Journal of Current Science and Technology, July - September 2025 Vol. 15 No. 3, Article 111

ISSN 2630-0656 (Online)

Cite this article: Poorani, K., Balakannan, S. P., & Karuppasamy, M. (2025). Mitigating data imbalance for robust diabetes diagnosis using machine learning and explainable artificial intelligence. *Journal of Current Science and Technology*, *15*(3), Article 111. https://doi.org/10.59796/jcst.V15N3.2025.111

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Mitigating Data Imbalance for Robust Diabetes Diagnosis Using Machine Learning and Explainable Artificial Intelligence

Poorani K[1], Balakannan SP[2,*], and Karuppasamy M[3]

[1]Department of Computer Applications, School of Computing, Kalasalingam Academy of Research and Education, Krishnankoil, Tamil Nadu, India.
[2]Department of Information Technology, School of Computing, Kalasalingam Academy of Research and Education, Krishnankoil, Tamil Nadu, India.
[3]Department of Computer Science and Engineering, RajaRajeswari College of Engineering, Bengaluru, Karnataka, India.

*Corresponding author; E-mail: balakannansp@gmail.com

Received 18 October 2024; Revised 21 December 2024; Accepted 13 January 2025; Published online 15 June 2025

Abstract

Diabetes is increasing at a global level and is associated with a high mortality rate. Early diagnosis can significantly reduce the risk of complications and save lives. This study proposes an efficient ensemble model for diabetes diagnosis using Machine Learning (ML). To address class imbalance in the dataset, a hybrid sampling technique Synthetic Minority Over-sampling Technique (SMOTE) combined with Edited Nearest Neighbors (ENN) is implemented. This combined method, referred to as SMOTE-ENN, enhances the model's ability to accurately predict diabetes outcomes by generating synthetic samples and removing noisy instances. Before implementing any ML model, it is necessary to prepare the data and find a suitable model. Data preprocessing techniques like normalization, and filling missing values are essential in preparing data for a ML model. The proposed approach implements suitable preprocessing techniques such as mean value imputation and encoding. Feature selection with mutual information is carried out to select important variables. The PIMA Indian Diabetic dataset is balanced using SMOTE-ENN, a sampling strategy in Python with the help of the imbalanced-learn library, which improves model performance. The dataset is split for analysis at 80:20 ratio (train: test). Before ML implementation, the data is prepared for model building. Then, various ML models are introduced, ranging from single classifiers to ensemble models. The proposed approach, SMOTE-ENN with ensemble ML models proves that stacking provides high accuracy (98.9%), precision (97.6%), recall (99.5%), and F1-score (98%). Explainable Artificial Intelligence is also used to interpret the results with the help of Local Interpretable Model-Agnostic Explanations (LIME). The proposed approach combines feature selection, data imbalance handling, and ensemble techniques to improve performance. Stacking with the proposed approach performs better than state-of-the-art algorithms.

Keywords**:** *diabetes prediction; explainable ai; machine learning; mutual information; random forest; SMOTE-ENN; stacking*

1. Introduction

Diabetes leads to various vascular complications, many of which are unnoticeable. It is a long-term disorder that affects different parts of the body, leading to serious complications and a high mortality rate. People with diabetes tend to have several diseases that become worse

with comorbidities like uncontrolled blood pressure, cholesterol, and smoking (World Health Organization, 2024).

International Diabetes Federation (IDF) Report 2022 on Type 1 diabetes reveals that 62% of all cases are in individuals aged 20 and above. The IDF Atlas (International Diabetes Federation, n.d.) shows that diabetes has affected 537 million adults worldwide. Every five seconds, one person dies from diabetes. Over three in four people with diabetes live in low- and middle-income countries. Diabetes is also a significant risk factor during the COVID-19 pandemic. Statistical reports indicate that many diabetes-related deaths occurred in individuals under the age of 60. The prevalence and mortality rates vary across regions. These reports highlight the importance of diabetic diagnosis at earlier stages.

Figure 1 indicates the number of deaths in various IDF regions among individuals under the age of 60, with the Western Pacific region being the most affected. This highlights that age is an important feature to consider. Diabetes progression may be due to genetic, environmental, or lifestyle factors (Suryasa et al., 2021). The rate of increase in diabetes cases is projected to be much higher in 2045 (Sun et al., 2022). The prevalence of undiagnosed diabetes is increasing due to a lack of public awareness. Currently, machine learning (ML) techniques are highly effective in diagnosing various diseases.

