INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Data Augmentation Approach for Type-2 Diabetes Prediction and Classification

B. Patel Hitesh[1], Brahmbhatt Keyur[2]*, Joshi Mahasweta[3]

1 Research Scholar, Gujarat Technological University, Ahmedabad-382424, Gujarat, India
2 Professor and Head, Information Technology Department, BVM Engineering College,V V Nagar-388120, Gujarat, India
3 Assistant professor, Computer Engineering Department, BVM Engineering College,V V Nagar-388120, Gujarat, India

OPEN ACCESS

Received: 30/12/2024

Accepted: 06/02/2025

Published: 26/04/2025

Citation: Hitesh BP, Keyur B, Mahasweta J (2025) Data Augmentation Approach for Type-2 Diabetes Prediction and Classification. Indian Journal of Science and Technology 18(14): 1096-1104. https://doi.org/10.17485/IJST/v18i14.3904

* **Corresponding author**.

keyur.brahmbhatt@bvmengineering.ac.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The main objective of this study is to develop a generalized machine learning framework for Type 2 diabetes prediction that integrates robust preprocessing, feature analysis, and advanced data augmentation techniques to enhance prediction accuracy. Methods: The proposed model of Type-2 Diabetes Prediction and Classification uses publicly available PIMA Diabetes Dataset includes missing value analysis, outliers detection, standardization as a pre-processing task, and a diabetic and non-diabetic class-wise data augmentation module using Adversarial Variational Auto-Encoder (AVAE) with Random Forest Classifier. The experimental results are evaluated and compared with Logistic Regression, Decision Tree, AdaBoost, SVC, Random Forest, GradientBoosting, and KNeighbors with cross-validation scores using mean accuracy. Findings: The experimental findings using RandomForestClassifier suggest a Type 2 diabetes prediction based on class-wise data augmentation has an accuracy level of 93.09%. Novelty: For medical dataset, class-wise data augmentation using Adversarial Variational Autoencoder plays crucial role while developing an accurate model.

Keywords: Type 2 Diabetes; Data Augmentation; AVAE; Random Forest Classifier; Cross Validation Scores

1 Introduction

Diabetes, particularly Type II diabetes, poses a significant global health challenge due to its chronic nature and prevalence. The rising global prevalence, projected to affect 592 million people by 2035, underscores the urgency of effective management and prevention strategies [(1)]. Effective early prediction and intervention are critical in mitigating the disease's progression and associated complications. Advances in machine learning have been pivotal in predicting diabetes onset and complications, enabling tailored preventive measures and better patient outcomes [(2)], [(3)].

Reza MS *et al.* [(4)] explores the use of Support Vector Machines (SVM) for Type II diabetes prediction, introducing a novel integrated kernel framework that combines Radial Basis Function (RBF) and RBF City Block kernels. Leveraging the PIMA

dataset, the proposed model addresses data challenges such as class imbalance, missing values, and outliers. The results demonstrate an impressive accuracy score of 85.5%, showcasing the model's superior performance compared to existing methods. These findings highlight the potential of the proposed approach in aiding clinical decision-making for early diabetes detection.

Kibria HB *et al.*[5] introduces an ensemble approach for diabetes prediction using a soft voting classifier, integrating six machine learning algorithms, including Random Forest and XGBoost, along with explainable AI tools like SHAP. Employing the PIMA Indian Diabetes dataset, the model applies advanced preprocessing techniques such as SMOTETomek for class balancing and median imputation for missing values. Achieving a balanced accuracy of 90% and an F1 score of 89%, the proposed model not only outperforms existing methods but also enhances interpretability for clinical application, aiding in early-stage diabetes detection.

Sampath P *et al.*[6] presents a robust framework for diabetes prediction by integrating preprocessing techniques like SMOTE for class balancing, outlier rejection, and feature selection with ensemble machine learning models. Using the Pima Indian Diabetes dataset, the study applies an ensemble of AdaBoost and XGBoost, optimized via grid search. The results demonstrate exceptional performance, achieving an accuracy of 90.4% and an AUC of 0.968, surpassing current state-of-the-art methods. These findings emphasize the model's potential in enhancing diabetes diagnosis and healthcare outcomes.

