feature as shown in Figure 1. Features that consistently lead to greater decreases in Gini impurity across multiple trees are deemed more important, while those that have little impact on impurity reduction are considered less important (Algehyne *et al*., 2022).

The measure of impurity importance referred to as the mean decrease in impurity, is denoted by the Gini Index algorithm, which is according to Gini theory. This theory asserts that given a collection of samples S with k classes $(C_i, i = 1,2,3,\dots,k)$, S can be partitioned into k subsets based on class distinctions. Let S_i represent a subset containing samples belonging to the class C_i, and s_i denote the number of samples in the subset S_i. The Gini Index of S can then be computed using Equation 1 (Algehyne *et al*., 2022).

$$Gini(S) = 1 - \sum_{i=1}^{m} P_i^2 \qquad (1)$$

Here, P_i represents the probability estimated as $\frac{s_i}{s}$ for any sample belonging to C_i. A Gini index of 0 indicates the minimum impurity, where all members in the set belong to the same class, signifying maximum useful information. Conversely, a maximum Gini index suggests the minimum useful information can be obtained.

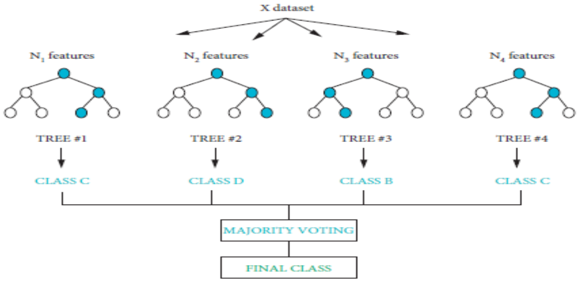

Figure 1: Process of GI-RF feature importance measure (Beghriche *et al*., 2021)

B. Imbalanced Data

Imbalanced data, characterized by heavily skewed class distributions where one or more classes significantly outnumber others, presents challenges for ML algorithms (Werner de Vargas *et al*., 2023). These algorithms often exhibit bias towards the majority class, resulting in suboptimal performance in predicting minority classes (Aruna and Nandakishore, 2022). Two basic types of approaches exist for handling imbalanced data: algorithm-driven and data-driven methods (Aruna and Nandakishore, 2022). The choice between these approaches depends on factors such as dataset characteristics, available computational resources, and the desired balance between predictive performance and interpretability (Viloria *et al*., 2020). Effectively addressing imbalanced data is crucial for developing robust ML models across various domains like healthcare, finance, and fraud detection. Common techniques for handling data imbalance include resampling (ROS, RUS) and SMOTE.

i. Algorithm-driven approach

Algorithm-driven approaches entail adapting the learning algorithm such as the Adaboost to better address class imbalance (Aguiar *et al*., 2023). This includes techniques such as cost-sensitive learning, where misclassification costs are adjusted to impose heavier penalties on errors within the minority class (Aguiar *et al*., 2023). Additionally, employing ensemble methods like bagging and boosting, which combine multiple models to enhance predictive accuracy on imbalanced datasets, is another viable approach.

ii. Data-driven approach

Conversely, data-driven approaches concentrate on adjusting the dataset to achieve a balanced class distribution (Werner de Vargas *et al*., 2023). A prevalent technique involves resampling, where the minority class is either oversampled to boost its presence or the majority class is undersampled to diminish its prevalence (Werner de Vargas *et al*., 2023). Oversampling methods include ROS and SMOTE, while undersampling techniques comprise RUS and NearMiss (Sharifai and Zainol, 2020).

- **Random oversampling (ROS) technique**

This method entails randomly augmenting the instances in the minority class to align with the majority class (Viloria *et al*., 2020). Although ROS is straightforward to execute, it may result in overfitting and failure to introduce new information.

- **Random undersampling (RUS) technique**

On the other hand, RUS randomly deletes the instances in the majority class to achieve balanced class distributions (Viloria *et al*., 2020). It removes instances from the majority class randomly until class equilibrium is attained. While RUS can alleviate computational complexity and processing time, it may cause the loss of significant information inherent in the majority class, potentially leading to underfitting and diminished model performance (Aruna and Nandakishore, 2022).

- **Synthetic minority oversampling technique (SMOTE)**

SMOTE is a widely used oversampling method developed to overcome the drawbacks of basic oversampling techniques like ROS (Aruna and Nandakishore, 2022). Instead of replicating instances from the minority class, SMOTE creates synthetic instances by interpolating between existing minority class samples (Aruna and Nandakishore, 2022). It starts by selecting a random minority class instance and identifying its k nearest neighbors. Then, a new instance is created by selecting a random point along the line connecting the chosen instance and one of its neighbors (Viloria *et al*., 2020). This process iterates until the desired balance between the minority and majority classes is attained.

