Journal of Artificial Intelligence in Medical Applications (JAIMA)
Vol. 1, No. 1, 2025, **pp**. 116–125
ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.742

Enabled-Detection Framework for Diabetic Disease based on Machine Learning Approach and Indian Pima Case Study

Karrar Hameed Abdulkareem[1,*], Zainab Hussein Arif [2], Mohammed Al-Mhiqani[3], Salem Bafjaish[4]

[1]College of Agriculture, Al-Muthanna University, Samawah 66001, Iraq; Khak9784@mu.edu.iq
[2]College of Nursing, University of Al-Qadisiyah , Al-Qadisiyah Province,58002, Iraq; zhussian94@gmail.com
[3]School of Computing and Engineering, University of Huddersfield, Huddersfield, UK, M.Al-Mhiqani@hud.ac.uk
[4]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar; saba56819@hbku.edu.qa

Received: 05/09/2025, Revised: 30/10/2025, Accepted: 18/11/2025, Published: 30/12/2025

ABSTRACT: Diabetes has become a growing health concern worldwide, posing serious risks to people's well-being. The condition develops when blood sugar levels remain consistently high, often as a result of factors such as physical inactivity, poor dietary habits, excess body weight, and other related influences. This study proposed a framework for detection of diabetes and non-diabetes cases based on machine learning approach and Pima Indian diabetes dataset (PID). The dataset consist of eight main indicators for detection within 764 samples. Random Forest (RF) is proposed as machine learning model for detection process. Due to avoid computational load and improve detection performance a feature selection method called Information Gain (IG) is used. Furthermore, due to misleading results challenge generated by unbalanced data of PID. This research employs SMOTE (Synthetic Minority Over-sampling Technique) to handle unbalanced data issue. The result of this study indicates that detection process is more effective based on feature selection method and balanced dataset. To mention, RF model scored 84% for accuracy, precision, recall, and F-score. Beside no over-fitting or underfitting issue is observed.

Keywords: Diabetes Disease, Random Forest, Diabetes Detection, Machine Learning, Pima Indian Dataset.

1. INTRODUCTION

Diabetes (DB) is a chronic and debilitating disease that places a significant treatment cost burden on healthcare systems worldwide [1]. In Type 1 DB, impaired beta-cell function in the pancreas leads to insufficient insulin production, resulting in persistent hyperglycemia. In Type 2 DB, the body is unable to effectively utilize the insulin that is available. Both types are associated with severe clinical complications, including neurological damage, retinal degeneration, kidney disease, and cardiovascular disorders [2]. Early detection and timely preventive measures are therefore crucial to lowering mortality and mitigating the long-term impact of the disease.

In recent years, machine learning (ML) algorithms have been increasingly applied to the prediction and early detection of diabetes, offering an alternative to conventional diagnostic approaches [3]. By analyzing data from routine physical examinations, ML models can support preliminary assessments and provide valuable insights for healthcare professionals. With the rising global prevalence of diabetes, a range of ML methods—including Random Forests (RF) [4], regression models [5], and ensemble techniques—have been developed for this purpose. Feature selection has also emerged as a critical step for improving efficiency and boosting classification accuracy [6]. To this end, approaches such as Support Vector Machines (SVM), Decision Trees (DT), Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA) have been widely used [7].

Investigating and optimizing ML algorithms for diabetes prediction is essential to improve accuracy and to enable preventive action before the onset of severe complications. Beyond technical improvements, ML frameworks can also support healthcare governance and resource management, ultimately benefiting patients, clinicians, telehealth systems, and hospital administrators [8]. Building on this foundation, the present research aims to enhance ML-based models for predicting the progression of diabetes, with particular emphasis on improving accuracy and other performance metrics. Prior studies have shown that ML has achieved promising results in predicting chronic diseases such as diabetes across various statistical measures, underscoring the importance of early detection and accurate classification [9]. For instance, one study employing the RF approach reported an accuracy of 75.32% when classifying diabetic datasets using Automatic Identification System techniques [10]. The authors in[8] implemented different machine learning approach on pima dataset such as logistic regression and RF classifiers. However, RF classifier scored only 77.4% as classification accuracy indicator. In the study[11], authors proposed RF classifier based on two scenarios with PIMA dataset. First, classification Process based on all feaures.Second classification based on subset of features .However, RF in both scenarios did not exceeded 79.5%. In [12] accurate results have been obtained which proves using the

proposed Bayes network to predict Type-2 diabetes is effective .However, the best accuracy is obtained based on pima dataset and mentioned classifier is 72.3%.

