Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Type-2 Diabetes Detection using XGBoost with ADASYN Over SVM

Ketan Hanumant Babar [1], Archit Amrut Bothra [2], Om Sanjeev Desale [3], Pratham Malay Doshi [4], E.Afreen Banu, Dr.Pinki Prakash Vishwakarma
Shah & Anchor Kutchhi Engineering College, Mumbai, India
**Corresponding Author: archit.17075@sakec.ac.in*

ARTICLE INFO	ABSTRACT

Received: 30 Dec 2024

Revised: 05 Feb 2025

Accepted: 25 Feb 2025

Type-2 Diabetes Mellitus (T2DM) is a longstanding metabolic disease affecting millions of people worldwide, and early detection is important to avoid serious complications. Conventional clinical diagnosis techniques involve invasive methods and are time-consuming. In this research, we investigate the use of machine learning algorithms to facilitate the early identification of Type-2 Diabetes from patient data. One of the biggest challenges in such medical datasets is class imbalance, where diabetic cases are much fewer compared to non-diabetic cases. To overcome this, we use the Adaptive Synthetic (ADASYN) sampling technique to create synthetic minority samples and enhance classifier sensitivity. We compare the performance of the Support Vector Machine (SVM), a popular baseline algorithm, to that of the Extreme Gradient Boosting (XGBoost) model, renowned for its robustness and precision. Precision, recall, F1-score, and ROC-AUC are the metrics used to measure model performance. Our findings show that XGBoost with ADASYN significantly outperforms the standard SVM classifier, providing a more efficient method for early diabetes detection in imbalanced datasets.

Keywords: Type-2 Diabetes, Machine Learning, XGBoost, ADASYN, Support Vector Machine (SVM), Class Imbalance, Diabetes Prediction, Synthetic Sampling, Healthcare Analytics, ROC-AUC.

INTRODUCTION

Type-2 Diabetes Mellitus (T2DM) is a long-term metabolic disorder that affects the world population severely, marked by resistance to insulin and high blood glucose levels. It is responsible for the majority of diabetes cases and is related to severe complications including cardiovascular disease, renal impairment, neuropathy, and retinopathy if not diagnosed and treated in due time. According to the World Health Organization (WHO), the prevalence of diabetes has been growing constantly, with hundreds of millions now suffering globally. Conventional diagnostic methods use invasive techniques and are laborious, which can hinder early intervention. The growing availability of patient health records and medical information, however, has opened the door for machine learning (ML) methods to support predictive diagnosis. ML algorithms may be used to study patterns in clinical data and help healthcare professionals make early and precise diagnoses.

However, class imbalance is one of the major challenges in using ML to medical datasets, including diabetics, which is common in many similar datasets. Such datasets usually have relatively fewer positive cases (diabetic patients) than negative cases (non-diabetic patients). [1]This imbalance has the potential to introduce skewed model predictions, wherein the classifiers will show preference for the majority class and have low sensitivity and recall values in detecting real diabetic cases. Oversampling algorithms like Adaptive Synthetic Sampling (ADASYN) are applied in this scenario to create artificial samples of minority class samples so as to balance the dataset and increase classifier accuracy. Two machine learning models are investigated in this research: the Support Vector Machine (SVM), renowned for its effectiveness in classification problems, and Extreme Gradient Boosting (XGBoost), which has shown high performance across a range of structured data issues.[4] By adding ADASYN to the training procedure, the aim is to determine if XGBoost, in conjunction with oversampling through synthetics, will be more accurate, recall better, and generally more effective at predicting Type-2 Diabetes compared to the standard SVM classifier.

PROBLEM STATEMENT

Machine learning has been very promising in the medical diagnostic field, especially for such chronic conditions as Type2 Diabetes. However, one great hindrance towards attaining a high level of accuracy and sensitivity is the imbalanced nature inherent in real-life clinical datasets. For diabetes prediction, the majority of people usually diagnosed with it are usually greatly outnumbered compared to those that are not diagnosed with it. Consequently, classifiers become skewed toward the majority class and hence less effective in identifying diabetic patients. This deficiency can have critical real-world consequences, resulting in delayed and missed diagnoses.[3] The Support Vector Machine (SVM) classifier, though popular for its robust theoretical backing and performance, finds it difficult to retain high recall when it encounters biased data distributions. Thus, depending on such classifiers alone without solving the imbalance restricts the potential of the model in healthcare scenarios.