Albadri RF *et at.*[7] introduces a novel hybrid machine learning approach that integrates Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) classifiers to enhance prediction accuracy. By leveraging the Pima Indian Diabetes dataset and incorporating features like glucose levels, BMI, age, and insulin, the model achieves impressive results. By rigorous testing with the holdout method and k-fold cross-validation yields accuracies of 88.5% and 90.1%, respectively. These findings highlight the model's robustness and potential for advancing clinical applications in diabetes risk assessment.

The studyAl-Dabbasa L *et al.*[8] introduces a novel framework for early detection of female Type-2 diabetes using advanced oversampling techniques, specifically SVM-SMOTE, combined with machine learning classifiers like XGBoost, Random Forest, and SVM. Experiments conducted on the Pima Indian Diabetes dataset revealed that the framework achieved 91% accuracy with XGBoost and SVM-SMOTE, outperforming prior methodologies by approximately 6.4%. These results underscore the potential of machine learning and data preprocessing techniques in improving diabetes prediction and early intervention strategies.

Ganie SM *et al.*[9] leverages ensemble learning with five boosting algorithms—XGBoost, CatBoost, AdaBoost, LightGBM and Gradient boosting—for diabetes prediction is using the Pima Indian Diabetes dataset. Preprocessing steps like normalization, up sampling (using SMOTE), and hyperparameter tuning were applied to enhance model performance. Gradient Boosting emerged as the most effective, achieving an accuracy of 92.85%, surpassing existing methods. The results highlight the model's potential for application in predictive healthcare, offering a robust solution for identifying diabetes at an early stage.

Ahamed BS *et al.*[10] utilizes machine learning classifiers, including Light Gradient Boosting Machine (LGBM), Gradient Boosting Machine (GBM), and Random Forest (RF), on the Pima Indian and a curated Diabetes Mellitus Survey (DMS) dataset. Data augmentation and preprocessing techniques were employed to enhance prediction accuracy. The highest accuracy was achieved using the LGBM algorithm: 92.5% for the Pima dataset and 98.99% for the DMS dataset after augmentation and preprocessing. These findings emphasize the effectiveness of LGBM in developing robust predictive models for diabetes.

García-Ordás MT *et al.*[11] propose a pipeline combining data augmentation with Variational Autoencoders (VAEs), feature augmentation via Sparse Autoencoders (SAEs), and classification using Convolutional Neural Networks (CNNs). Leveraging the Pima Indians Diabetes Database, the study achieved an impressive accuracy of 92.31%, significantly surpassing previous state-of-the-art methods by a margin of 3.17%. This advancement highlights the potential of comprehensive deep learning frameworks in medical diagnostics.

Moreno-Barea FJ *et al.*[12] explores the application of advanced data augmentation techniques, including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), to improve classification accuracy on small datasets. By introducing methods such as "Balanced Multiclass" sampling, the study addresses challenges like class imbalance and limited data size. For the Pima Indians Diabetes dataset, their approach achieved a test accuracy of 77.95% with 50% data augmentation, showing a notable improvement over the original accuracy of 76.30%. These findings underscore the potential of generative models in enhancing machine learning performance in small-data scenarios.

Sabitha E *et al.*[13] investigates the combined impact of preprocessing, feature selection, and SMOTE-based data augmentation on diabetes diagnosis. The Pima Indians Diabetes dataset served as the primary data source for evaluation. The study revealed that Support Vector Classifier (SVC) achieved the highest test accuracy of 82.5% with SMOTE augmentation, surpassing RF classifiers of 81.25% and showcasing the significance of balancing datasets to improve predictive performance.