Predicting diabetes progression at an early stage could improve patients' quality of life and protect them from chronic conditions. Medical data is typically high-dimensional, containing various features, missing values, and class imbalance. Such data can be effectively handled using machine learning models.

Machine learning models are increasingly being applied in the healthcare sector to enhance disease diagnosis and reduce mortality (Pechprasarn et al., 2024; Pechprasarn et al., 2023). Selecting relevant features from the dataset is crucial for minimizing model complexity and improving predictive accuracy. In this study, mutual information is employed for feature selection due to its ability to quantify the dependency between input variables and the target outcome. To address class imbalance, a common challenge in medical datasets, this work utilizes SMOTE-ENN, a hybrid sampling technique. SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic samples for the minority class, while ENN (Edited Nearest Neighbors) removes noisy or ambiguous examples from the majority class. This combination results in a more balanced and cleaner dataset, which contributes to improved model generalization. For classification, stacking is employed as an ensemble learning method. Stacking combines multiple base learners with a meta-learner to build a more robust predictive model. This approach reduces both bias and variance more effectively than individual classifiers. The meta-learner is trained on the outputs of the base models, allowing it to learn from their strengths and compensate for their errors. Stacking is particularly beneficial for handling complex, non-linear data patterns, making it well-suited for healthcare applications such as diabetes prediction.

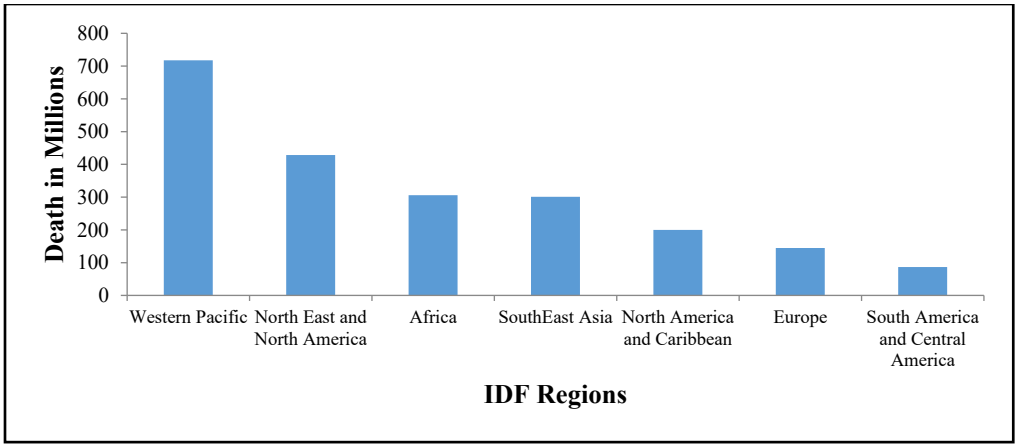

Figure 1 Statistical Report on Diabetes Death Before age 60

1.1 Literature Review

Various studies have been conducted in the domain of disease prediction and classification. Chou et al., (2023) employed several machine learning algorithms including Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Decision Tree (DT), Logistic Regression (LR), and Random Forest (RF) for diabetes prediction and found that Random Forest outperformed the other models in terms of predictive performance. Jaiswal et al., (2021) showed that ML models help to diagnose diseases. Models such as SVM, ANN, and deep learning techniques are eliminating the barrier to disease prediction. Tuppad, & Patil (2022) highlighted the uses of ML for clinical decision support in diabetes. Khanam, & Foo (2021) compared ML algorithms such as LR, SVM, and NN for diabetic classification and found that NN achieved 88.6% accuracy.

Larabi-Marie-Sainte et al., (2019) in their review stated that ML and Deep Learning (DL) algorithms are used for classification and their accuracy ranges from 64-78%. Kishor, & Chakraboty's (2024) studies reveal that implementing data preprocessing along with data balancing helps build highly accurate machine learning models. Xu et al., (2020) implemented SMOTE and ENN for class imbalance in medical datasets which helped improving performance. Nnamoko et al., (2020) and Tasneem et al., (2024) highlighted the importance of outlier detection interquartile range and SMOTE. The use of SMOTE-ENN for data balancing improved efficiency of the model (Yang et al., 2022). Feature selection with mutual information improves the machine learning model, resulting in efficient outcomes (Tasin et al., 2023).