The study in [13] explored diagnostic solutions for the disease by examining data patterns and applying classification analysis using the Naïve Bayes algorithm. Nevertheless, NB achieved only 79% as accuracy rate. In[14]authors indicated that the machine learning method focus on classifying diabetes disease from high dimensional medical dataset such as Pima dataset. Therefore, they proposed SVM as a machine learning method as the classifier for diagnosis of diabetes. Nonetheless, SVM got only 78% of accuracy for data classification. A recent study [15] utilized the PIMA dataset retrieved from the Kaggle repository and applied Principal Component Analysis (PCA) as part of the preprocessing stage. The findings showed that the Random Forest (RF) model achieved strong results, with an accuracy of 80%, precision of 82%, error rate of 20%, and sensitivity of 88%.

In contrast, many earlier studies using the PID dataset reported relatively poor classification performance. This limitation can largely be attributed to the insufficient use of data preprocessing techniques prior to building machine learning models, particularly when working with a limited number of samples. As a result, the outcomes were often suboptimal. To address this gap, we emphasize the importance of exploratory data analysis as a means to enhance the quality of data used in prediction models. Moreover, although balancing datasets is a critical step for improving predictive performance, this aspect has been overlooked in much of the existing research. Thus, the main contributions of this work as follows:

- Propose a methodology of diabetics detection based on Pima Indian dataset and RF model.
- Improve the detection performance based on Information Gain (IG) method as pattern for feature selection process.
- Propose a SMOTE (Synthetic Minority Over-sampling Technique) algorithm for handling unbalancing data issue relevant to PID dataset.

2. THE PROPOSED METHODOLGY

The proposed framework begins with the original imbalanced dataset, which undergoes preprocessing steps such as data normalization to ensure uniformity in scale and distribution. To reduce computational overhead and improve detection accuracy, feature selection techniques including Information Gain (IG) and ANOVA are applied. This step helps identify the most relevant indicators for diabetes detection, ensuring that the classification models focus only on meaningful attributes. After feature selection, classification is performed using machine learning models, specifically the Random Forest (RF) classifier and SVM with RBF kernel, which are chosen for their strong predictive capabilities. The performance of these classifiers is then assessed using evaluation metrics such as accuracy, confusion matrix, and ROC analysis. To address the issue of class imbalance inherent in the dataset, the framework incorporates data sampling techniques. Over-sampling approaches, including the ADASYN algorithm, are applied to generate synthetic samples for the minority class, thereby producing a more balanced dataset. This balanced dataset is reintroduced into the classification pipeline, leading to improved performance across accuracy, recall, and other evaluation measures. By combining preprocessing, feature selection, robust classification models, and advanced sampling strategies, the framework offers a systematic and effective solution for enhancing diabetes detection. In addition, the modular structure of the framework ensures that each stage can be independently refined or replaced with newer methods as technology evolves. For example, more advanced classifiers or feature selection techniques could be integrated without disrupting the overall workflow. This flexibility makes the framework not only effective for diabetes detection but also adaptable for other medical diagnostic tasks where data imbalance and feature relevance are critical issues. By emphasizing scalability and adaptability, the study highlights the potential of the framework as a generalizable tool for medical decision support systems. Figure 1 displays the interconnected phases of proposed framework.