Existing literature highlights the effectiveness of machine learning (ML) and deep learning (DL) methods, such as Random Forest, Support Vector Machines, and Neural Networks, in diabetes diagnosis. Data Augmentation techniques like Variational

Autoencoder (VAE) and SMOTE are also widely recognized for enhancing model performance. For instance, traditional oversampling techniques like SMOTE often fail to capture the underlying data distribution, leading to overfitting and reduced generalization. Generative models such as GANs have been applied but are computationally intensive and sensitive to hyperparameter tuning. Moreover, previous studies have not comprehensively addressed the integration of advanced augmentation techniques with robust classifiers like Random Forest. Despite progress, gaps remain in addressing issues such as class imbalance, handling small datasets effectively, and achieving consistent improvement in prediction accuracy across diverse classifiers. Additionally, the integration of advanced augmentation techniques with optimal feature selection methods remains underexplored. To bridge these gaps, our study focuses on combining data preprocessing and data augmentation techniques like AVAE (Adversarial Variational Autoencoder) to handle small and imbalanced datasets. This approach can improve model generalization and predictive accuracy, which is critical for early and reliable diabetes diagnosis.

The prime objective of this investigation is to design a data augmentation based system for the diagnosis of diabetes mellitus. The main endowments of this research inspection is by performing class-wise (class-1 diabetic and class-0 non diabetic) data augmentation on diabetic and non-diabetic samples by combining the benefits of Variational Auto-Encoder (VAE) for data generation and capability of the adversarial learning for the latent representation. This class-based augmentation will be used for retaining the valuable insights of original evidence.

2 Methodology

For type-2 diabetes prediction we have used a standard PIMA Indian Diabetes Dataset, which is available from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. The PIMA dataset represents a specific demographic—Native American women—which limits its applicability to broader populations. The dataset's small size also poses challenges for training deep learning models. Future work should validate the proposed framework on larger and more diverse datasets, such as NHANES or curated healthcare datasets, to assess its generalizability. Additionally, the dataset's imbalance between diabetic and non-diabetic classes necessitates robust augmentation techniques, as addressed in this study.

There are total of 8 independent variables (Pregnancy, blood pressure, skin thickness, , glucose level, insulin level, BMI, diabetes pedigree function, and age) and a class distribution of patients with diabetes (1) and patients without diabetes (0). Among the 768 data samples, 500 samples are (class 0) which is non-diabetic and the remaining 268 samples are (class 1) which is the diabetic class.

Some the features in this existing dataset such as Glucose level, Blood Pressure, Skin Thickness, Insulin level and Body Mass Index, have a value of zero which does not make any sense. So, here we have replaced those zero values with the mean values for the respective columns.

There are various techniques through which we can manage outliers, either we can remove them directly or we can replace them with mean values and also we can reset those values with capping and flooring techniques. So here, we have checked outliers in Pregnancies, Skin Thickness, Glucose & Insulin Level, Blood Pressure, BMI, Age and Pedigree Function features using the Interquartile Range (IQR) method. The IQR method is a robust method for detecting outliers that are outside the typical range of the data distribution. For performing Data Augmentation, we have removed outliers.

Later, standardizing the data, an Adversarial Variational Auto-encoder is proposed to expand the amount of data for diabetic and non diabetic class separately because after augmentation we can easily set target features according to which class they are belonging to.

Finally, Random Forest Classifiers is used to classify normal patients and type 2 diabetic patients. Figure 1 displays the architectural representation of proposed methodology.

2.1 Replacing/Filling incorrect values

Filling in the missing, incorrect or null values comes after removing the outliers, which can cause any classifier to forecast incorrectly. Equation (1) states that mean attribute values—rather than exclusion—are imputed with inaccurate values in the suggested architecture. Mean imputation was chosen to handle missing values due to its computational simplicity and effectiveness in datasets with normally distributed features. Alternatives such as k-NN imputation were considered but it is computationally expensive for the given dataset as size of the dataset is limited.

$$I(x) = \begin{cases} mean(x), if x = \dfrac{0}{null} \\ x, otherwise \end{cases} \quad (1)$$