C. Machine Learning

i. Logistic regression (LR) ML algorithm

LR is a statistical method used for binary classification tasks (Dritsas and Trigka, 2022). It models the relationship between independent variables (features) and a binary dependent variable (target) using the logistic function, which transforms the output into a probability score between 0 and 1 (Kumari *et al*., 2021). It is simple, interpretable, and efficient for large datasets (Kumari *et al*., 2021). While it assumes a linear relationship between features and the log odds of the target, it may not perform well with highly imbalanced classes or nonlinear decision boundaries (Kumari *et al*., 2021). Nonetheless, it is widely used in fields like healthcare, marketing, and finance for its effectiveness in predicting binary outcomes (Kumari *et al*., 2021).

ii. Naive Bayes (NB) ML algorithm

NB is a simple yet powerful probabilistic classification algorithm based on Bayes' theorem with an assumption of independence between features. It is widely used for text classification tasks, spam filtering, and recommendation systems (Mushtaq *et al*., 2022). NB calculates the probability of each class given a set of features using Bayes' theorem. It assumes that the presence of a particular feature in a class is independent of the presence of other features, hence the term "naive." Despite this simplifying assumption, NB often performs remarkably well in practice, especially with large datasets (Mushtaq *et al*., 2022). One of the main advantages of NB is its speed and efficiency, making it suitable for real-time prediction tasks. It also requires a small amount of training data to estimate the parameters necessary for classification (Kishor and Chakraborty, 2021). However, NB may not perform well if the independence assumption does not hold for the data, or if features are highly correlated (Mushtaq *et al*., 2022). Additionally, it is not suitable for tasks requiring a detailed understanding of relationships between features (Mushtaq *et al*., 2022).

iii. Random Forest (RF) ML algorithm

RF is a versatile and powerful ensemble learning algorithm used for classification and regression tasks. It operates by constructing multiple decision trees during training and outputting the mode (for classification) or mean prediction (for regression) of the individual trees (Sadeghi *et al*., 2022). Each decision tree in RF is trained on a random subset of the training data and a random subset of the features (Sadeghi *et al*., 2022). This randomness helps to de-correlate the trees and reduces the risk of overfitting. During prediction, the output of the RF is determined by aggregating the predictions of all the individual trees. In classification, it uses a majority vote, while in regression, it takes the average of the individual tree predictions (Sadeghi *et al*., 2022). RF is known for its high accuracy, robustness to overfitting, and ability to handle large datasets with high dimensionality. It can also provide estimates of feature importance, making it valuable for understanding the underlying relationships in the data. However, RF tends to be computationally intensive and may not be as interpretable as simpler models like decision trees. Additionally, it may not perform well on datasets with highly imbalanced classes or when there are strong interactions between features (Sadeghi *et al*., 2022).

iv. Ensemble soft voting classifier

Ensemble soft classifiers are a type of ensemble learning method that combines multiple base classifiers to make predictions, where each base classifier outputs probabilities rather than discrete class labels as in hard voting classifiers (García-Ordás *et al*., 2021; Kumari *et al*., 2021). These probabilities represent the confidence of the classifier in assigning each instance to different classes. In ensemble soft classifiers, the final prediction is made by aggregating the probabilities produced by the individual base classifiers (García-Ordás *et al*., 2021). This aggregation can be done in various ways, such as averaging the probabilities or using more sophisticated techniques like stacking or gradient boosting. One of the key advantages of ensemble soft classifiers is their ability to capture more nuanced information about the data compared to traditional hard classifiers, which only output discrete class labels (García-Ordás *et al*., 2021). By considering the probabilities assigned to each class, ensemble soft classifiers can provide more insight into the uncertainty associated with each prediction. Moreover, ensemble soft classifiers tend to be more robust to noise and outliers in the data, as they can incorporate information from multiple classifiers to make a more informed decision (Kumari *et al*., 2021). Figure 2 shows how a hard and soft voting classifier works.

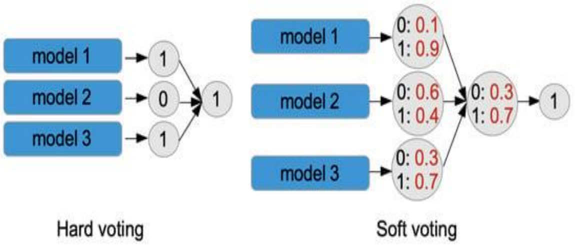

Figure 2: An illustration of a hard and soft voting classifiers (Manconi *et al*., 2022)

D. Related Works Numerous studies have explored the application of feature importance measures and methods for handling imbalanced data in predicting DM disease. These methods use either one or a collection of learning algorithms. Additionally, the use of ensemble methods has been profound. For example, Ramadhan *et al*., (2021) integrated the random oversampling technique to address the imbalanced data. Additionally, the Gini Index Random Forest was employed for feature importance assessment and data augmentation. Subsequently, both the RF and LR algorithms were utilized for classification. The results revealed a notable improvement of 20% and 24%, respectively. Kumari *et al*., (2021) employ a soft voting classifier, combining three ML algorithms for Diabetes prediction. Experimental results demonstrate that the model achieved improved classification performance for new instances of Diabetes patients. However, the model utilized all 9 features without assessing the potential presence of irrelevant or redundant features. Additionally, the dataset exhibited severe imbalance, with no attempts made to address it, significantly compromising the overall effectiveness of the model.