 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.742

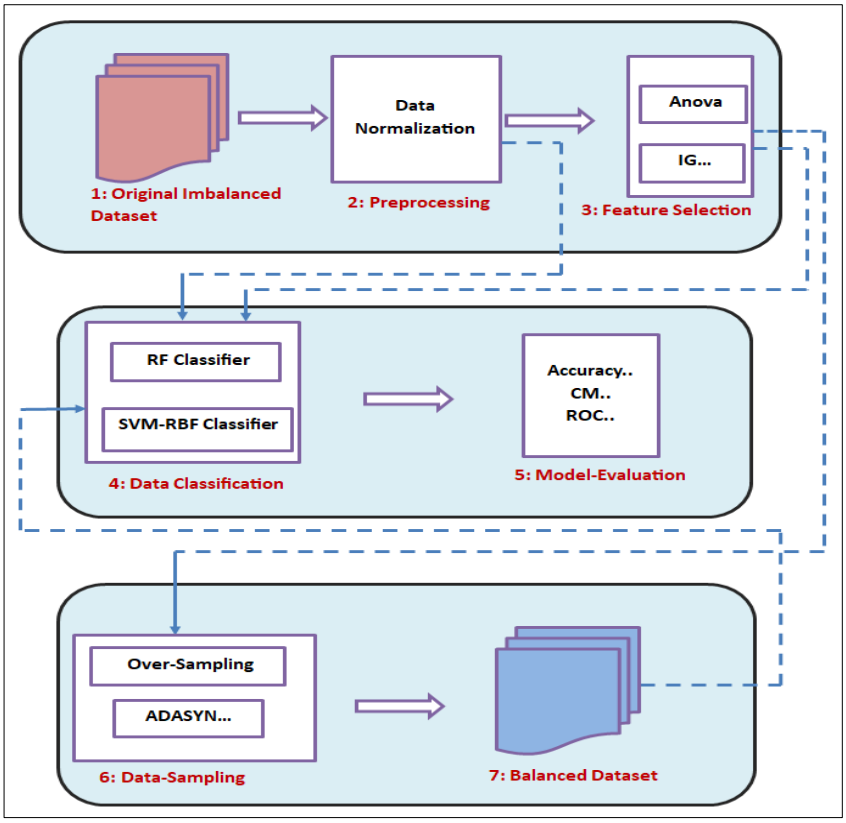
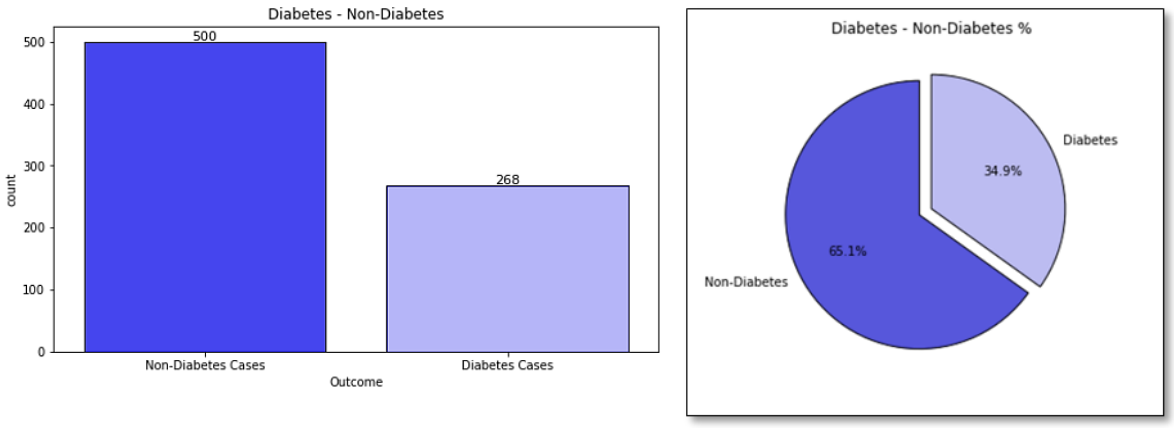


Figure 1. Proposed Framework for Diabetic Detection

2.1 Dataset

The proposed classification framework was trained and tested using a benchmark dataset known as the Pima Indian Diabetes dataset, administered by the National Institute of Diabetes and Digestive and Kidney Diseases [16]. The dataset was obtained from Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database), where it is publicly available under a CC0: Public Domain License. It is fully anonymized and does not include any personally identifiable information about the subjects. The dataset contains 769 records with eight diagnostic features: number of pregnancies, blood glucose concentration, blood pressure, skinfold thickness, insulin level, body mass index (BMI), diabetes pedigree function (DPF), and age. The outcome variable, indicating the presence or absence of diabetes, was used as the target class for prediction. Furthermore, the number of people with diabetes and non-diabetes is displayed Figure 2.



(a) (b)

Figure 2. Diabetes vs Non-diabetes casesto (a) Number of samples per each class and (b) Percentages of diabetes vs non-diabetes

2.2 Preprocessing:

Standardization, also known as Z-score normalization, is applied to rescale features so that they follow a normal distribution with a mean of zero and a variance of one. As shown in Equation (1), this process also helps reduce skewness in the data distribution.

$$st(d)\frac{d-d^{-}}{\propto} \tag{1}$$

Where d is the n-dimensional instances of the feature vector, $d \in R^n, d^- \in R^n,$ and $\propto \in R^n$; are the standard deviation and mean of attributes[17].

2.3 Feature Selection:

In this study, a feature selection technique is employed to identify the most relevant attributes for the classification process. The approach is based on Information Gain (IG) [18], a concept derived from Information Theory [19], which evaluates and ranks attributes according to their contribution to the prediction of the target class. Unlike traditional correlation measures such as the Pearson Linear Correlation Coefficient [20], which can only capture linear relationships, IG is capable of detecting both linear and nonlinear dependencies between attributes. The mathematical formulation of IG is presented as follows:

$$IGA(X|Y) = H(X) - H(X|Y) \tag{2}$$

Therefore, a feature X is strongly correlated to feature Y than to feature V if IG(X | Y) > G(V |Y)[21].

2.4 Classification models

2.4.1. Support vector machine (SVC)

Support Vector Classification (SVC) is a widely used and efficient supervised learning algorithm, known for its strong generalization ability in both classification and regression tasks [22, 23]. The method works by constructing a hyperplane that separates data points, with the closest points to this boundary referred to as support vectors. In the case of linear SVC, the algorithm distinguishes between two classes in an *n*-dimensional space using a maximum-margin hyperplane of dimension *n–1*. Among the possible separating hyperplanes, the one that maximizes the margin between classes is selected. In practice, separating data points is not always straightforward, as some observations may fall within an overlapping or "grey" region. To handle such cases, SVC introduces a regularization parameter that allows a balance between maximizing the margin and minimizing classification errors. Furthermore, SVC employs kernel functions—including linear, polynomial, sigmoid, and radial basis function (RBF) kernels—to map data from lower-dimensional to higher-dimensional spaces, enabling the algorithm to effectively manage non-linear decision boundaries [24].