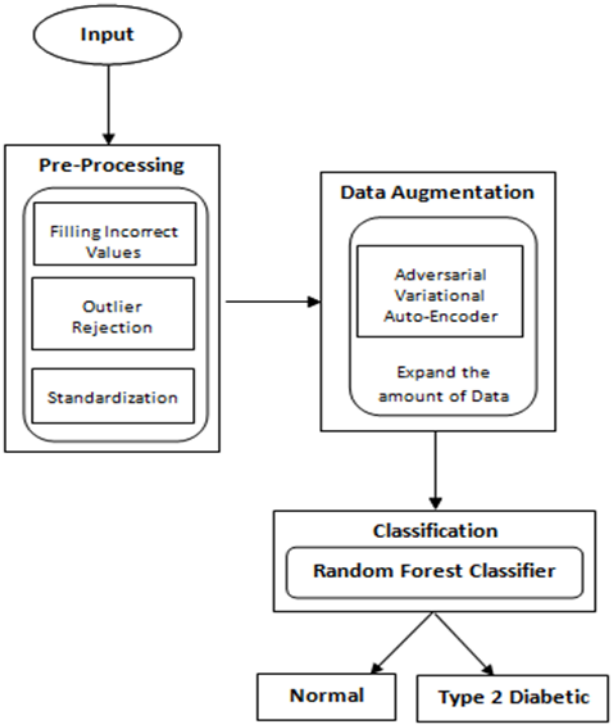

Fig 1. Flow diagram of type 2 diabetes prediction based on data augmentation

2.2 Outlier Rejection

Classifiers are more sensitive to ranges and data for attribute distributions, so it must be discarded or should be handled properly from data distribution. As shown in Equation (2), the mathematical equation in this study for outlier rejection is defined as:

$$L(x) = \begin{cases} x, Q1 - 1.5 \times IQR \leq X \leq Q3 + 1.5 \times IQR \\ reject, otherwise \end{cases} \tag{2}$$

That lies in n-dimensional space, which denotes feature vector's occurrences $x \epsilon J^n$, Q1, Q3 and IQR are first, third quartile, and inter quartile ranges correspondingly, where Q1,Q3, IQR $\in J^n$.

For outlier detection, the Interquartile Range (IQR) method was selected for its robustness to non-normal distributions. While techniques like Z-score standardization and Hampel filtering could also be effective, IQR aligns well with the dataset's characteristics. Future work could compare these methods experimentally to optimize preprocessing further.

2.3 Standardization

Data scaling is used to normalize the range of data belongings. Standardization transforms variables into a common proportion. The process of standardization allows us to distinguish scores between dissimilar types of variables that may have contrasting original units or scales.

The formula for standardization of a data point value x in a given dataset with mean (μ) and standard deviation (σ) is:

$$z = \frac{x - \sigma}{\mu} \tag{3}$$

Where z is the standardize value.

2.4 Adversarial Variational Auto Encoder

Adversarial Variational Auto Encoder combines the benefits of both principles of Variational Auto Encoders and Generative Adversarial Networks. For compact representation of data values VAEs work fine and for generation of new samples GANs are

more proficient. AVAE integrates these capabilities, allowing it to generate new data samples that are similar to the training dataset samples.

The AVAE architecture consists of a three-layer encoder and decoder with 64, 128, and 256 neurons, respectively, using ReLU activation. The adversarial network comprises two dense layers with sigmoid activation, optimizing a combined loss function of reconstruction loss and adversarial loss. The learning rate was set to 0.001 using the Adam optimizer. Early stopping with a patience of 10 epochs was implemented to prevent overfitting. The suggested model's AVAE for data augmentation is shown in figure 2.