Beghriche *et al*., (2021) employed a DNN to classify Diabetes disease. Subsequently, the model's performance was assessed against various other learning algorithms. Additionally, grid search was implemented for hyperparameter tuning across the ML algorithms utilized in the study. Findings indicated that the DNN surpassed other algorithms in performance. Kishor and Chakraborty, (2021) employs the Fast Correlation statistical measure to select pertinent features, effectively discarding irrelevant ones. Subsequently, the SMOTE technique is applied to address the imbalance in the dataset. The refined dataset is then subjected to various ML algorithms, with the RF demonstrating superior performance. Ergün and O.İlhan, (2021) utilized multiple ML algorithms to classify diabetes disease, employing 5-fold cross-validation for each algorithm. The findings indicate the CNN model's superiority in the classification task. Ghosh *et al*., (2021) study applies four learning algorithms to classify the PIMA diabetes dataset, initially utilizing all features and subsequently employing the Minimum Redundancy Maximum Relevance (mRmR) statistical measure for FS. Findings consistently demonstrate the RF's superior performance in both approaches. García-Ordás *et al*., (2021) introduce a pipeline founded on Deep Learning methodologies to predict DM disease. It incorporates data augmentation via a variational autoencoder (VAE), feature augmentation using a sparse autoencoder (SAE), and classification via a convolutional neural network (CNN). Employing a comprehensive deep-learning pipeline for both data preprocessing and classification has exhibited considerable promise in diabetes detection, surpassing existing state-of-the-art proposals.

Dritsas and Trigka, (2022) employed the use of 14 ML algorithms for the classification of Diabetes disease. The study utilized four feature importance ranking methods: Pearson correlation, Gain Ratio, Naïve Bayes, and Random Forest. Additionally, the SMOTE technique was employed to address data imbalance. Experimental results indicate that KNN and RF outperformed other methods in terms of accuracy. Laila *et al*., (2022) utilized the chi-square statistical method to choose relevant features from a pool of 17 features. Subsequently, three ensembles i.e., AdaBoost, Bagging, and RF approaches were employed for the classification task, among which the RF consistently demonstrated superiority in classifying diabetes disease. However, the study did not include information regarding the size and distribution of the dataset and its corresponding classes. Consequently, the reliability of the study's findings is questionable. Mushtaq *et al*., (2022) employed correlation analysis to select pertinent features from the PIMA diabetes dataset, with Tomek and SMOTE techniques utilized to address data imbalance. Subsequently, various learning algorithms, finely tuned through multiple K-folds, were employed for classification, with RF demonstrating superiority. Additionally, a soft voting classifier, combining NB, RF, and Gradient Boosting, outperformed the initial experiment phase for classification tasks.

Sadeghi *et al*., (2022) apply Deep Neural Networks (DNN), Extreme Gradient Boosting (XGBoost), and RF for diabetes classification tasks. Acknowledging the data's imbalanced nature, both upsampling and downsampling techniques were employed to mitigate negative impacts on overall performance and generalization. Findings reveal that with the original imbalanced dataset, the DNN outperforms XGBoost and RF significantly. Conversely, with the balanced dataset, RF demonstrates superior performance. Mustofa *et al*., (2023) use a single ML algorithm i.e., the RF with the PIMA Indians Diabetes dataset. Thorough hyperparameter tuning was carried out on the RF algorithm using k-fold cross-validation, varying k values of 3, 5, 7, and 10. Additionally, experimentation with different numbers of trees was undertaken to enhance performance. However, the study overlooked the imbalanced dataset and neglected to perform FS, potentially compromising the robustness and generalizability of the findings.

THE PROPOSED DM PREDICTION METHOD

A. Dataset Description

The PIMA Diabetes Mellitus dataset, sourced from the UCI repository, is a widely recognized benchmark dataset extensively utilized for assessing ML models. It has been cited in numerous studies, including works by Mustofa *et al*., (2023), Kumari *et al*., (2021), and Mushtaq *et al*., (2022), among others. It comprises 768 samples from both positive and negative cases with 9 clinical features as described in Tables 1.

Table 1: The PIMA Diabetes Mellitus Dataset

	Instances	Features	Classes
Majority class (Negative)	500	9	Binary(0,1)
Minority class (Positive)	268		
Total	**768**	**9**	

One thing of note is its significant class imbalance, with 500 samples (>65%) representing the negative class and only 268 (<35%) representing the positive class. This pronounced class imbalance highlights the importance of addressing class imbalance issues throughout model development and evaluation processes.

B. Proposed Method Design

Figure 3 presents the overall architecture of the proposed method.

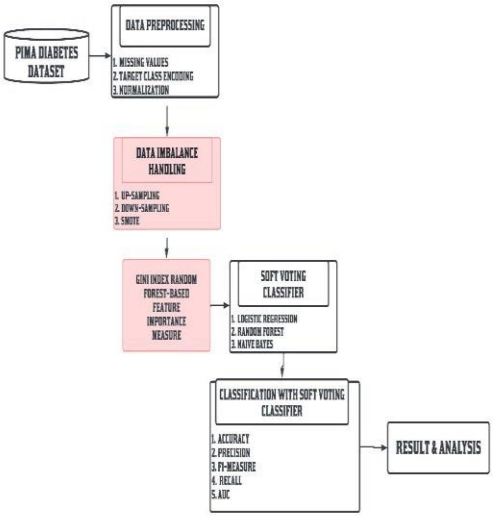

Figure 3: Architecture of the proposed method of DM prediction

i. Data Preprocessing

The process starts with the original PIMA Diabetes Mellitus benchmark dataset outlined in subsection 3.2. Extensive preprocessing was conducted on the datasets, involving tasks such as imputing missing values, converting categorical values to numeric, and ultimately normalizing the features to standardize them within a specified range, typically ranging between 0 and 1.