2.4.2. Random forest (RF)

Random Forest (RF) [25] is a widely used ensemble learning method that supports classification, regression, and various other predictive tasks. The technique builds a collection of decision trees, each trained on different subsets of the data, to collectively solve a given problem. In constructing a decision tree, RF selects node splits randomly from the *n* best candidates, thereby introducing diversity among trees. For prediction, each tree produces an output, and the final result is obtained through aggregation, typically by majority voting for classification or averaging for regression [24]. In regression problems, RF generates *N* independent regression trees, denoted as $h_n(s)$, based on the input variable *s*. The overall prediction is then calculated as the average of the outputs from all *N* trees. To further enhance diversity and reduce correlation among trees, RF employs a bootstrapping technique, where each tree is trained on a randomly resampled subset of the data [26]. This process strengthens the robustness of the model, reduces overfitting, and improves generalization. The mathematical formulation is expressed in Equation (3).

$$RF - \text{classification} = \frac{1}{N}\sum_{N=1}^{N} h_n(n) \tag{3}$$

2.5 SMOTE algorithm.

The Synthetic Minority Over-sampling Technique (SMOTE), introduced by Chawla [27], is a widely used method for addressing class imbalance in datasets. Instead of simply duplicating minority samples, SMOTE generates new synthetic examples by performing random linear interpolation between an existing sample and its nearest neighbors. By creating these artificial minority samples, the algorithm reduces the imbalance ratio and improves the performance of classifiers on skewed datasets [28]. The generation process can be expressed as:

$$P_{ij} = x_i + rand(0,1) \times (x_{ij} - x_i) \quad (4)$$

where *rand(0,1)* represents a random number uniformly distributed between 0 and 1. This process continues until the desired level of balance between classes is achieved [29].

3. RESULTS AND DISCUSSION

This section presents the results based on adopted methodology into previous section. The results divided into different scenarios. Mainly results of detection based on raw data of Pima Indian diabetes dataset. Then results based on features selection method with balanced and unbalanced dataset. All results are evaluated based on well-known metrics such as accuracy, precision, recall, F-score, ROC, and confusion matrix. Furthermore, to generalize the results and avoid bias scenario we implemented SVM classifier within RFB function in each detection scenario.

3.1 All Features and Unbalanced Dataset Results

The results in this section are obtained based on detection with all eight features (Pregnancies, Glucose, BloodPressure, Insulin, BMI, SkinThickness ,DiabetesPedigreeFunction, Age, Outcome).The details of detection for diabetes is displayed into Table 1.

Table 1. Results of detection based on all Features

Classifier	No of features	Accuracy
RF	8	74.6%
Rbf svm	8	76%

According to Table 1, both RF and RBF-SVM have showed poor detection performance based on all features of PID dataset. However, RBF-SVM is outperformed RF into this scenario.

3.2: Results of Diabetes Detection based on Unbalanced Dataset and Features Selection

The features selection process is very important step during feature engineering phase. This study we have adopted Information Gain (IG) method as features selector method. For more generalize and non-bias results we have implement ANOVA method and compare it results with IG. The results of feature selection methods are presented into Table 2.

Table 2. Feature Selection Methods

	#	ANOVA		#	Gain ratio
1	Glucose	213.162	1	Glucose	0.085
2	BMI	71.772	2	Age	0.041
3	Age	46.141	3	BMI	0.039
4	Pregnancies	39.670	4	Insulin	0.030
5	DiabetesPedigreeFunction	23.871	5	Pregnancies	0.021
6	Insulin	13.281	6	SkinThickness	0.018
7	SkinThickness	4.304	7	DiabetesPedigreeFunction	0.011
8	BloodPressure	3.257	8	BloodPressure	0.007

As showed into Table 2, IG gives the highest weight for Glucose indictor while lowest score for BloodPressure. Features such as Age, BMI, Insulin, and Pregnancies also have scored good value according to IG preference. However, each of SkinThikness and DiabetesPedigreeFucntion feature have scored low preference rate. However, compare to results of IG, ANOVA method should same performance in terms of best and worst diabetes indicator. Therefore, the results of IG acceptable.

According to results of IG, the detection task is implemented only on five features and we have excluded each of SkinThickness, BloodPressure and DiabetesPedigreeFunction in this phase.

The full details for results of diabetes detection in this phase are presented into Table 3, Figure 3, and Figure 4.