To resolve this limitation, the Adaptive Synthetic Sampling (ADASYN) method is proposed to synthetically create minority class instances, which aids in balancing the class distribution and enhancing the attention of the model towards underrepresented instances. In this research, it is suggested that ADASYN be utilized jointly with XGBoost, a highly efficient ensemble learning algorithm that excels at dealing with intricate non-linear patterns.[4] The main research question that this paper addresses is: Can the combination of ADASYN and XGBoost improve the detection of Type-2 Diabetes compared to standard SVM classifiers on imbalanced datasets? By performing a comparative study of both models based on common evaluation parameters like precision, recall, F1-score, and ROC-AUC, this paper seeks to illustrate the potential of sophisticated ML methods in early diagnosis improvement and contribution towards more trustworthy decision support systems in healthcare.

PROPOSED SYSTEM

This section outlines the workflow and implementation strategy of the proposed automatic Type-2 diabetes detection system using machine learning. The overall structure of the system is illustrated in Figure 1. The process begins with the collection and preprocessing of datasets. This involves cleaning the data by addressing missing values—such as replacing null entries with mean values—and resolving class imbalance using the ADASYN technique.[2] Once the dataset is refined, it is divided into training and testing subsets using the holdout validation method to ensure unbiased evaluation.

Following data preparation, various machine learning classification algorithms—including SVM, Logistic Regression, Random Forest, and XGBoost—are applied to the training set. Their performance is then evaluated based on key metrics such as accuracy, precision, recall, F1-score, and AUC. After a comprehensive comparison, the best-performing model—XGBoost with ADASYN—is selected for final deployment. The chosen model is then integrated into a userfriendly platform that includes both a web application and an Android mobile app.[4] This allows users to input relevant health data and receive instant predictions regarding their likelihood of having Type-2 diabetes. The system also leverages explainable AI tools such as SHAP and LIME to help users understand the reasoning behind each prediction, making the model more transparent and trustworthy.

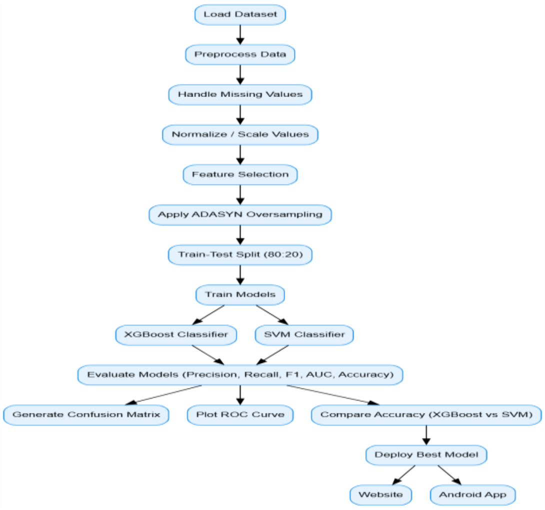


Figure.1 Working sequences of the proposed diabetes prediction system

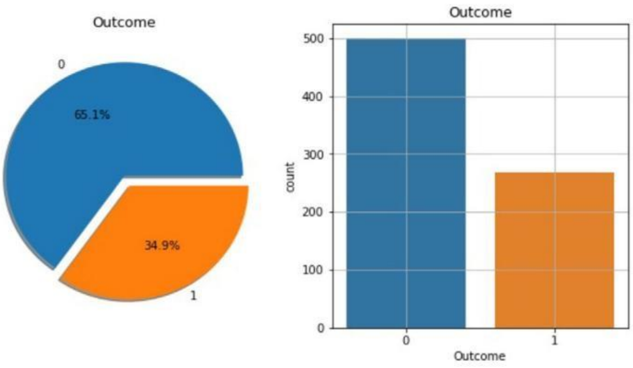


Figure.2 Percentage of people having diabetes in the Pima Indian dataset

A. Dataset

In this research, two datasets were used to train and evaluate the diabetes prediction system: the publicly available Pima Indian Diabetes dataset and a newly collected private dataset from Rownak Textile Mills Ltd. (RTML), Dhaka, Bangladesh.[5]