Ganie, & Malik (2022) used bagging, boosting, and voting techniques to predict diabetes, with bagging using a decision tree achieving 97% accuracy. Hasan et al., (2020) relied on ensemble learning techniques and employed the Area Under the Curve (AUC) as a performance metric, optimizing model performance through hyperparameter tuning using the grid search method. Abdollahi, & Nouri-Moghaddam (2022) showed that ensemble learning using a genetic algorithm is more effective for diabetes classification, with bagging and boosting classifiers achieving 98% accuracy.

Kalagotla et al., (2021) implemented stacking techniques for classification using Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Logistic Regression (LR) as base learners. They also applied the Adaptive Boosting (AdaBoost) technique to the selected features in order to enhance classification performance. Singh, & Singh (2020) developed a learning system based on Non-dominated Sorting Genetic Algorithm II (NSGA-II) stacking, incorporating multi-objective optimization to enhance the performance of base learners for diabetes prediction. Selection of meta-classifiers in stacking is essential. Studies have used Random Forest as a classifier, which improves the efficiency of stacking models (Ali et al., 2022).

2. Objectives

1. To propose an efficient ensemble model for diabetes diagnosis using machine learning technique.
2. To address class imbalance in the dataset by implementing a hybrid sampling method combining SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbors).
3. To develop and evaluate a stacking ensemble model for diabetes prediction, and to compare its performance against existing classification methods.

3. Methodology

This work involves data preprocessing with normalization, data imbalance handling, train-test splitting, Machine learning (ML) model training, and classification of diabetes. The main contribution of this work is data imbalance handling using SMOTE-ENN which helps in preparing balanced data for machine learning model implementation. Several researchers have utilized other data imbalance techniques such as SMOTE, SMOTETomek, class weighting, and synthetic data generation; however, SMOTE-ENN is employed in this work. Performing feature selection using mutual information before data imbalance handling reduces model cost and complexity. Incorporating the stacking ensemble method for diabetes prediction improves model accuracy to a notable extent. While single classifiers perform well, improving accuracy in disease prediction remains a challenge, which is addressed in the proposed approach. Explainable AI significantly aids in the interpretation of results.

The key contributions of this work are categorized as

i. Data Preprocessing
ii. Machine Learning models
iii. Ensembling methods
iv. Explainable AI interpretation
v. Performance Evaluation.

The PIMA Indian diabetes dataset has been used for prediction. It contains eight attributes relevant to diabetes risk, with the final attribute representing the outcome. This work used machine learning techniques to classify diabetes by ensembling concepts stacking with feature importance and data balancing through SMOTE and ENN. The proposed efficient ensemble approach demonstrates that the combination of SMOTE-ENN and mutual information provides higher accuracy compared to other models.

The S-E-MI approach identifies the most relevant features for classification, which are later combined with ensemble stacking to achieve high classification performance. The proposed Ensemble Stacking with SMOTE-ENN-Mutual Information (S-E-M) is described in detail, and its performance is compared with existing methods. The workflow of the proposed S-E-M model for efficient diabetes prediction is illustrated in Figure 2.

Feature analysis is performed to analyze the major features contributing to the model. Those features are essential for assessing the risk of diabetes complications. Various validation metrics are used for analysis. The process involves class balancing, 10-fold cross-validation, train-test splitting, and stacking of DT, NB, SVM, and kNN.

3.1 Data Collection

The PIMA Diabetic dataset is taken from the UCI Machine Learning Repository for analysis. It contains data from PIMA Indian individuals diagnosed with diabetes. Eight different attributes are involved in the study. The outcome variable indicates whether or not a person has diabetes.

3.2 Data Preprocessing

3.2.1 Handling Missing Values

Missing or null values can affect model efficiency by causing deviations in the outcome. Therefore, handling missing values through removal or imputation is necessary for further processing. In this study, mean imputation is used to fill missing values.

Table 1 presents the mean values of the PIMA dataset. It highlights the most frequently occurring features in the dataset. Preprocessing refers to the process of preparing data ML classification. This work involves preprocessing, followed by data transformation and classification.