Table 3. Results of detection based on unbalanced dataset

Model	No of features	Accuracy	Precision	Recall	F-score
RF	5	75%	69%	70%	70%
SVM-RBF	5	77%	73%	70%	71%

Table 3, showed the detection results based on selected features. There is a small ratio of improvements in each of RF and SVM. However, the performance of these classifiers based on other evaluation metrics is not adequate. For instance, the gap between Precision that link to how many cases are positive and accuracy metric reaches 6% which is quite high. On other hand, F-score metric that reveals harmonic mean of precision and recall have same ratio of gap within accuracy measure. Thus, there is a misleading result that not reflects the real performance of selected detection model. All other insights about the problem of detection based on unbalanced dataset are cleared through ROC and confusion matrix figures.

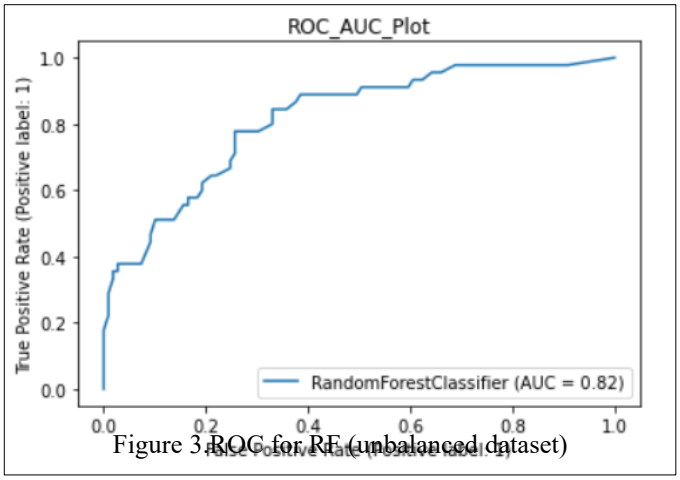


Figure 3 ROC for RF (unbalanced dataset)

Figure 3. ROC curve for RF (unbalanced dataset)

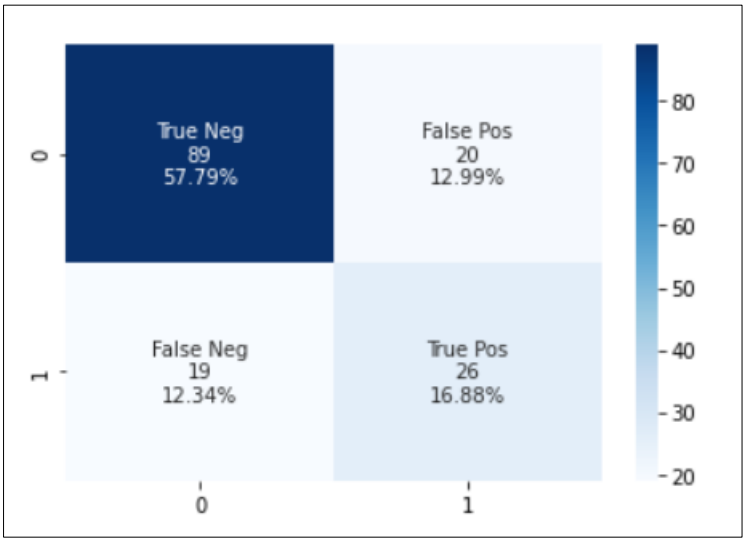


Figure 4. Confusion matrix for RF (unbalanced dataset)

3.3: Results of Diabetes Detection based on balanced Dataset and Features Selection

According to figure 2, the difference into number of samples between diabetes and non-diabetes cases is quite high. Where only 268 diabetes samples were founded while 500 samples are provided as non-diabetes cases. This issue may lead to misleading classification results and over-fitting problem that well-known into machine learning area. However, in order to solve this issue we have proposed Synthetic Minority Over-sampling Technique (SMOTE).The results SMOTE are showed into Figure 3.

 ISSN: 3106-1494 | **DOI**: https://doi.org/10.65511/jaima.v1i1.742

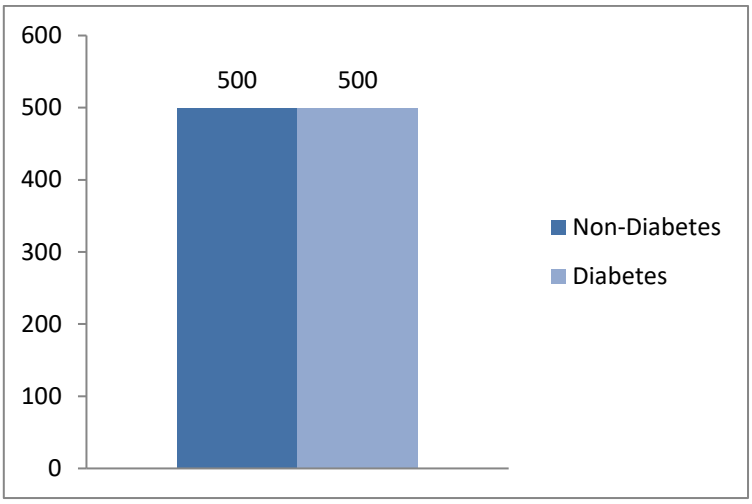


Figure 5. Balanced dataset based on SMOTE.