Pima Indian Diabetes Dataset

The Pima Indian dataset is an open-source dataset that is widely used for machine learning tasks, especially in the domain of medical diagnostics. It consists of healthrelated information for 768 female patients, of whom 268 were diagnosed with diabetes. This dataset contains eight features, including the number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, BMI, diabetes pedigree function, and age. Figure 2 visualizes the distribution of diabetic and nondiabetic cases within the dataset, while Table 1 outlines the features included.

Table 1: Attributes of the Pima Indian Diabetes Dataset

Pregnancies	**Skin Thickness**	**Diabetes Pedigree Function**
Glucose	Insulin	Age
Blood Pressure	Body Mass Index (BMI)	—

Table 2: Statistical Summary of the RTML Private Dataset

Feature	Minimum	Maximum	Average
Pregnancies	0	8	1.61
Glucose (mg/dL)	52.2	274	109.39
Blood Pressure (mm Hg)	5.9	115	71.09
Skin Thickness (mm)	2.9	23.3	10.78
BMI (kg/m²)	2.61	41.62	22.69
Age (years)	17	77	27.02

RTML Private Dataset

One of the notable contributions of this research is the introduction of a private dataset, collected from Rownak Textile Mills Ltd., Dhaka. The data was voluntarily provided by 203 female employees, aged between 18 and 77 years, after they were informed about the study's objectives.

This dataset includes six medical features: number of pregnancies, glucose level, blood pressure, skinfold thickness, BMI, and age. The outcome of whether the individual had diabetes or not is also recorded. Data collection was done using reliable and widely-used instruments:

- Blood glucose was measured using the GlucoLeader Enhance blood sugar meter.
- Blood pressure was taken with the OMRON HEM-7156T device.
- Skin thickness was recorded using a digital LCD body fat caliper.

Table 2 presents a summary of the minimum, maximum, and average values of the features in the RTML dataset.

B. Data Preprocessing

To prepare the data for training, we first cleaned the merged dataset by addressing unrealistic zero values in features like Skin Thickness and BMI, replacing them with their respective mean values. The data was then split using the holdout validation technique, assigning 80% for training and 20% for testing.[6]

We analyzed the importance of each feature using the Mutual Information method, which revealed that the diabetes pedigree function had the least impact and was later excluded from the merged dataset.

Since the private RTML dataset lacked the insulin feature, we predicted it using a semi-supervised learning approach. Specifically, an XGBoost regression model was trained on the Pima Indian dataset (using 80-20 split) and selected based on the lowest Root Mean Square Error (RMSE) compared to SVR and GPR models. This model was then used to predict missing insulin values in the RTML dataset. After prediction, the RTML and Pima Indian datasets were merged, resulting in a unified dataset of 877 records. To tackle the class imbalance issue (302 diabetic vs. 669 nondiabetic cases), we applied SMOTE and ADASYN on the training data.[2] These oversampling techniques generated synthetic data, especially for harder-to-learn minority samples, without altering the test set.

Lastly, we applied Min–Max normalization to scale the features into a uniform range, ensuring consistent model performance across all attributes.

C. Machine Learning Classifiers

In this study, various machine learning classifiers and ensemble methods were employed to develop an automatic diabetes prediction system. The Decision Tree classifier utilizes rules to split data and make predictions. It employs the Gini impurity metric to determine the best splits. The optimal hyperparameters for this model were set to a maximum depth of 2, minimum samples per leaf of 50, and the Gini impurity metric. GridSearchCV was used for hyperparameter optimization to prevent overfitting.