ii. Feature Importance Measure Using the GI-RF

Following the preprocessing of the data to prepare it for ML training, the GI-RF FIM was employed to evaluate the importance of each feature. Figure 4 illustrates a bar chart depicting the ranking of features in the dataset, with those considered most significant receiving higher scores. Consequently, for this study, we focused solely on the top 5 features: Glucose, BMI, Age, DiabetesPedigreeFunction, and BloodPressure. This selection aligns with previous studies in the literature, as observed in Algehyne *et al*., (2022).

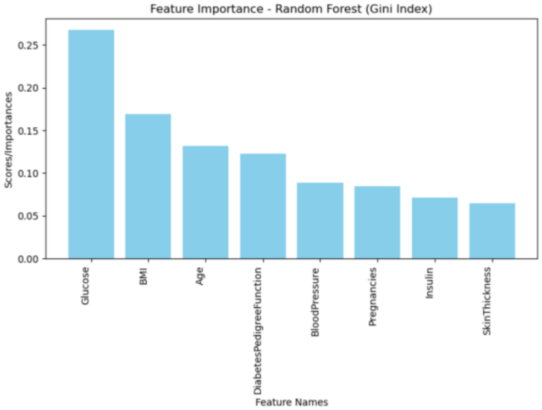

Figure 4: Scores of individual features using the GI-RF

iii. Data Imbalance Handling

The subset of the dataset comprising only the top 5 features, was subsequently processed individually through ROS, RUS, and SMOTE techniques to tackle its imbalance problem. Employing their respective methods, ROS and SMOTE augmented the minority class to generate a new dataset featuring 100 samples and 5 features. Conversely, RUS generated a new dataset comprising 536 samples and 5 features.

iv. Prediction Using the Soft Voting Classifier

The datasets, balanced using RUS, ROS, and SMOTE, were subsequently fed into the soft voting classifier proposed by Kumari *et al*., (2021) for training and prediction. The soft voting classifier is from a combination of three ML algorithms i.e., RF, LR, and NB. The output from each of these algorithms is combined through a weighted average, where the weights are determined during the training phase based on the individual performance of each algorithm on the dataset. This collective decision-making process enhances the overall predictive capability of the classifier, leveraging the strengths of each underlying ML algorithm.

v. Evaluation and Statistical Test:

The metrics covered below are used to evaluate each of the three developed approaches, and the T-test and Standard Deviation (SD) are used to determine the methods' statistical significance and stability, respectively.

i. **Accuracy:** calculates the percentage of cases that are correctly classified out of all instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

ii. **Precision:** measures the accuracy of positive predictions made by a classification mode

$$Precision = \frac{TP}{TP + FP} \qquad (3)$$

iii. **F1-score:** calculates the precision and recall harmonic mean.

$$F1 = \frac{2 * TP}{2 * TP + FN + FP} \qquad (4)$$

iv. **Recall:** calculates the percentage of accurately recognized true positive cases.

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

v. **Area Under Curve (AUC):** measures the model's ability to rank true positives higher than false positives across different threshold values,

where TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives respectively.

vi. Parameter Values and Settings

The parameters and configurations employed in the experiments for each of the ML algorithms, as well as other crucial configurations, are outlined in Table 2. These settings adhere to standard practices established in the literature.

Table 2: Parameter values and settings utilized in the proposed method

Random Forest Classifier	$n\_estimators = 100$
Logistic Regression	*Max_iteration = 1000*
Naive Bayes	$Defualt\ settings$
Other Settings	$train\_test\_spilt = 70{:}30$, $number\ of\ algorithm\ run = 10$

I. EXPERIMENTS AND RESULTS DISCUSSION

In this section, we present the outcomes derived from our conducted experiments and engage in a thorough discussion regarding them. Note that all experiments in this study are conducted in a Jupyter Notebook Integrated Development Environment (IDE). The implementation process is performed using the Python programming language on a computer that has an Intel(R) Core(TM) i7-6600U CPU with a processing speed of 2.80 GHz and a RAM capacity of 8.00 GB.

The three versions of the proposed model i.e., each with a different data imbalance handling technique, are compared against each other, employing all metrics discussed in sub-section 3.2.5. Subsequently, the model with superior performance is chosen for further comparison with existing studies in the literature. Each result has its SD displayed next to it, denoted by the notation $x \pm y$. where $'y'$ denotes the associated result's standard deviation (SD) value and $'x'$ denotes the average result (in percentage) over ten separate runs. The best outcomes across all measurement categories are indicated using boldface.