3.2.2 Normalization

Normalization is one of the data preprocessing concepts that involves handling data where the feature does not contain extreme outliers. Z-scores are typically used when there are only a few outliers.

$$x' = (x - \mu)/\sigma \qquad \text{(eq. 1)}$$

Normalization is applied because features such as pregnancies, insulin, diabetes pedigree function, and age are right-skewed. Other variables have varied distributions.

Table 1 Mean Values of PIMA Dataset

Attribute	Mean Value
1	3.845
2	120.859
3	69.105
4	20.536
5	79.799
6	31.993
7	0.427
8	33.241

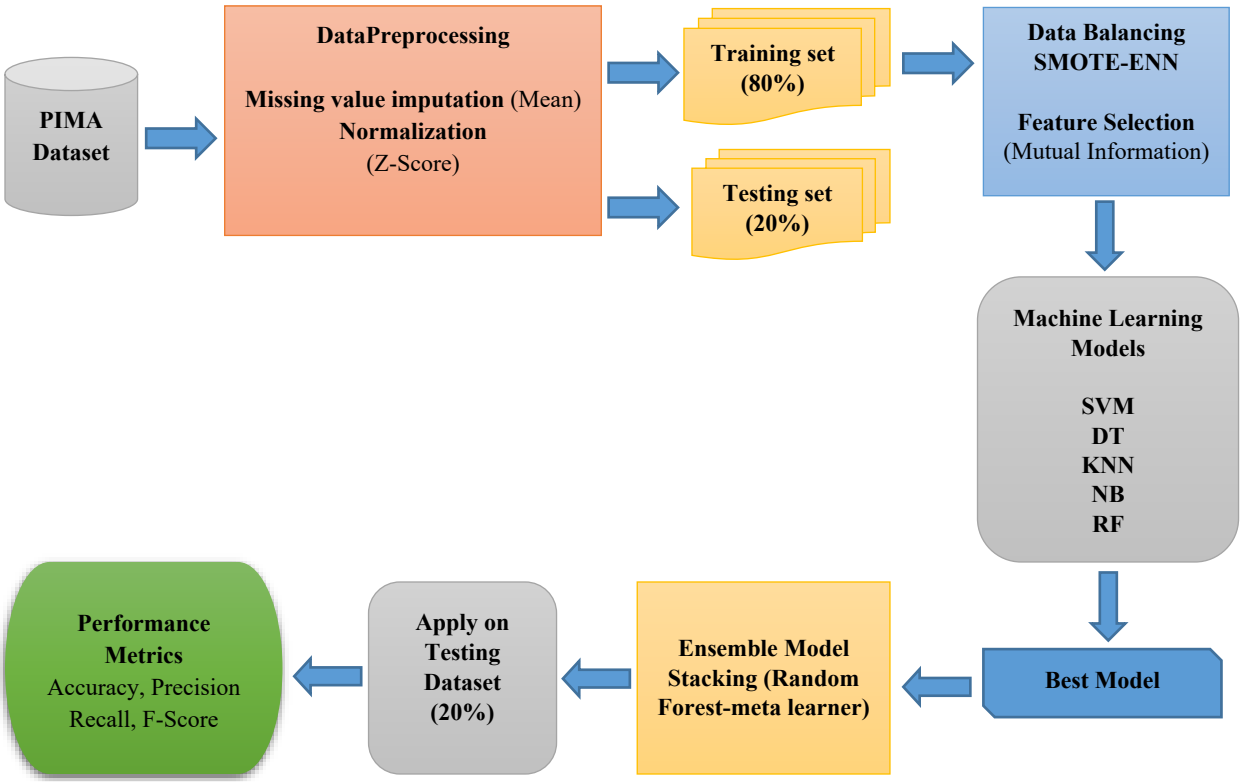

Figure 2 Efficient S-E-M Model for Diabetes Prediction

3.3 Data Balancing

Before addressing data imbalance, the distribution of the dataset is analyzed. The eight variables were widely distributed from low to high, where the outcome variable is highly imbalanced. To build highly accurate ML models, it is necessary to balance the data before testing. Data distribution has been thoroughly examined, and data balancing is recommended. The least commonly known ENN, commonly used in combination with SMOTE, is applied for more effective data balancing. The proposed model implements a hybrid SMOTE-ENN model for handling an imbalance dataset. Both oversampling (SMOTE) and under sampling (ENN) are combined to provide a hybrid approach. This model generates synthetic data for the minority class using SMOTE, resulting in a balanced dataset. With balanced data, samples from the majority class adjacent to the minority class are rejected by under-sampling ENN. This concept reduces both overfitting and noise.