According to Figure 3, the issue with class diabetes case has resolved where this class has increased to contain 500 cases. Therefore, mentioned class had equal samples with non-diabetes class.
Same set of selected features that applied in non-balanced scenario have used into this stage. The full details for results based on different classification measures are presented into Table 4,Figure 6,and Figure 7.

Table 4. Results of detection based on balanced dataset

Model	No of features	Accuracy	Precision	Recall	F-score
RF	5	**84%**	**84%**	**84%**	**84%**
SVM-RBF	5	79%	78%	78%	78%

Table 4 showed the improvement level of detection performance based on balanced dataset using SMOTE algorithm. For instance the accuracy of SVM-RBF have increased by 3%.However, the significant impact appeared into RF detection model where improvement ratio reached to 9%.Furthermore,the highest detection rate scored is 84% by RF in evaluation metrics and there is no misleading results have noticed. Therefore, the best detection model for diabetes and non-diabetes cases is RF. Also, as shown into Figure 6 the performance of RF words detection of true positive have improved significantly comparing to detection level based on unbalanced dataset.

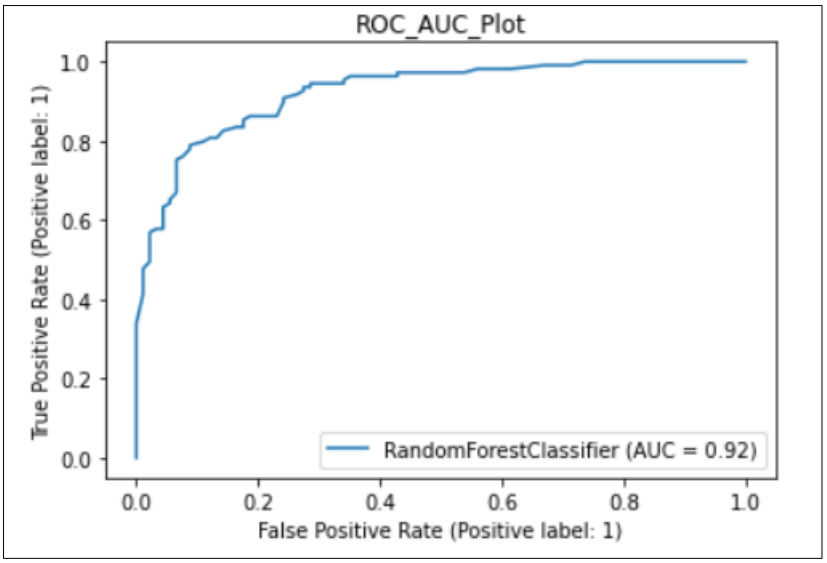


Figure 6. ROC for RF(balanced dataset)

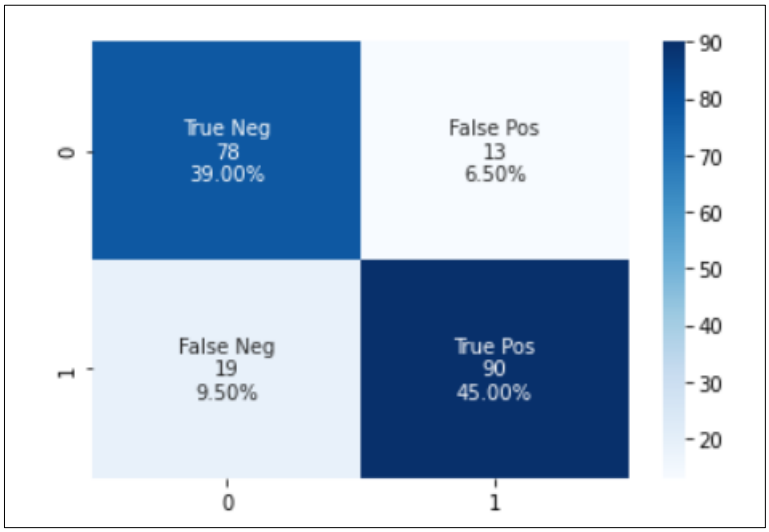


Figure 7. Confusion matrix for RF (balanced dataset)

4. COMPARISON WITH STATE OF THE ART STUDIES

In the common scientific research in order to generalize the obtained results it is essential to compare these results with outcomes of previous studies. Thus, findings of proposed study have compared with four baseline works. The comparison conducted based on dataset, number of features and obtained accuracy as detailed into Table 5.