A. Performance Comparison of the Three Different Data Imbalance Handling Techniques.

Table 3 presents the performances of the three versions of the proposed method, each employing a different data imbalance handling technique. Notably, ROS demonstrates significantly superior performance across four metrics including accuracy (84%), F1 score (84%), recall (88%), and AUC(84%), with a slight exception in precision (81%) as also shown in Figure 5. Moreover, it has been established that the average performance of ROS across all metrics surpasses that of RUS and SMOTE by 10% and 6% respectively. This higher performance of ROS can be attributed to its efficacy in augmenting the minority class instances, thereby mitigating the class imbalance and enabling the model to better capture the underlying patterns in the data.

Table 3: Comparison of performance among three data imbalance handling techniques using RF-GI selected features.

Metrics	ROS (%)	RUS (%)	SMOTE (%)
Accuracy	**84.00±1.17E-16**	75.16±1.17E-16	78.33±1.17E-16
Precision	81.48±0	**81.69±2.34E-16**	77.56±0
F1	**84.62±1.17E-16**	74.36±1.17E-16	78.83±1.17E-16
Recall	**88.00±1.17E-16**	68.24±1.17E-16	80.13±1.17E-16
AUC	**84.00±1.17E-16**	75.57±0	78.32±0

Furthermore, ROS preserves information by replicating existing minority class instances, ensuring that original data points are retained and reducing the risk of information loss present in RUS and potential noise introduction in SMOTE. It is also straightforward and computationally efficient, as it involves randomly duplicating minority class samples without the need for complex calculations involved in SMOTE. Also, it reduces the risk of overfitting by directly duplicating minority class instances, thus avoiding the introduction of potentially noisy synthetic samples as in SMOTE. Additionally, ROS maintains the original class distribution, providing a balanced representation that helps classifiers learn effectively from both classes, unlike RUS which may lead to the underrepresentation of important patterns within the majority class.

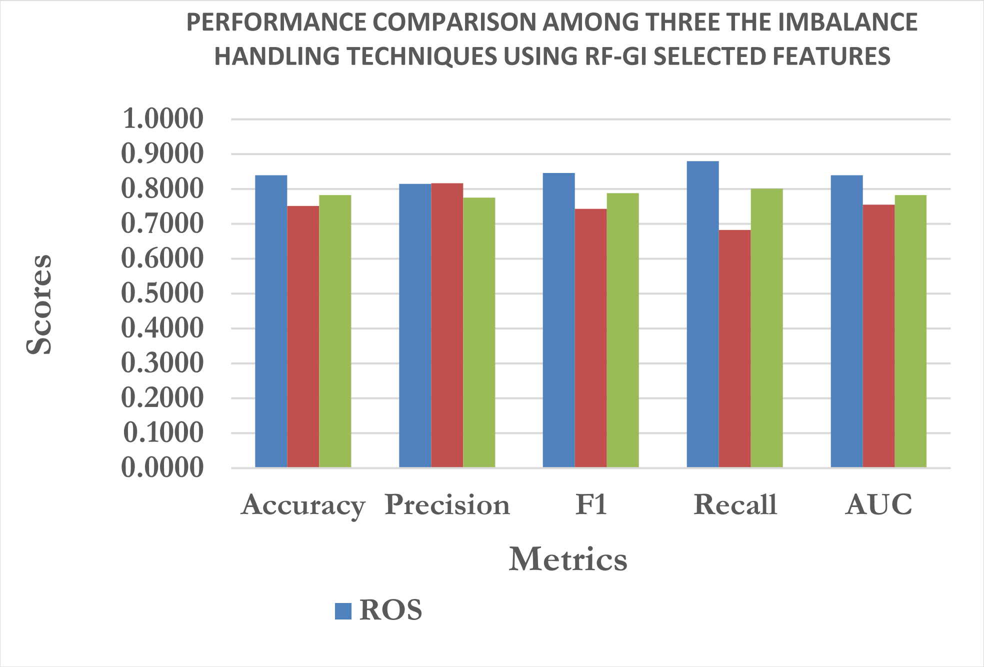

Figure 5: Performance comparison among the three imbalance handling techniques.

To reinforce the superiority of ROS, a T-test was conducted to determine the presence of a statistically significant difference between these performances. Employing a critical value of 0.05, p-values below this threshold were deemed indicative of statistical significance. Table 4 presents the p-value obtained from pairwise comparisons of the mean performances of the methods. Notably, all methods demonstrate statistically significant differences with p-values significantly below the 0.05 threshold.

Table 4: Statistical analysis results of T-tests demonstrating significant differences among three data handling techniques utilizing RF-GI selected features

Accuracy	**ROS Vs. RUS**	**ROS Vs SMOTE**	**RUS Vs SMOTE**
	2.938E-264	8.7037E-261	3.0559E-256
Precision	**ROS Vs. RUS**	**ROS Vs SMOTE**	**RUS Vs SMOTE**
	4.2733E-118	5.9701E-261	9.7094E-130
F1	**ROS Vs. RUS**	**ROS Vs SMOTE**	**RUS Vs SMOTE**
	2.0116E-265	5.9701E-261	6.2902E-259
Recall	**ROS Vs. RUS**	**ROS Vs SMOTE**	**RUS Vs SMOTE**
	1.5137E-270	2.3804E-263	1.4156E-266
AUC	**ROS Vs. RUS**	**ROS Vs SMOTE**	**RUS Vs SMOTE**
	3.0834E-135	1.0773E-133	1.4156E-266

In general, employing ROS for addressing data imbalance alongside GI-RF selected features demonstrates superior performance compared to alternative methods i.e., RUS and SMOTE. Consequently, it has been chosen for further comparison with existing studies discussed in the literature.