Algorithm for combining SMOTE and ENN

Step 1: Start
Step 2: Initiate Processing Dataset
Step 3: Random selection of X_j
Step 4: Verify kNN
Step 5: Generate Xnew=Xi+|X-Xj| *δ
Step 6: Check whether the dataset is balanced.
Step 7: If yes, move to step 8 or move to step 3
Step 8: Perform noise removal with ENN (undersampling)
Step 9: Stop

3.4 Feature Selection

Feature selection improves model performance by eliminating non-informative features that may interfere with the target variable. There are several feature selection techniques, mutual information is being used here.

3.4.1 Mutual Information

MI measures the dependencies between variables based on their correlations. Here, each feature's score is

calculated based on its dependency with the target variable. It is a score-based system. A higher score indicates stronger dependency, while a lower score reflects weaker relevance and may be excluded from further analysis.

$$I(X;Y)=H(X)-H(X|Y) \quad \text{(eq. 2)}$$

Table 2 Mutual Information Score

Variables	Mutual Information Score
Plasma Glucose	0.15
BMI	0.12
Age	0.10
No. of Pregnancies	0.09
2-hr serum insulin	0.08
Diabetes pedigree function	0.07
Diastolic BP	0.05
Triceps	0.04

Table 2 represents the mutual information obtained for selecting major features for model building. The MI score of 0.05 or less is considered the least important and is excluded from further processing.

3.5 Data Classification

This section presents the classification process using Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), and k-Nearest Neighbors (kNN), followed by the implementation of stacking to combine these models.

3.5.1 Support Vector Machine (SVM)

Support Vector Machine is used to reveal the optimal hyperplane in the n-dimensional plane. A hyperplane is used to find various classes in the model. When the threshold value is near 0, then the classes are negative, when it is nearer to one, then the classes are positive.

$$y=wX+b \quad \text{(eq. 3)}$$

3.5.2 Decision Tree (DT)

Decision Tree performs both classification and regression. It forms a tree, with the highest node as the root and features as internal nodes. Each internal node is a test-over feature, and the branch indicates the outcome of the internal node. This uses entropy for feature selection over the nodes.

$$\text{Entropy (p)} = -\sum_{i=1}^{N} pi \log2\ pi \quad \text{(eq. 4)}$$

3.5.3 Naïve Bayes (NB)

Naïve Bayes is a probabilistic classifier. It uses supervised learning methods for classification. It uses probability to give results. Probability is between 0 and 1, if the result is 0 then the case is negative, if it is 1, the case is positive. The method of probability is stated below.

$$P(X|Y) = P(X \text{ and } Y) / P(X) \quad \text{(eq. 5)}$$

3.5.4 k Nearest Neighbour (kNN)

The k-Nearest Neighbors algorithm is widely used for both classification and regression tasks. It classifies a given query point based on the majority class among its nearest neighbors in the feature space. The proximity between data points is typically calculated using the Euclidean distance metric, which helps determine the similarity between instances.

Euclidean Distance

Euclidean distance is a commonly used metric to measure the straight-line distance between two points in an N-dimensional space. It is particularly effective when the dataset is balanced, as imbalanced data can distort distance measurements and negatively impact model performance. The equation to solve the distance is given below:

$$d=\sqrt{\sum_{i=1}^{N} (Xi-Yi)^2} \quad \text{(eq. 6)}$$

3.5.5 Random Forest (RF)

The Random Forest (RF) algorithm constructs an ensemble of decision trees during the training process. Each tree independently contributes to the final prediction, and the overall result is determined by aggregating the outputs typically through majority voting for classification tasks. This ensemble approach enhances accuracy and reduces the risk of overfitting compared to individual decision trees (Mohamad Suffian et al., 2024). The quality of each split within the trees is commonly evaluated using the Gini Index, as defined by the following equation:

$$\text{Gini Index}=1-\sum_{i=1}^{n} (Pi)2 \quad \text{(eq. 7)}$$

3.6 Stacking

Stacking is an ensemble technique used to improve the model's performance. It usually combines various base learners and has a meta-learner (Ali et al., 2022). In the proposed system, SVM, DT, kNN, and NB are used as base learners, while Random Forest (RF) serves as the meta-learner. The results of base classes are combined and given as input to the Meta-classifier, Random Forest.