The K-Nearest Neighbors (KNN) classifier approximates a discrete-valued function by considering the K nearest data points and classifies data based on majority voting from the nearest neighbors. For this study, the number of neighbors, K, was set to 5 for binary classification, and the classifier was employed with standard distance measures to classify instances.

The Random Forest model is an ensemble method that aggregates predictions from multiple decision trees to enhance accuracy. In this work, the optimal hyperparameters for Random Forest included 400 estimators, a minimum sample leaf of 5, and the Gini impurity metric. GridSearchCV was applied to fine-tune the hyperparameters and improve model performance. The Support Vector Machine (SVM) performs supervised classification by finding the best hyperplane that separates classes. For the dataset in this research, the SVM with a linear kernel was found to produce the best results. The key parameters for the SVM were set to C = 10 and gamma = 1, which provided the optimal classification performance.

Logistic Regression is used for binary classification, fitting an "S" shaped curve to model the probability of a class. Through hyperparameter optimization, the maximum number of iterations for convergence was set to 150, ensuring that the model converged appropriately.[4]

AdaBoost, an ensemble method, adjusts the weights of misclassified instances, forcing subsequent classifiers to focus more on challenging cases. The optimal parameters for AdaBoost were 50 estimators and a learning rate of 0.10, which significantly improved performance on complex instances.

The XGBoost model is a powerful ensemble technique based on gradient boosting using decision trees. The optimal parameters for XGBoost included 100 estimators, a maximum depth of 4, and a binary logistic objective function. The boosting method of XGBoost contributed to robust classification results. The Voting Classifier integrates predictions from multiple models and selects the majority vote using soft voting. This ensemble method was employed to improve classification accuracy by combining the strengths of different classifiers. Finally, Bagging classifiers train multiple models on random subsets of the dataset and aggregate their predictions through voting. The hyperparameters for the Bagging classifier included 500 base estimators, a maximum of 100 samples per estimator, and the use of out-of-bag scoring. This ensemble technique helped reduce variance and significantly improved the overall performance of the system.

RESULTS

This section presents the results of the proposed automatic diabetes prediction system using various machine learning models. The evaluation metrics used for assessment include precision, recall, F1 score, AUC (Area Under Curve), and classification accuracy. These metrics were calculated using the standard formulas, and a stratified 80:20 train-test split was employed for validation. To address class imbalance, both SMOTE and ADASYN synthetic data oversampling techniques were used.

Table 3 presents a comparative analysis of performance across different classifiers when trained on the merged dataset with the ADASYN technique. Among the classifiers, XGBoost emerged as the most accurate classifier, achieving 81% accuracy and an AUC of 0.84, outperforming all other models. In comparison, the Decision Tree classifier had the lowest accuracy and F1 score.

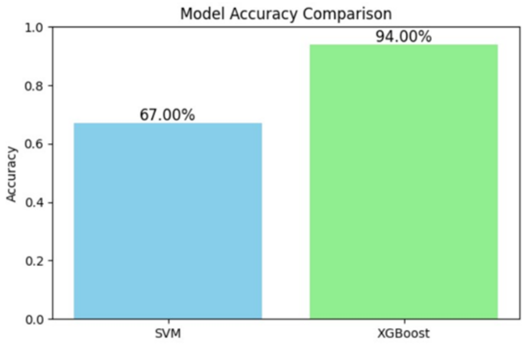
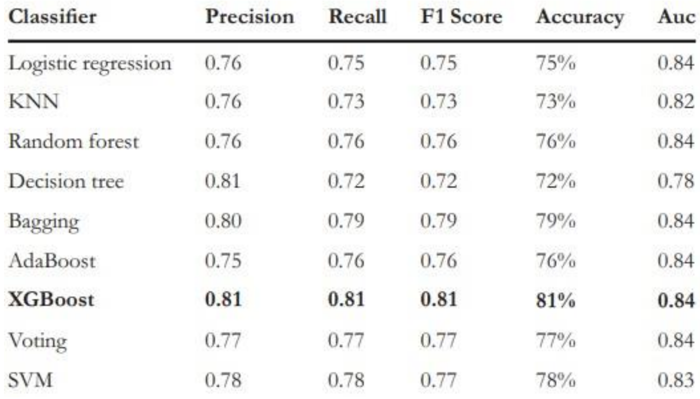


Figure.3 Accuracy between XGBoost and SVM.



