2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM 2023)
16-17 June 2023, Gazipur-1707, Bangladesh

Machine Learning-based Diabetes Prediction: A Cross-Country Perspective

Sadia Afrin Shampa*, Md. Saiful Islam*, Ayatun Nesa†
* Institute of Information and Communication Technology (IICT),
Bangladesh University of Engineering and Technology (BUET) Dhaka, Bangladesh
† Dept. Of Laboratory Medicine, BIRDEM General Hospital

2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM) | 979-8-3503-1600-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/NCIM59001.2023.10212596

***Abstract*—High blood sugar levels characterize diabetes and is a chronic disease with long-lasting effects on human health. Accurately predicting diabetes occurrence presents challenges due to the limited availability of labeled data and outliers or missing values in diabetes datasets. In diabetes research, machine learning (ML) algorithms are extensively employed to analyze datasets and predict the onset of the disease. In this study, diabetes data from Bangladesh, India, and Germany were examined using various ML models. The experimental results demonstrate that the Bangladesh dataset performs better using boosting ML algorithms such as AdaBoost, CatBoost, Gradient Boost, and XGBoost. These algorithms effectively predict the occurrence of diabetes. Additionally, satisfactory performance was observed with basic models like Random Forests and Decision Trees, as evaluated by performance metrics. Early detection of diabetes plays a crucial role in mitigating associated risk factors and severity. ML algorithms have emerged as valuable tools in diabetes prediction, leveraging the available data to make accurate predictions. The study's findings underscore the potential of boosting ML algorithms, such as AdaBoost, CatBoost, Gradient Boost, and XGBoost, in predicting diabetes based on the Bangladesh dataset. Furthermore, the study acknowledges the acceptable performance of basic models like Random Forests and Decision Trees in evaluating diabetes data. In conclusion, this study contributes to the understanding of diabetes prediction by analyzing datasets from multiple countries. The results highlight the effectiveness of ML algorithms, particularly boosting algorithms, in accurately predicting diabetes occurrence. This knowledge can aid researchers, healthcare professionals, and policymakers in implementing strategies for early detection and management of diabetes, ultimately improving patient outcomes and overall public health.**

***Index Terms*—Diabetes, Machine Learning Models, Missing values, Outliers.**

I. Introduction

Diabetes is a long-lasting condition that affects essential functions of the human body. The International Diabetes Federation (IDF) estimates that by 2022, approximately 537 million adults worldwide will have diabetes [1]. According to IDF, about 13.13 million people in Bangladesh suffer from diabetes [2]. The number of diabetic patients in Bangladesh is rapidly increasing so there is a large amount of data available on diabetic patients, which has sparked the interest in exploring the data from a broader perspective.

In recent years these data have been analyzed using various Machine Learning (ML) algorithms to find out latent meanings from the data for predicting diabetes [1], diabetic retinopathy severity [2], [3], hypoglycaemic events [4] of the diabetic patients. However, diabetes prediction has become a challenging task for researchers as the data used are often high dimensional, and also the availability of data for its confidential nature. Few studies have been conducted for diabetes prediction using ML approaches. For example, in [5], the authors investigated diabetes, focusing on some external factors such as Glucose, BMI, age, insulin, etc. Hasan et al. [6] proposed a framework for diabetic prediction based on a weighted ensemble of different ML models on the publicly available dataset (PIMA Indian Diabetes Dataset). A similar study using ML models is also carried out in [7] on the same dataset. These studies have investigated publicly available data, and in most cases, the amount of data used is insufficient due to the availability and confidential nature of data. Though very few studies are carried out in the context of Bangladesh [8], [9] using ML approaches, none of the research is conducted in cross country perspective with Bangladeshi data using ensemble ML approaches. Therefore, further studies are required to investigate these issues using ensemble ML approaches in the context of Bangladesh and other countries data.

In this paper, the study focuses on predicting diabetes using machine learning techniques. The study also explores how the prediction results vary across countries like India and Germany. An in-depth analysis and comparison of diabetes prediction is conducted across different countries. The objectives of the research are as follows. Firstly, to propose other ML models to classify diabetes based on around fifteen thousand patient data from Bangladesh. Secondly, to find out best performed ML models for predicting diabetes. Thirdly, to analyze and investigate diabetes prediction across three different data sets of three other countries (Bangladesh, India, and Germany).