Table 5.Comparison with baseline works

Study	Dataset	Classifier	No of features	Accuracy
[30]	PID	RF	8	75.8%
[12]	PID	NB	8	72.3%
[11]	PID	RF	**3**	75.2%
[11]	PID	RF	5	73.9%
[10]	PID	RF	8	75.32%
Our study	PID	**RF**	**6**	**81%**
Our study	PID	**RF**	**5**	**84%**

Table 5 revealed that all previous works have scored low detection performance for diabetes against non-diabetes cases. Where all studies not exceeded 75% of accuracy metric. Furthermore, the best accuracy reach to 75.8% as stated into study[30] .However, even with minimal number of features that use for detection into study[11] still the detection rate very low. To mention our proposed study have outperformed other studies with less number of diabetes indicators and high detection rate. For instance, based on 6 features have scored 81% as classification accuracy. In better scenario, 84% have scored based only 5 features which considered as best detection accuracy for diabetes over non diabetes cases with less computational load as well as effective classification performance.

The comparative analysis with baseline studies highlights the strength of the proposed framework in improving diabetes detection accuracy. Previous works using the Pima Indian Dataset with Random Forest or Naïve Bayes classifiers reported accuracy levels below 76%, even when employing the full set of eight features. In contrast, the proposed approach demonstrates superior performance by achieving 81% accuracy with only six features and further improving to 84% with five features. This indicates that the combination of effective feature selection and data balancing not only reduces computational complexity but also enhances the model's predictive capability. By focusing on the most informative indicators, the framework avoids noise from irrelevant attributes, resulting in a more efficient and accurate classification process. The results also underline the broader contribution of this study to the field of medical data analysis. Achieving higher accuracy with fewer features addresses two key challenges: the scarcity of balanced medical datasets and the need for computationally efficient models suitable for real-world healthcare applications. The strong performance of the Random Forest classifier in this study suggests that, when paired with appropriate preprocessing and sampling strategies, traditional machine learning methods can still outperform baseline benchmarks. This improvement demonstrates the potential of the framework to support early and reliable diabetes detection, offering practical value in clinical decision-making and preventive healthcare systems.

Although the proposed framework has shown promising results, several limitations should be acknowledged. First, the study is based solely on the Pima Indian Diabetes Dataset (PID), which is relatively small in size and limited to a specific population group. As such, the results may not fully generalize to other demographic groups or larger, more diverse datasets. Future research should therefore validate the framework on broader datasets to confirm its robustness.
Second, while the use of feature selection and oversampling techniques improved performance, these methods can sometimes introduce bias or artificial patterns not present in real-world data. In particular, oversampling approaches such

as ADASYN may inflate the minority class, potentially leading to optimistic accuracy estimates when compared to natural class distributions. Third, the study relies mainly on Random Forest and SVM classifiers. Although these models performed well, the exploration of additional algorithms, including deep learning methods, could offer further improvements in predictive power and adaptability. Finally, the evaluation metrics were focused on standard measures such as accuracy, precision, recall, and F-score; a more comprehensive assessment including sensitivity, specificity, and real-world cost–benefit analysis would provide deeper insights into clinical applicability.

5. Conclusion

Recently, diabetic disease has become world-wide disease and the rate rapidly increased day by day. There is a great effort by medical association to handle such disease. However, the computer aided system is very effective in this direction. Therefore, this research proposed detection framework for diabetes and non-diabetes cases. The proposed study included different stages such as preprocessing, features selection classification based on machine learning approach, and evaluation based on well-known evaluation metric. This study has proposed RF model as base for detection stage. Furthermore, IG and SMOTE methods were used for handle issue related to feature selection and unbalance data. The results showed that proposed study succeed in handling selecting of most effective features as well as misleading results that generated by unbalanced Pima Indian diabetes dataset. In the same direction, RF has strong performance in terms of detection for diabetes and non-diabetes case with low computational load. Also, comparing with pervious works this study outperformed other studies in terms of number selected features as well as accuracy rate of detection for diabetes disease. For further improvements for current study another dataset can be aggregated with Pima dataset in order to increase detection rate and implement this study into real-time application.

FUNDING INFORMATION

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

COFLICTS OF INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY STATEMENTS

The dataset used in this study is publicly available at the following link: https://www.kaggle.com/uciml/pima-indians-diabetes-database