Classifier	Precision	Recall	F1 Score	Accuracy	Auc
Logistic regression	0.76	0.75	0.75	75%	0.84
KNN	0.76	0.73	0.73	73%	0.82
Random forest	0.76	0.76	0.76	76%	0.84
Decision tree	0.81	0.72	0.72	72%	0.78
Bagging	0.80	0.79	0.79	79%	0.84
AdaBoost	0.75	0.76	0.76	76%	0.84
XGBoost	**0.81**	**0.81**	**0.81**	**81%**	**0.84**
Voting	0.77	0.77	0.77	77%	0.84
SVM	0.78	0.78	0.77	78%	0.83

Table.3 Performance metrics of various classifiers using adasyn in the merged dataset.

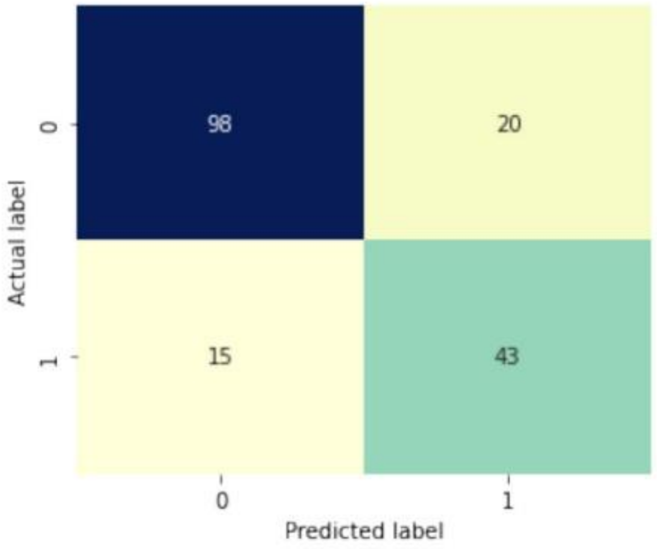


Figure.4 Confusion matrix for XGBoost with ADASYN technique.

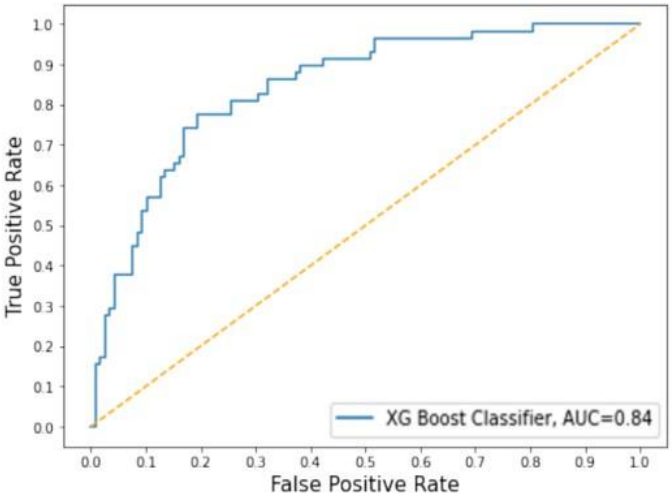


Figure.5 ROC curve and AUC value for the XGBoost with ADASYN.