This document is structured as follows. In the following section, a description of the research methodology used to analyze diabetes data is provided (Section II). The subsequent section offers a detailed analysis and discussion of the results obtained from various machine learning models (Section III). Section **??** discusses the findings and outcomes of the adopted methods for predicting diabetes in this paper. The final section concludes the article with a discussion of limitations and overall conclusions.

TABLE I: Summary of the dataset

Country	No of Patients /Samples	Normal	Pre-diabetic	Diabetic
Bangladesh	14401	1552	1871	10978
India (PIMA)	768	500	-	268
Germany	2000	3	-	1997

TABLE II: Summary of the dataset after oversampling

Country	No of Patients /Samples	Normal	Pre-diabetic	Diabetic
Bangladesh	14401 (Before Sampling)	1552	1871	10978
	32934 (After Oversampling)	10978	10978	10978

II. Research Method

This section briefly discusses the steps that have been followed in this study. The research methodology includes two major phases such as building the ML models and data analysis and visualization. (see Figure 1). The research methodology includes the following:

A. Data Acquisition

The diabetes data used for analysis in this study was obtained from Popular Diagnostic Center. The dataset includes information from approximately 15,000 patients, including those who have been diagnosed with diabetes (positive), those who are prediabetic, and those who do not have diabetes (negative). The dataset contains 25 different feature values, including Glucose level, Insulin, BMI, Creatinine, Cholesterol, and Glycohemoglobin, among others. To supplement this dataset, additional data was collected from two other sources: PIMA Indian diabetes dataset, which is publicly available that contains information on 768 patients, and a dataset from a hospital in Frankfurt, Germany, which includes data from 2000 patients. The summary of the dataset is presented in Table-I.

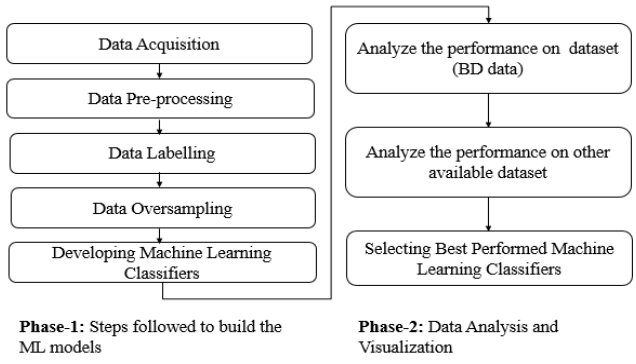


Fig. 1: Diabetes Prediction model

B. Data Pre-processing

The data preprocessing stage involves transforming and cleaning the collected dataset to ensure it is in an efficient and usable format. This preprocessing includes several key steps, such as identifying and rejecting any outliers in the dataset, filling in any missing values, standardizing the data, oversampling, and selecting the most relevant features for analysis. To identify the optimal features, three standard techniques for feature selection are used: Principal Component Analysis (PCA), Independent Component Analysis (ICA), and correlation-based techniques. These methods are compared to evaluate their performance on the collected dataset.

C. Data labeling

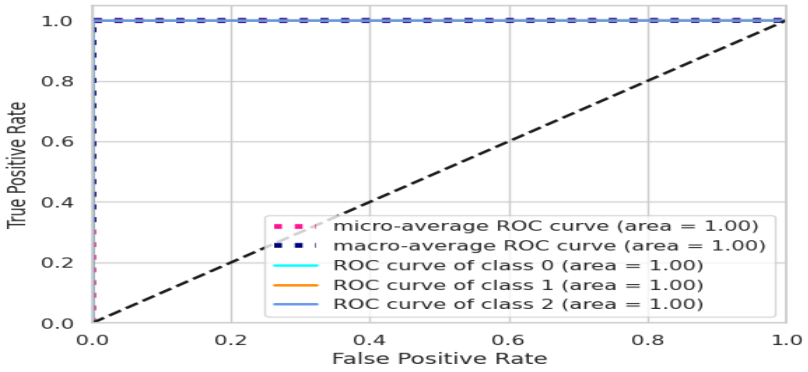
The Bangladeshi diabetes dataset collected from various hospitals and diagnostic centers was preprocessed and labeled for diabetes classification. The dataset was labeled under three categories: 1) Diabetic, 2) Pre/Undiagnosed Diabetic, and 3) Normal. The labeling process involved two main features, namely the fasting plasma glucose test and the oral glucose tolerance test. The fasting glucose test was conducted after an 8-hour fast in the morning, while the oral glucose tolerance test involved measuring the blood glucose level after an overnight fast and then after drinking a sugary drink. A fasting glucose value of 126 mg/dL or higher was considered indicative of diabetes, a value between 100-125 mg/dL indicated pre-diabetes, and a value less than 100 mg/dL was considered normal. Similarly, an oral glucose tolerance test value of 200 mg/dL or higher was indicative of diabetes, a value between 140-199 mg/dL indicated pre-diabetes, and a value less than 140 mg/dL was considered normal. The labeling information is summarized in Table-II.