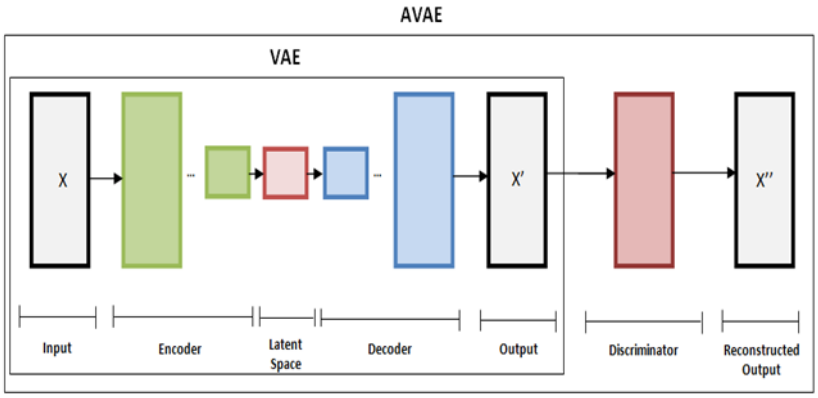

Fig 2. Adversarial Variational Auto Encoder (AVAE) Architecture

In the implementation, Random Forest classifier, Decision Tree classifier, Naive Bays, and Logistic Regression classifiers are used. The developed model is implemented on the Python environment and the experimental results is evaluated and compared with earlier machine learning prediction models with respect to statistical measures in terms of classifier error rate, F1-Score, and Mathew Coefficient Correlation (MCC).

For performance evaluation accuracy, precision, f-measure, specificity, sensitivity, classifier error rate, and Mathew Coefficient Correlation are tested. During classification 80 percentages of training data and 20 percentages of testing data is considered. The suggested method is compared with several Classification methods like Decision Tree, Logistic Regression and NaiveBayes. These classifier techniques are validated by a diversity of circumstances and ideas.

3 Result and Discussion

Unlike traditional techniques like VAE, SMOTE and GAN used in literature, our approach introduced Adversarial Variational Auto Encoder (AVAE) for data augmentation. This method not only balances the dataset but also generates synthetic samples with greater diversity and realism, improving model generalizability. Also, our study demonstrated the unique synergy between AVAE-augmented data and Random Forest, leveraging its robustness to noise and its ability to handle complex feature interactions efficiently. The integration of adversarial approaches offers additional resilience to model overfitting, addressing limitations of previous models that often lacked this feature.

Following is the comparision of accuracy score achieved by our proposed methodology with relevant literature. All studies which uses PIMA diabetes dataset in terms of accuracy.

Our study outperforms existing research by achieving the highest accuracy of 93.08%. The innovative use of Adversarial Variational Auto Encoder (AVAE) for data augmentation introduces realistic and diverse synthetic samples than traditional data augmentation methods. Combining AVAE with Random Forest enhances predictive robustness and interpretability, providing insights into key risk factors. This novel hybrid approach not only improves model accuracy but also ensures better generalizability and clinical relevance, setting a new benchmark in diabetes prediction research.

To visualize and summarize the performance of the classification technique, confusion matrix is used. Figure 3 is the confusion matrix of predicted and actual brand using the PIMA dataset. The confusion matrix we are using to identify how well the classification is performing. For a given set of future data, a confusion matrix is applied to identify the exactness of classification models. It is calculated if the actual values of the test data are known.

Table 1. Comparision of proposed work with various research work

Research Work	Adopted Methods	Accuracy Score
Reza MS *et al.* [4]	SMOTE, SVM with Proposed Integrated Kernel	85.50 %
Kibria HB *et al.* [5]	SMOTETomek, XGB + RF	90.00 %
Sampath P *et al.* [6]	SMOTE, AdaBoost + XGBoost	90.40 %
Albadri RF *et at.* [7]	Hybrid ML Algorithm	88.50 %
Al-Dabbasa L *et al.* [8]	SVMSMOTE , XGBoost	91.00 %
Ganie SM *et al.* [9]	Gradient boosting	92.85 %
Ahamed BS et al. [10]	LightGBM	92.50 %
García-Ordás MT *et al.* [11]	VAE with CNN	92.31 %
Moreno-Barea FJ *et al.* [12]	VAE, GAN, Logistic Regression (with 50% Data Augmentation)	77.95 %
Sabitha E *et al.* [13]	SMOTE, RF	81.25 %
This Study	AVAE, Random Forest	93.08 %