4. Results

Feature importance was evaluated using several supervised machine learning algorithms while Random Forest gave much more efficient features of importance relevant to outcome.

Figure 3 represents the feature importance of the diabetic dataset with random forest, which highlights glucose, BMI, diabetes pedigree, and age as the most significant features, along with their respective importance scores. The feature importance highlights that glucose is the most vital feature for the progression of diabetes. Secondly, BMI (Body Mass Index), which calculates the height and weight of a person determines the diabetic condition. Diabetes pedigree and age are the next most important features in diabetes prediction. The features are ranked accordingly to provide the important score level. The results demonstrate that implementing certain feature selection and data preprocessing techniques helps in improving model performance.

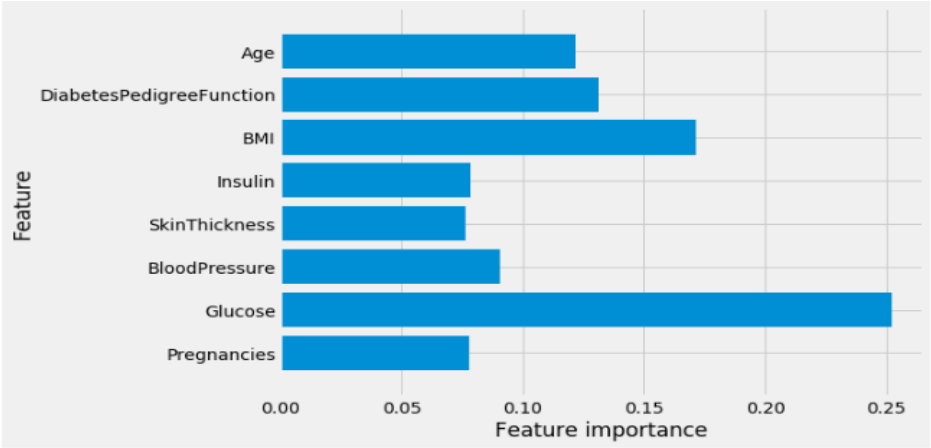

Figure 3 Feature Importance using Random Forest

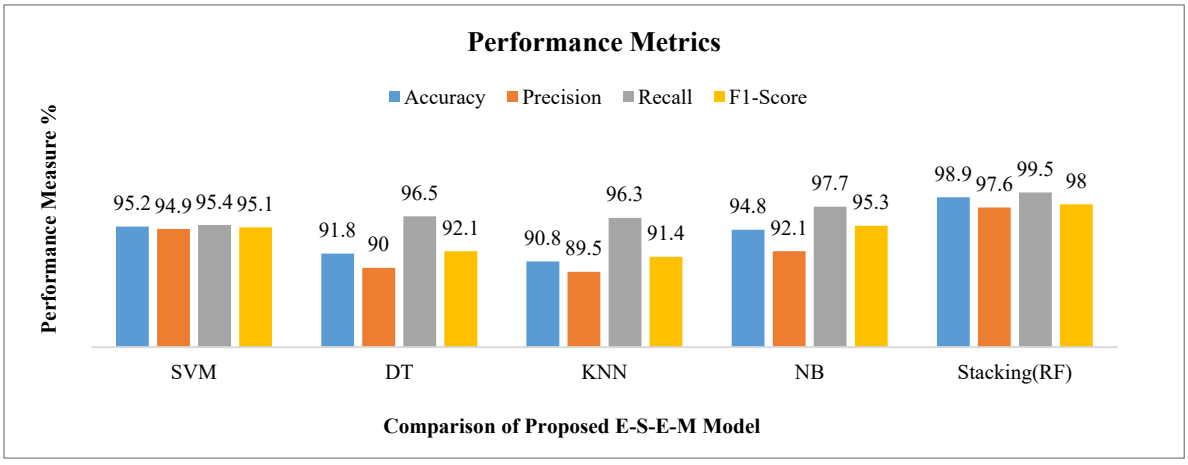

Figure 4 Performance Measures of E-SEM model

Figure 4 shows that SVM, RF, LR, and KNN perform well, and the implementation of stacking with models using RF as a metaclassifier improves all the performance metrics. Implementing stacking exactly provides better results compared to all individual ML models. As a result, stacking achieves an accuracy of 98.9%, precision of 97.6%, recall of 99.5% and F1-score of 98%. The improved accuracy and recall across all models represent the major strength of the proposed ESEM model.

Table 3 presents a comparative analysis between the proposed model and existing approaches in terms of classification accuracy. The results show that the proposed E-S-E-M model outperforms prior methods. The integration of stacking significantly improves model efficiency, particularly in terms of accuracy and precision.

5. Explainable Artificial Intelligence

Explainable AI using Local Interpretable Model-Agnostic Explanations (LIME) is employed to understand how the model makes predictions. Figure 6 demonstrates the feature of importance score generated by LIME. From Figure 5, it is understood that the conditions and probabilities for diabetes classifications are well defined. When glucose levels exceed 100, age is above 40, and BMI is greater than 32, the probability of developing diabetes is higher. Explainable artificial intelligence proves highly useful in interpreting the model's results.

6. Conclusion