D. Data Oversampling

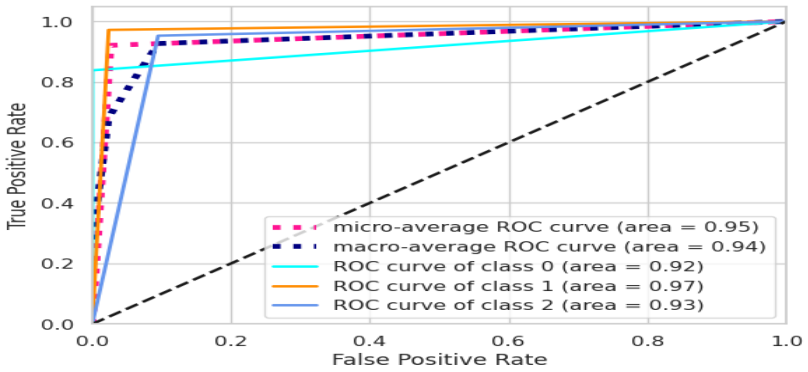
The labeled dataset used in this study had a class imbalance problem, with varying quantities of diabetes, pre-diabetes, and normal labels. The distribution of labeled data indicated that the number of diabetes patients was higher than that of pre-diabetes and normal patients (See Table-I). This class imbalance can create overfitting problems during the training process in ML algorithms. In this study, an oversampling method called adaptive synthetic (ADASYN) [10] employs SMOTE [11] sampling algorithm to produce more synthetic data for minority class examples that were more difficult to learn by classifiers. The method was applied to ensure similar numbers of instances for all three classes in both the train and test sets. The class frequency of the dataset after using the oversampling method is presented in Table II.

III. Analyzing Machine Learning Models

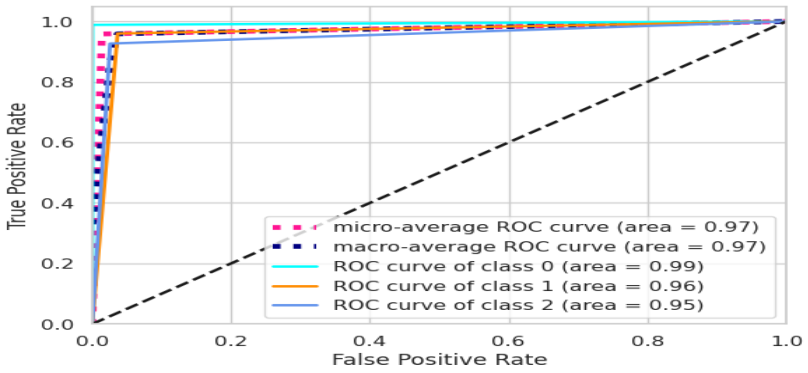
This section explores various machine learning (ML) models to classify diabetes into different labels. The algorithms are trained by performing hyper-parameter tuning using. To develop and analyze these models, we used the Python programming language and the sci-kit-learn library [12]. We randomly split the manually labeled dataset into an 80-20



(a)



(b)



(c)

Fig. 2: ROC Curves for the developed models: (a) CatBoost, (b) Naive Bayes and (c) SVM