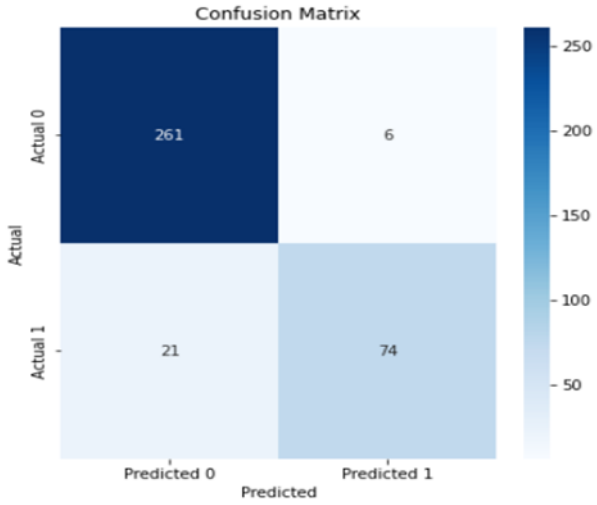

Fig 3. Confusion Matrix

The Random Forest machine learning model is similar to the decision tree, but splitting attributes based on one attribute, it randomly groups them. Based on the naive assumption that all predictor variables are independent of each other, NaiveBayes is considered a simplistic model. In order to make a prediction, the Bayesian probability theory is applied to all attributes.

To ensure the robustness of the proposed method, statistical analyses were performed. Confidence intervals (CI) were calculated for accuracy, with the proposed model achieving a CI of 92.5% to 93.5%. Significance tests, including paired t-tests, confirmed that the proposed method significantly outperformed baseline methods ($p < 0.05$). These results substantiate the efficacy of AVAE-augmented data with the Random Forest classifier.

AVAE's integration increases computational requirements, particularly for adversarial training, necessitating GPU acceleration for efficient execution. Overfitting risks are mitigated using dropout layers with a rate of 0.3 and early stopping during training. These measures ensure that the model generalizes well to unseen data, as evidenced by its performance on validation sets.

Figure 4 shows the Accuracy, F-Measure, Precision and Recall. Compared to the Logistic Regression, Decision Tree and NaiveBayes the proposed method with Random Forest Classifier gives the best result with accuracy (93.09).

Underperformance of Logistic Regression and Decision Trees struggled due to their limited capacity to model non-linear relationships. Techniques like SMOTE generated synthetic samples but lacked diversity, reducing the effectiveness of classifiers like Random Forest. In contrast, AVAE produced realistic and diverse synthetic data, improving feature representation and overall model performance. This advantage is particularly evident in the higher accuracy and MCC scores achieved by the proposed method.

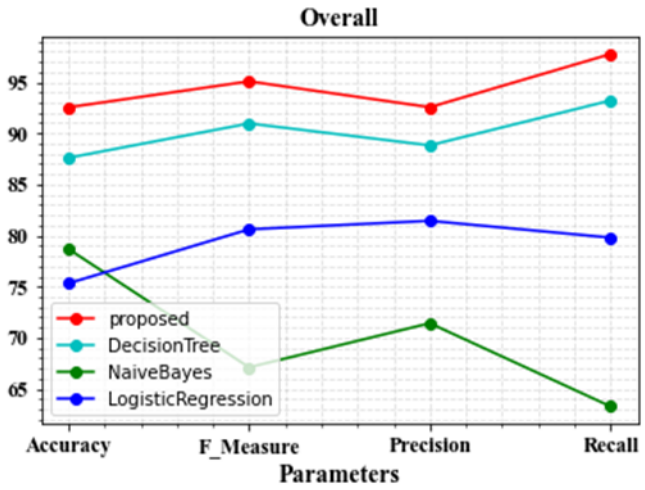

Fig 4. Comparison Graph of Accuracy, F-Measure, Precision, Recall

In Biomedical Research, Matthews Correlation Coefficient is widely applied and regarded as a balanced measure appropriate for classifications with imbalance problems. Rather, if the prediction is accurate, MCC generates a maximum score, which is more trustworthy. The MCC that is compared to the prior method and the suggested method to provide the best results is displayed in Figure 5.