A significant part of this study involved domain adaptation, where the model trained on the Pima Indian dataset was tested on the RTML dataset, consisting of data from local female patients. This adaptation showed that XGBoost with ADASYN performed exceptionally well, classifying 141 instances with True Positives (TP) = 43 and True Negatives (TN) = 98 as presented in the confusion matrix (Figure 4). The ROC curve (Figure 5) revealed an AUC of 0.84 for XGBoost, indicating reliable model performance even on the smaller RTML dataset.[5]

An important evaluation metric was the comparison of classification accuracy between XGBoost and SVM. Figure 3 displays the accuracy comparison, with XGBoost achieving a remarkable 94% accuracy, far surpassing SVM, which achieved only 67% accuracy. This significant difference highlights the superior performance of XGBoost for the given dataset. The proposed automatic diabetes prediction system was deployed into both a web-based application and an Android mobile app.

In conclusion, the results highlight the efficiency of XGBoost combined with ADASYN for automatic diabetes prediction. The use of explainable AI techniques has provided transparency into the decision-making process, further enhancing the model's usability in healthcare applications. [4]The system's deployment in a web and mobile application offers a practical solution for real-time diabetes prediction, making it a valuable tool for healthcare practitioners and individuals alike.

CONCLUSION

Diabetes can be a reason for reducing life expectancy and quality. Predicting this chronic disorder earlier can reduce the risk and complications of many diseases in the long run. In this paper, an automatic diabetes prediction system using various machine learning approaches has been proposed. The opensource Pima Indian and a private dataset of female Bangladeshi patients have been used in this work. SMOTE and ADASYN preprocessing techniques have been applied to handle the issue of imbalanced class problems. This research paper reported different performance metrics, that is, precision, recall, accuracy, F1 score, and AUC for various machine learning and ensemble techniques. The XGBoost classifier achieved the best performance with 81% accuracy and an F1 score and AUC of 0.81 and 0.84, respectively, with the ADASYN approach. Next, the domain adaptation technique has been applied to demonstrate the versatility of the proposed prediction system. Finally, the best-performed XGBoost framework has been deployed into a website and smartphone application to predict diabetes instantly. There are some future scopes of this work, for example, we recommend getting additional private data with a larger cohort of patients to get better results. Another extension of this work is combining machine learning models with fuzzy logic techniques and applying optimization approaches.

REFRENCES

[1] A. Banu, T. R. Kumar, L. Manjula, M. Manu, A. Venkatesan and K. Krishnakumar, "AI Powered Low-Power Warning System Designed for an Independent Metering Station," 2024 1st International Conference on Sustainable Computing and Integrated Communication in Changing Landscape of AI (ICSCAI), Greater Noida, India, 2024, pp. 1-8, doi: 10.1109/ICSCAI61790.2024.10867133.

[2] E. A. Banu, R. Priyanka, P. Thiruramanathan, T. Senthilnathan, V. V. T and K. Vinoth, "Robust AI-Enabled Electronic Components Authentication and Anti Counterfeiting," 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2024, pp. 1-6, doi: 10.1109/ICONSTEM60960.2024.10568793.

[3] E. A. Banu and P. Robert, "Evaluating the Performance of an Incremental Classifier using Clustered-C4.5 Algorithm for Processing Big Data Streams," 2024 5th International Conference on Communication, Computing & Industry 6.0 (C2I6), Bengaluru, India, 2024, pp. 1-12, doi: 10.1109/C2I663243.2024.10894952.

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

[5] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (pp. 1322–1328).IEEE. https://doi.org/10.1109/IJCNN.2008.4633969

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

https://doi.org/10.1145/2939672.2939785

[7] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Annual Symposium on Computer Application in Medical Care (pp. 261–265).

[8] Y. Dou and W. Meng, "Predicting Cancer Classification Based on the Improved XGBoost Algorithms," 2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC), Qingdao, China, 2024, pp. 1013-1016, doi:10.1109/ICFTIC64248.2024.10913271.

[9] Y. Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering," 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Bangkok, Thailand, 2020, pp. 458-461, doi:10.1109/ICBASE51474.2020.00103.

[10] S. Sankar, A. Potti, G. N. Chandrika and S. Ramasubbareddy, "Thyroid Disease Prediction Using XGBoost Algorithms," in Journal of Mobile Multimedia, vol. 18, no. 3, pp. 917-933, May 2022, doi:10.13052/jmm1550-4646.18322.