The proposed study concludes that machine learning is highly effective for diabetes classification. Handling data imbalance using SMOTE-ENN significantly enhances model performance. In addition, feature selection using mutual information and the incorporation of stacking are major contributions that improve classification accuracy. The proposed S-E-M model achieves higher accuracy compared to existing approaches. The key takeaway from this work is that proper data preprocessing and effective handling of data imbalance are essential for building efficient machine learning models.

The primary limitation of the proposed work is the high computational cost of stacking for high-dimensional datasets and the associated risk of overfitting. However, the PIMA diabetes dataset is relatively simple and does not pose significant complexity. Performing feature selection prior to data imbalance handling helps reduce both the computational cost and model complexity. The areas of future research include stacking with neural network for other diseases such as heart disease, stroke, and chronic kidney diseases.

Table 3 Comparative Analysis of Existing Works in terms of Accuracy

Authors & Year	Feature Selection Techniques	Model Used	Accuracy (%)
Kalagotla et al., (2021)	Correlation-Based	Adaboost for Feature Selection+ MLP, SVM, and LR	80.2
Singh, & Singh (2020)	Multi-objective Optimization	NSGA-II (Meta classifier -KNN)	83.8
Abdollahi, & Nouri-Moghaddam (2022)	Genetic Algorithm	ST-GA	98
Tasin et al., (2022)	SMOTE+ADASYN	XGBoost+ADASYN	81
Saihood, & Sonuç (2023)	-	ML+RF	95.5
Proposed E-S-E-M Model	Mutual Information	SMOTE-ENN +MI+Stacking	98.9

Figure 5 Probabilities & Feature Value by LIME

Funding

This work does not receive any funding.

Conflict of Interest

The authors declare no conflict of interest.

7. References

Abdollahi, J., & Nouri-Moghaddam, B. (2022). Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. *Iran Journal of Computer Science*, *5*(3), 205-220. https://doi.org/10.1007/s42044-022-00100-1

Ali, M., Haider, M. N., Lashari, S. A., Sharif, W., Khan, A., & Ramli, D. A. (2022). Stacking classifier with random forest functioning as a meta classifier for diabetes diseases classification. *Procedia Computer Science*, *207*, 3459-3468. https://doi.org/10.1016/j.procs.2022.09.404

Chou, C. Y., Hsu, D. Y., & Chou, C. H. (2023). Predicting the onset of diabetes with machine learning methods. *Journal of Personalized Medicine*, *13*(3), Article 406. https://doi.org/10.3390/jpm13030406

Ganie, S. M., & Malik, M. B. (2022). An ensemble machine learning approach for predicting type-II diabetes mellitus based on lifestyle indicators. *Healthcare Analytics*, *2*, Article 100092. https://doi.org/10.1016/j.health.2022.100092

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516-76531. https://doi.org/10.1109/ACCESS.2020.2989857