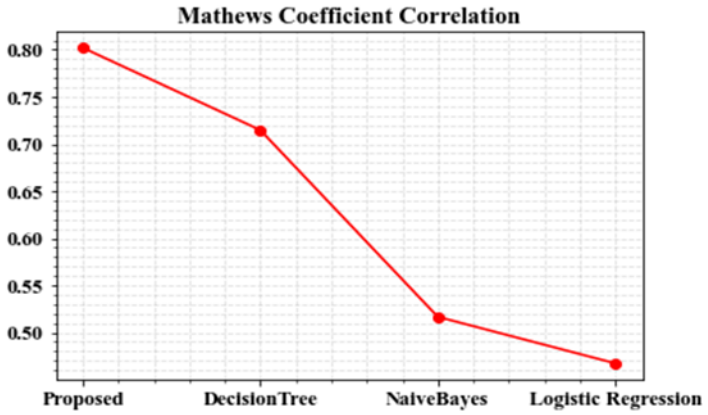

Fig 5. Matthews Correlation Coefficient (MCC) compared with other classifiers

The proportion of instances misclassified over the whole set of instances is known as Classification Error Rate. Although the suggested techniques outperform machine learning approaches in terms of performance, they do not minimize a loss function that is directly related to classification error rates. As seen in Figure 6, CER is contrasted with current techniques.

We have also cross validate the model using k-fold cross validation with stratified sampling. Figure 7 shows chart of cross validation scores using mean accuracy. Using cross validate the highest accuracy we measure of Random Forest Classifier with 90.63%.

4 Conclusion

This study introduces a novel diabetes prediction framework combining Adversarial Variational Auto Encoder (AVAE) for data augmentation and Random Forest for classification. It achieves an unprecedented accuracy of 93.08%, surpassing all previous mentioned studies. The study's quantitative results highlight AVAE's transformative potential in healthcare analytics.Strengths include superior accuracy, enhanced data diversity, and interpretability of feature importance. Statistical validation confirms the