B. Performance Comparison of the proposed method with existing methods in the literature

Table 5 showcases the comprehensive comparison of the proposed method with four other studies extensively discussed in the literature. Empty fields from the table mean the authors of the respective studies did not employ the use of the specific metrics in their study. Thus, we only carried out a comparison with those metrics in which their scores are recorded. The basis of comparison of our method with these studies lies in the use of one or more ML algorithms for DM prediction, the utilization of feature importance measures, and data imbalance handling techniques. Most importantly, the use of the PIMA Indians Diabetes Dataset as a benchmark dataset.

As observed from the table, the proposed methods surpassed their competitors in terms of accuracy and AUC, both achieving a score of 84%. However, concerning the recall, our method attained a competitive score comparable to that of Ramadhan, Adiwijaya, and Romadhony (2021), with both reaching a score of 88%. Regarding the F1 score, the study by Ramadhan, Adiwijaya, and Romadhony (2021) attained the highest score of 86%. Lastly, concerning precision, the research conducted by Mustofa *et al*. (2023) notably outperforms all others, achieving a score of 87%.

Table 5: Performance comparison of the proposed method with the existing method in the literature.

Methods	Accuracy (%)	Precision (%)	F1(%)	Recall (%)	AUC (%)
Proposed Method	**84.00**	81.48	84.62	**88.00**	**84.00**
(Mustofa *et al*., 2023)	75.00	**87.00**	-	80.00	-
(Mushtaq *et al*., 2022)	81.50	-	-	-	81.50
(Kumari *et al*., 2021)	79.08	73.13	71.56	70.00	80.98
(Ramadhan, Adiwijaya, and Romadhony, 2021)	-	83.00	**86.00**	**88.00**	-

Another thing of note can be observed in the F1 score and recall metrics which are particularly useful in evaluating the performance of models on imbalanced datasets because they prioritize the correct identification of instances from the minority class. The proposed method demonstrates a notable performance enhancement in these metrics compared to the approach of Kumari *et al*., (2021). Specifically, there is a substantial difference of over 8% and 9% between its accuracy, F1, and recall scores respectively. These disparities indicate that our method effectively addresses the imbalanced nature of the dataset, achieving precise scores of 84% in both accuracy and F1, along with an even higher score of 88% in the recall.

Moreover, the results recorded from all evaluated metrics highlight the importance of feature importance measures compared to utilizing the entire dataset. This is revealed in the comparison of the proposed method's performance with that of Kumari *et al*., (2021), where they utilized the complete dataset without prior feature selection. The results across all metrics demonstrate a notable increase, including accuracy (5%), precision (8%), F1 score (13%), recall (18%), and AUC (4%) as shown in Figure 6.

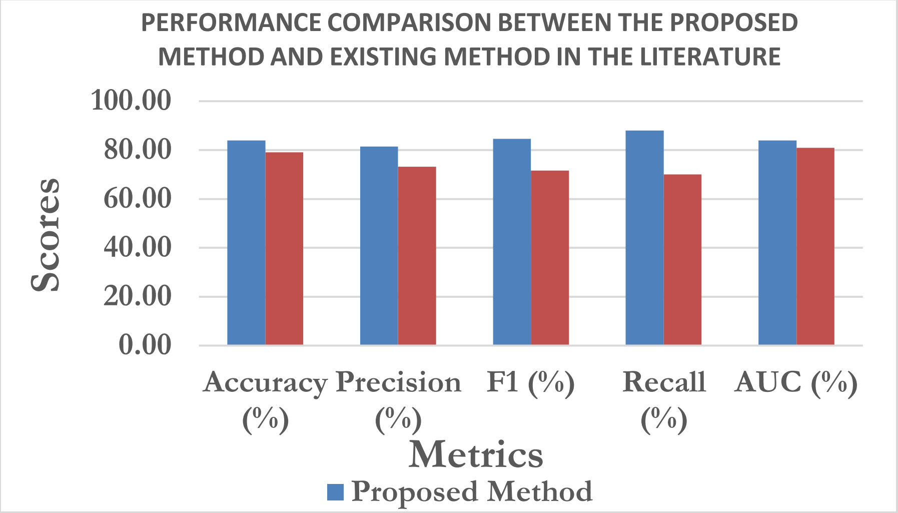

Figure 6: Performance comparison between the proposed method and that of Kumari *et al*., (2021)

Overall, the proposed method which combines the GI-RF for feature importance measurement with ROS to address data imbalance for DM prediction holds significant importance in several aspects. Firstly, the utilization of the Gini index algorithm within the random forest provides a robust means of determining feature importance, ensuring that the most relevant features are identified for predictive modeling. This enhances the accuracy and reliability of the predictive model by focusing on the most informative attributes. Additionally, the integration of ROS techniques effectively mitigates the issue of imbalance in the employed dataset. By generating synthetic samples of the minority class, the imbalance is addressed, thereby preventing the predictive model from being biased towards the majority class and improving its ability to accurately predict instances of DM.