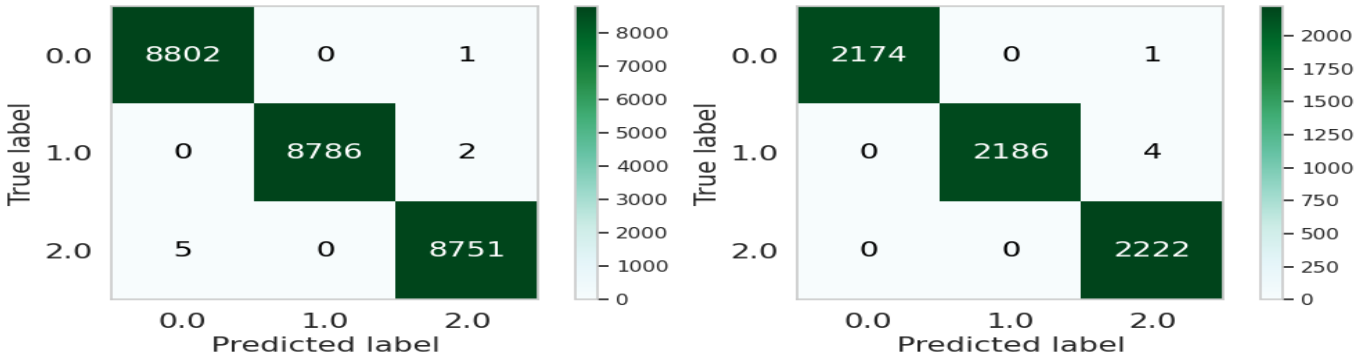
train-test distribution, where 80% of the data was used as training data and 20% of the data was used as test data.

A. Decision Tree

Decision tree (DT) is a popular hierarchical ML technique for classification and prediction. This non-parametric supervised ML approach is applied to our diabetes dataset to predict diabetes by using decision nodes in the test function to identify local regions through a series of iterations of separation. Each internal node in the tree tests an attribute, and each branch represents a test result. Finally, each leaf node specifies a class tag. Decision trees are easy to understand and interpret, making them popular in applications where model transparency is essential.

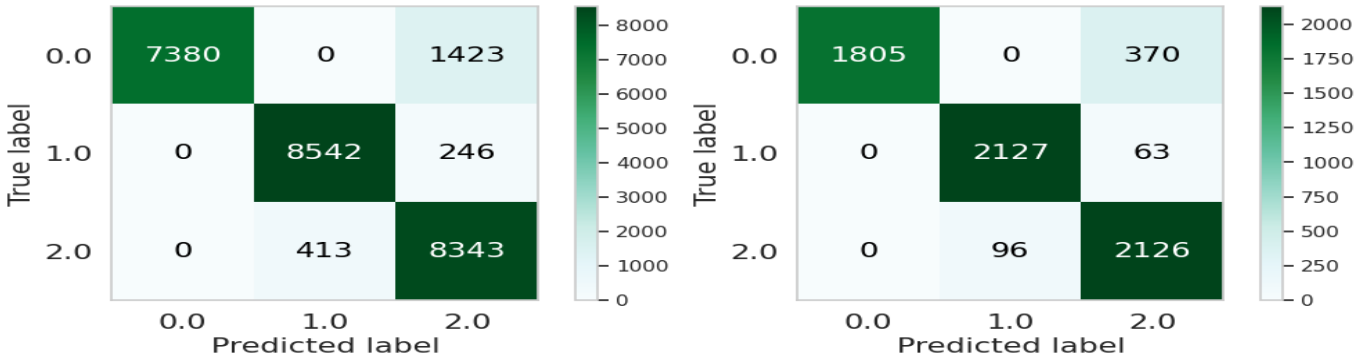
B. Naive Bayes

Naive Bayes (NB) is a probabilistic machine learning technique based on Bayes' theorem. It assumes the presence of one feature in a class is independent of the existence of other features. This feature independence assumption simplifies the model and makes it easy to build and apply to large datasets. Naive Bayes has been shown to perform well compared to even more complex machine learning models. The NB algorithm is computationally efficient and can handle large datasets with high-dimensional features.