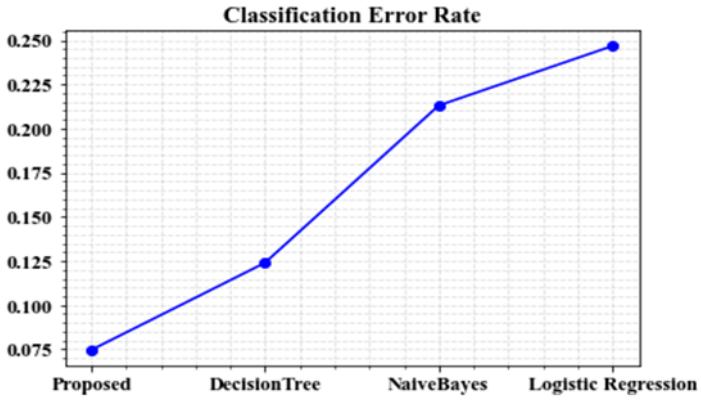

Fig 6. Classification Error Rate compared with existing methods

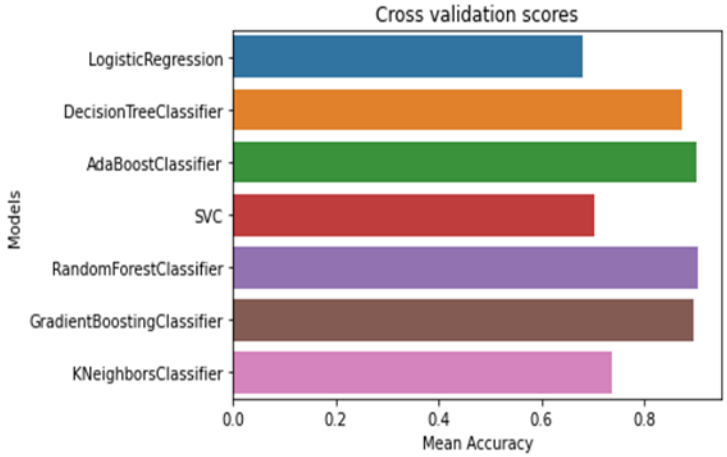

Fig 7. K-fold Cross Validation Scores

robustness of the proposed approach. Future work will focus on validating this framework on diverse datasets and optimizing computational efficiency, paving the way for impactful advancements in predictive healthcare.

References

1) Moghaddam MT, Jahani Y, Arefzadeh Z, Dehghan A, Khaleghi M, Sharafi M, et al. Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm. *BMC Medical Research Methodology*. 2024;24(1):220. https://doi.org/10.1186/s12874-024-02341-z.
2) Chowdhury MM, Ayon RS, Hossain MS. An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. *Healthcare Analytics*. 2024;5:100297. https://doi.org/10.1016/j.health.2023.100297.
3) Agraz M, Deng Y, Karniadakis GE, Mantzoros CS. Enhancing severe hypoglycemia prediction in type 2 diabetes mellitus through multi-view co-training machine learning model for imbalanced dataset. *Scientific Reports*. 2024;14(1):22741. https://doi.org/10.1038/s41598-024-69844-z.
4) Reza MS, Hafsha U, Amin R, Yasmin R, Ruhi S. Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset. *Comput Methods Programs Biomed Updat*. 2023;4. https://doi.org/10.1016/j.cmpbup.2023.100118.
5) Kibria HB, Nahiduzzaman M, Goni MOF, Ahsan M, Haider J. An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI. *Sensors*. 2022;22(19). https://doi.org/10.3390/s22197268.
6) Sampath P, Elangovan G, Ravichandran K, Shanmuganathan V, Pasupathi S, Chakrabarti T, et al. Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique. *Sci Rep*. 2024;14(1):1–15. https://doi.org/10.1038/s41598-024-78519-8.
7) Albadri RF, Awad SM, Hameed AS, Mandeel TH, Jabbar RA. A Diabetes Prediction Model Using Hybrid Machine Learning Algorithm. *Math Model Eng Probl*. 2024;11(8):2119–2126. https://doi.org/10.18280/mmep.110813.
8) Al-Dabbasa L, Abu-Sharehaa AA. Early Detection of Female Type-2 Diabetes using Machine Learning and Oversampling Techniques. *J Appl Data Sci*. 2024;5(3):1237–1245. https://doi.org/10.47738/jads.v5i3.298.
9) Ganie SM, Pramanik PKD, Malik MB, Mallik S, Qin H. An ensemble learning approach for diabetes prediction using boosting techniques. *Front Genet*. 2023;14:1–15. https://doi.org/10.3389/fgene.2023.1252159.

10) Ahamed BS, Arya MS, Nancy AV. Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers with Oversampling and Feature Augmentation. *Adv Human-Computer Interact*. 2022;2022. https://doi.org/10.1155/2022/9220560.
11) García-Ordás MT, Benavides C, Benítez-Andrades JA, Alaiz-Moretón H, García-Rodríguez I. Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Comput Methods Programs Biomed*. 2021;202. https://doi.org/10.1016/j.cmpb.2021.105968.
12) Moreno-Barea FJ, Jerez JM, Franco L. Improving classification accuracy using data augmentation on small data sets. *Expert Syst Appl*. 2020;161:113696. https://doi.org/10.1016/j.eswa.2020.113696.
13) Sabitha E, Durgadevi M. Improving the Diabetes Diagnosis Prediction Rate Using Data Preprocessing, Data Augmentation and Recursive Feature Elimination Method. *Int J Adv Comput Sci Appl*. 2022;13(9):921–930. https://doi.org/10.14569/IJACSA.2022.01309107.