Conclusion and Future Works

The research aims to develop a more accurate DM prediction model using an ensemble soft voting classifier, incorporating Random Forest, Logistic Regression, and Naïve Bayes algorithms. Specifically, the study focuses on utilizing the Gini Index Random Forest algorithm (GI-RF) to determine feature importance. Additionally, it evaluates three imbalance handling techniques: Random Oversampling (ROS), Random Undersampling (RUS), and Synthetic Minority Oversampling Technique (SMOTE). Initially, the GI-RF algorithm is employed to select the top 5 most informative features from the PIMA Indians Diabetes Dataset, which originally comprised 8 features. Subsequently, the subset of data containing only these 5 selected features is subjected to each of the three data imbalance handling techniques individually. This process aims to assess the effectiveness of each method in balancing the dataset and improving model performance. Finally, the performance of each variation of the proposed method, incorporating different imbalance handling techniques, is intensively compared. Evaluation metrics such as accuracy, precision, recall, F1 score, and AUC are utilized for this comparison. Additionally, the performance of the proposed method is benchmarked against four existing studies in the literature, providing further context and validation.

Based on the findings, it is recommended to adopt ROS as the preferred technique for handling data imbalance in DM prediction models. Future research should explore additional ML algorithms and feature importance measures to further enhance model performance. Moreover, conducting real-world validation studies to assess the clinical applicability and generalizability of the proposed model would provide valuable insights for healthcare decision-making.

REFERENCES

Aguiar, G., Krawczyk, B., and Cano, A. (2023). A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. *Machine Learning*, 1–79.

Algehyne, E. A., Jibril, M. L., Algehainy, N. A., Alamri, O. A., and Alzahrani, A. K. (2022). Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia. *Big Data and Cognitive Computing*, *6*(1).

Alrefai, N., and Ibrahim, O. (2022). Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets. *Neural Computing and Applications 2022*, 1–16.

Aruna, S., and Nandakishore, L. V. (2022). Empirical Analysis of the Effect of Resampling on Supervised Learning Algorithms in Predicting the Types of Lung Cancer on Multiclass Imbalanced Microarray Gene Expression Data. *EAI/Springer Innovations in Communication and Computing*, 15–27.

Azeez, T. A., Durotoluwa, I. M., and Makanjuola, A. I. (2023). Diabetes Mellitus as a risk factor for stroke among Nigerians: A systematic review and meta-analysis. *International Journal of Cardiology: Cardiovascular Risk and Prevention*, *18*(February), 200189.

Bandyopadhyay, R., Das Sharma, A., Dasgupta, B., Ghosh, A., Das, C., and Bose, S. (2023). A New hybrid Feature selection-Classification model to Improve Cancer Sample Classification Accuracy in Microarray Gene Expression Data. *2023 International Conference on Computer, Electrical and Communication Engineering (ICCECE)*, 1–7.

Beghriche, T., Djerioui, M., Brik, Y., Attallah, B., and Belhaouari, S.

B. (2021). An Efficient Prediction System for Diabetes Disease Based on Deep Neural Network. *Complexity*, *2021*.

Chang, V., Bailey, J., Xu, Q. A., and Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, *35*(22), 16157–16173.

Chen, T. C., Alizadeh, S. M., Albahar, M. A., Thanoon, M., Alammari, A., Guerrero, J. W. G., Nazemi, E., and Eftekhari-Zadeh, E. (2023). Introducing the Effective Features Using the Particle Swarm Optimization Algorithm to Increase Accuracy in Determining the Volume Percentages of Three-Phase Flows. *Processes 2023, Vol. 11, Page 236*, *11*(1), 236.

Deng, X., Li, M., Wang, L., and Wan, Q. (2022). RFCBF: Enhance the Performance and Stability of Fast Correlation-Based Filter. *International Journal of Computational Intelligence and Applications*, *21*(2).

Dritsas, E., and Trigka, M. (2022). Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. *Circulation*, *134*(19), 1441–1443.

Elseddawy, A. I., Karim, F. K., Hussein, A. M., and Khafaga, D. S. (2022). Predictive Analysis of Diabetes-Risk with Class Imbalance. *Computational Intelligence and Neuroscience*, *2022*.

ERGÜN, Ö. N., and O.İLHAN, H. (2021). Early Stage Diabetes Prediction Using Machine Learning Methods. *European Journal of Science and Technology*, *29*, 52–57.

Gao, W. (2020). *New Ant Colony Optimization Algorithm for the Traveling*. *13*(1), 44–55.

García-Ordás, M. T., Benavides, C., Benítez-Andrades, J. A., Alaiz-Moretón, H., and García-Rodríguez, I. (2021). Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Computer Methods and Programs in Biomedicine*, *202*.

Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., and Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. *Procedia Computer Science*, *192*, 467–477.

Holland, J. H. (1984). *GENETIC ALGORITHMS AND ADAPTATION*. 317–333.

Huda, R. K., and Banka, H. (2020). New efficient initialization and updating mechanisms in PSO for feature selection and classification. *Neural Computing and Applications*, *32*(8), 3283–3294.

Isuwa, J., Abdullahi, M., and Abdulrahim, A. (2022). *Hybrid particle swarm optimization with sequential one point flipping algorithm for feature selection*. *July*, 1–18.