International Diabetes Federation. (n.d.). *Diabetes around the world in 2021*. IDF Diabetes Atlas. Retrieved from https://diabetesatlas.org/

Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. *Primary Care Diabetes*, *15*(3), 435-443. https://doi.org/10.1016/j.pcd.2021.02.005

Kalagotla, S. K., Gangashetty, S. V., & Giridhar, K. (2021). A novel stacking technique for prediction of diabetes. *Computers in Biology and Medicine*, *135*, Article 104554. https://doi.org/10.1016/j.compbiomed.2021.104554

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), 432- 439. https://doi.org/10.1016/j.icte.2021.02.004

Kishor, A., & Chakraborty, C. (2024). Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE. *International Journal of System Assurance Engineering and Management*, *15*(10), 4649-4657. https://doi.org/10.1007/s13198-021-01174-z

Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini, R., & Saba, T. (2019). Current techniques for diabetes prediction: review and case study. *Applied Sciences*, *9*(21), Article 4604. https://doi.org/10.3390/app9214604

Mohamad Suffian, M. S. Z. B., Kamil, S., Ariffin, A. K., Azman, A. H., Ibrahim, I. M., & Suga, K. (2024). A surrogate model's decision tree method evaluation for uncertainty quantification on a finite element structure via a fuzzy-random approach. *Journal of Current Science and*

Technology, *14*(3), Article 50. https://doi.org/10.59796/jcst.V14N3.2024.50

Nnamoko, N., & Korkontzelos, I. (2020). Efficient treatment of outliers and class imbalance for diabetes prediction. *Artificial Intelligence in Medicine*, *104*, Article 101815. https://doi.org/10.1016/j.artmed.2020.101815

Pechprasarn, S., Suechoey, N., Pholtrakoolwong, N., Tanedvorapinyo, P., & Toboonliang, Y. (2024). Optimizing lung cancer diagnosis with machine learning and feature selection methods. *Journal of Current Science and Technology*, *14*(3), Article 55. https://doi.org/10.59796/jcst.V14N3.2024.55

Pechprasarn, S., Wattanapermpool, O., Warunlawan, M., Homsud, P., & Akarajarasroj, T. (2023). Identification of important factors in the diagnosis of breast cancer cells using machine learning models and principal component analysis. *Journal of Current Science and Technology, 13*(3), 642–656. https://doi.org/10.59796/jcst.V13N3.2023.700

Saihood, Q., & Sonuç, E. (2023). A practical framework for early detection of diabetes using ensemble machine learning models. *Turkish Journal of Electrical Engineering and Computer Sciences, 31*(4), 722–738. https://doi.org/10.55730/1300-0632.4013

Singh, N., & Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybernetics and Biomedical Engineering*, *40*(1), 1-22. https://doi.org/10.1016/j.bbe.2019.10.001

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., ... & Magliano, D. J. (2022). IDF Diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, *183*, Article 109119. https://doi.org/10.1016/j.diabres.2021.109119

Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2021). Health and treatment of diabetes mellitus. *International Journal of Health Sciences*, *5*(1), 1-5. https://doi.org/10.53730/ijhs.v5n1.2864

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1-2), 1-10. https://doi.org/10.1049/htl2.12039

Tasneem, S., Younas, M., & Shafiq, Q. (2024). Identifying key learning algorithm parameter of forward feature selection to integrate with ensemble learning for customer churn prediction. *VFAST Transactions on Software Engineering*, *12*(2), 56-75. https://doi.org/10.21015/vtse.v12i2.1811

Tuppad, A., & Patil, S. D. (2022). Machine learning for diabetes clinical decision support: a review. *Advances in Computational Intelligence*, *2*(2), Article 22. https://doi.org/10.1007/s43674-022-00034-y

World Health Organization. (2024). *Diabetes*. Retrieved form https://www.who.int/news-room/fact-sheets/detail/diabetes

Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *Journal of Biomedical Informatics*, *107*, Article 103465. https://doi.org/10.1016/j.jbi.2020.103465

Yang, F., Wang, K., Sun, L., Zhai, M., Song, J., & Wang, H. (2022). A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis. *BMC Medical Informatics and Decision Making*, *22*(1), Article 344. https://doi.org/10.1186/s12911-022-02075-2