REFERENCES

[1] K. G. M. M. Alberti, P. Zimmet, and J. J. D. M. Shaw, "International Diabetes Federation: a consensus on Type 2 diabetes prevention," vol. 24, no. 5, pp. 451-463, 2007.
[2] A. D. A. J. D. care, "Diagnosis and classification of diabetes mellitus," vol. 33, no. Supplement_1, pp. S62-S69, 2010.
[3] E. S. Almutairi and M. F. Abbod, "Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia," vol. 4, no. 1, pp. 37-55, 2023.
[4] S. Malik, S. Harous, and H. El-Sayed, "Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women," in *International Symposium on Modelling and Implementation of Complex Systems*, 2020, pp. 95-106: Springer.
[5] E. Pekel Özmen and T. J. J. o. F. Özcan, "Diagnosis of diabetes mellitus using artificial neural network and classification and regression tree optimized with genetic algorithm," vol. 39, no. 4, pp. 661-670, 2020.
[6] "WOA-COVID-19: Whale Optimization Algorithm for Selection of Multi-Examination Features based on COVID-19 Infections," *Mesopotamian Journal of Computer Science,* vol. 2025, pp. 172–185, %08/%06 2025.
[7] J.-W. Mao, Y. He, and Z.-T. Liu, "Speech emotion recognition based on linear discriminant analysis and support vector machine decision tree," in *2018 37th Chinese control conference (CCC)*, 2018, pp. 5529-5533: IEEE.
[8] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," vol. 2021, no. 1, p. 9930985, 2021.
[9] S. S. Bhat, M. Banu, G. A. Ansari, and V. Selvam, "A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms," *Healthcare Analytics,* vol. 4, p. 100273, 2023/12/01/ 2023.
[10] A. Mahabub, "A robust voting approach for diabetes prediction using traditional machine learning techniques," *SN Applied Sciences,* vol. 1, no. 12, p. 1667, 2019/11/25 2019.
[11] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Computing and Applications,* vol. 35, no. 22, pp. 16157-16173, 2023/08/01 2023.
[12] Y. Guo, G. Bai, and Y. Hu, "Using bayes network for prediction of type-2 diabetes," in *2012 International conference for internet technology and secured transactions*, 2012, pp. 471-472: IEEE.
[13] A. Iyer, S. Jeyalatha, and R. J. a. p. a. Sumbaly, "Diagnosis of diabetes using classification mining techniques," 2015.
[14] V. A. Kumari, R. J. I. J. o. E. R. Chitra, and Applications, "Classification of diabetes disease using support vector machine," vol. 3, no. 2, pp. 1797-1801, 2013.
[15] A. Ahmed *et al.*, "Machine learning algorithm-based prediction of diabetes among female population using pima dataset," in *Healthcare*, 2024, vol. 13, no. 1, p. 37: MDPI.
[16] Pima Indians Diabetes Database [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[17] M. W. Nadeem, H. G. Goh, V. Ponnusamy, I. Andonovic, M. A. Khan, and M. Hussain, "A Fusion-Based Machine Learning Approach for the Prediction of the Onset of Diabetes," *Healthcare*, vol. 9, no. 10. doi: 10.3390/healthcare9101393

[18] Z. Gao, Y. Xu, F. Meng, F. Qi, and Z. Lin, "Improved information gain-based feature selection for text categorization," in *2014 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE)*, 2014, pp. 1-5: IEEE.

[19] C. E. J. T. B. s. t. j. Shannon, "A mathematical theory of communication," vol. 27, no. 3, pp. 379-423, 1948.

[20] H. Zhou, Z. Deng, Y. Xia, and M. Fu, "A new sampling method in particle filter based on Pearson correlation coefficient," *Neurocomputing,* vol. 216, pp. 208-215, 2016/12/05/ 2016.

[21] I. Emmanuel, Y. Sun, and Z. Wang, "A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method," *Journal of Big Data,* vol. 11, no. 1, p. 23, 2024/02/01 2024.

[22] H. T. Abbas *et al.*, "Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test," *PLOS ONE,* vol. 14, no. 12, p. e0219636, 2019.

[23] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Applied Computing and Informatics,* vol. 18, no. 1-2, pp. 90-100, 2020.

[24] N. Nipa, M. H. Riyad, S. Satu, Walliullah, K. C. Howlader, and M. A. Moni, "Clinically adaptable machine learning model to identify early appreciable features of diabetes," *Intelligent Medicine,* vol. 4, no. 1, pp. 22-32, 2024/02/01/ 2024.

[25] L. J. M. l. Breiman, "Random forests," vol. 45, no. 1, pp. 5-32, 2001.

[26] Z. N. J. J. o. A.-Q. f. C. S. Nemer and Mathematics, "Oil and gas production forecasting using decision trees, random forst, and XGBoost," vol. 16, no. 1, pp. 9–20-9–20, 2024.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. J. J. o. a. i. r. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," vol. 16, pp. 321-357, 2002.

[28] Y. J. R. J. o. A. S. E. Mi and Technology, "Imbalanced classification based on active learning SMOTE," vol. 5, pp. 944-949, 2013.

[29] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Scientific Reports,* vol. 11, no. 1, p. 24039, 2021/12/15 2021.

[30] "Performance Analysis of Diabetes Detection Using Machine Learning Classifiers," *International Journal of Management and Data Analytics (IJMADA),* vol. 4, no. 1, pp. 43-54, 10/13 2024.