Isuwa, J., Abdullahi, M., Ali, Y. S., Kim, J., Hassan, I. H., and Buba, J. R. (2023). Optimizing Microarray Cancer Gene Selection using Swarm Intelligence : Recent Developments and An Exploratory Study Optimizing Microarray Cancer Gene Selection using Swarm Intelligence : Recent. *Egyptian Informatics Journal*, *24*(4), 100416.

Jun Dou, Song, Y., Wei, G., and Zhang, Y. (2022). Fuzzy information decomposition incorporated and weighted Relief-F feature selection: When imbalanced data meet incompletion. *Information Sciences*, *584*, 417–443.

Kennedy, J., and Eberhart, R. (1995). Particle swarm optimization. *Proceedings of ICNN'95 - International Conference on Neural Networks*, *4*, 1942–1948.

Kishor, A., and Chakraborty, C. (2021). Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE. *International Journal of Systems Assurance Engineering and Management*.

Kumari, S., Kumar, D., and Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*(January), 40–46.

Laila, U. e., Mahboob, K., Khan, A. W., Khan, F., and Taekeun, W. (2022). An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study. *Sensors*, *22*(14), 1–15.

Liu, Y., Liu, Y., Yu, B. X. B., Zhong, S., and Hu, Z. (2023). Noise-robust oversampling for imbalanced data classification. *Pattern Recognition*, *133*, 109008. https://doi.org/10.1016/J.PATCOG.2022.109008

Manconi, A., Armano, G., Gnocchi, M., and Milanesi, L. (2022). A Soft-Voting Ensemble Classifier for Detecting Patients Affected by COVID-19. *Applied Sciences (Switzerland)*, *12*(15).

Mansoori, A., Sahranavard, T., Hosseini, Z. S., Soflaei, S. S., Emrani, N., Nazar, E., Gharizadeh, M., Khorasanchi, Z., Effati, S., Ghamsary, M., Ferns, G., Esmaily, H., and Mobarhan, M. G. (2023). Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: a cohort study analysis. *Scientific Reports*, *13*(1), 1–11.

Mushtaq, Z., Ramzan, M. F., Ali, S., Baseer, S., Samad, A., and Husnain, M. (2022). Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques. *Mobile Information Systems*, *2022*.

Mustofa, F., Safriandono, A. N., Muslikh, A. R., and Setiadi, D. R. I. M. (2023). Dataset and Feature Analysis for Diabetes Mellitus Classification using Random Forest. *Journal of Computing Theories and Applications*, *1*(1), 41–48.

Ramadhan, N. G., Adiwijaya, Romadhony, A., Lu, H., Uddin, S., Hajati, F., Moni, M. A., Khushi, M., Mustofa, F., Safriandono, A. N., Muslikh, A. R., Setiadi, D. R. I. M., Howlader, K. C., Satu, M. S., Awal, M. A., Islam, M. R. A., Islam, S. M. S., Quinn, J. M. W., Moni, M. A., … Xu, J. (2021). Efficient diabetes mellitus prediction with grid based random forest classifier in association with natural language processing. *Complexity*, *22*(3), 1441–1443.

Ramadhan Nur Ghaniaviyanto, Adiwijaya, and Romadhony, A. (2021). Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus based on Random Forest. *International Journal of Advanced Computer Science and Applications*, *12*(7), 223–228.

Sadeghi, S., Khalili, D., Ramezankhani, A., Mansournia, M. A., and Parsaeian, M. (2022). Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods. *BMC Medical Informatics and Decision Making*, *22*(1), 1–12.

Sahu, B., Dehuri, S., and Jagadev, A. (2018). A Study on the Relevance of Feature Selection Methods in Microarray Data. *The Open Bioinformatics Journal*, *11*(1), 117–139.

Sharifai, G. A., and Zainol, Z. (2020). Feature selection for high-dimensional and imbalanced biomedical data based on robust correlation based redundancy and binary grasshopper optimization algorithm. *Genes*, *11*(7), 1–26.

Song, X. fang, Zhang, Y., Gong, D. wei, and Sun, X. yan. (2021).

Feature selection using bare-bones particle swarm optimization with mutual information. *Pattern Recognition*, *112*, 107804.

Tan, K. R., Seng, J. J. B., Kwan, Y. H., Chen, Y. J., Zainudin, S. B., Loh, D. H. F., Liu, N., and Low, L. L. (2023). Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A Systematic Review. *Journal of Diabetes Science and Technology*, *17*(2), 474–489.

Viloria, A., Lezama, O. B. P., and Mercado-Caruzo, N. (2020). Unbalanced data processing using oversampling: Machine Learning. *Procedia Computer Science*, *175*, 108–113.

Wang, S., Chen, Y., Cui, Z., Lin, L., and Zong, Y. (2024). *Diabetes Risk Analysis based on Machine Learning LASSO Regression Model*. *4*(1), 58–64.

Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., and Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems*, *65*(1), 31–57.

Yusuf, A., O, A., DO, S., A, L., T, B., and M, O. (2023). DIABETES MELLITUS CARE IN NIGERIA: THE HEIGHTS AND THE HURDLES. *The Official Journal of LAUTECH …*, *2*.