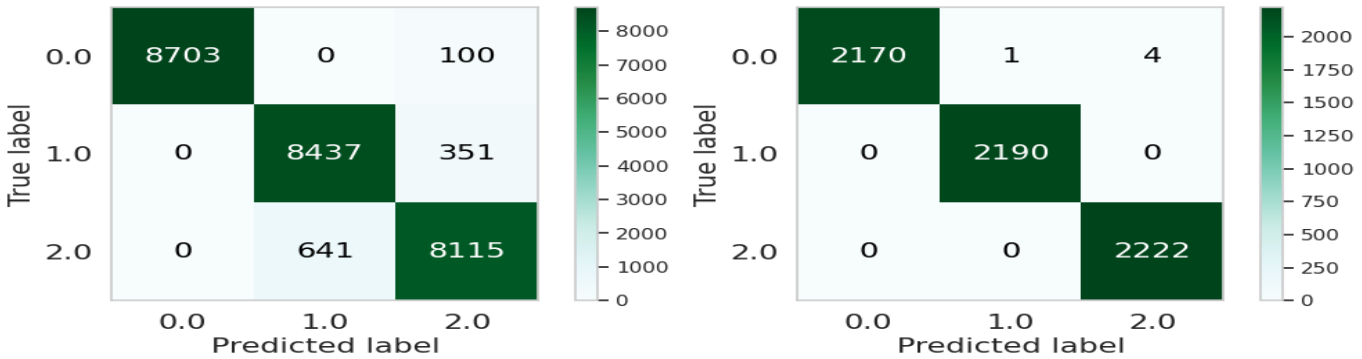
Train Data Test Data

(a)



Train Data Test Data

(b)



Train Data Test Data

(c)

Fig.3: Confusion matrices for developed models: (a) CatBoost, (b) Naive Bayes and (c) SVM

TABLE III: Performanc of ML models on of the datasets

Algorithm			AB	CB	XB	GB	SVM	RF	DT	KNN	ANN	NB
Bangladeshi Data	Train Data	Accuracy	1	1	1	1	0.934	1	1	1	0.999	0.866
		Precision	1	1	1	1	0.846	1	1	1	0.997	0.816
		Recall	1	1	1	1	0.812	1	1	1	0.998	0.92
		F1_score	1	1	1	1	0.825	1	1	1	0.998	0.839
	Test Data	Accuracy	1	0.999	1	1	0.936	1	1	0.994	0.998	0.862
		Precision	1	0.997	1	1	0.841	1	1	0.983	0.995	0.81
		Recall	1	1	1	1	0.812	1	1	0.985	0.996	0.916
		F1_score	1	0.998	1	1	0.826	1	1	0.984	0.995	0.836
PIMA Data	Train Data	Accuracy	0.825	0.955	0.873	0.913	0.799	1	1	1	0.806	0.754
		Precision	0.812	0.961	0.866	0.913	0.794	1	1	1	0.79	0.73
		Recall	0.796	0.941	0.851	0.894	0.75	1	1	1	0.777	0.717
		F1_score	0.803	0.95	0.857	0.902	0.763	1	1	1	0.783	0.722
	Test Data	Accuracy	0.792	0.831	0.779	0.792	0.831	0.818	0.753	0.701	0.805	0.805
		Precision	0.768	0.81	0.753	0.768	0.817	0.797	0.725	0.676	0.782	0.782
		Recall	0.777	0.816	0.758	0.768	0.797	0.797	0.729	0.69	0.787	0.787
		F1_score	0.772	0.813	0.755	0.768	0.805	0.797	0.727	0.68	0.784	0.784
Germany Data	Train Data	Accuracy	1	1	1	0.994	0.914	0.899	0.817	0.813	0.823	0.756
		Precision	1	1	1	0.995	0.918	0.893	0.811	0.797	0.81	0.73
		Recall	1	1	1	0.992	0.891	0.882	0.773	0.782	0.789	0.717
		F1_score	1	1	1	0.993	0.903	0.887	0.786	0.789	0.798	0.722
	Test Data	Accuracy	0.99	0.99	0.985	0.97	0.9	0.89	0.815	0.815	0.81	0.785
		Precision	0.993	0.984	0.985	0.974	0.901	0.88	0.802	0.788	0.784	0.753
		Recall	0.984	0.993	0.98	0.956	0.861	0.858	0.746	0.768	0.756	0.724
		F1_score	0.988	0.988	0.982	0.964	0.877	0.868	0.764	0.777	0.767	0.735

C. Support Vector Machine

Support Vector Machine (SVM) is a supervised ML technique for classification and regression tasks. This algorithm creates hyperplanes in high-dimensional or infinite-dimensional spaces to differentiate between different classes of data points. The objective is to find a hyperplane that maximizes the margin between the data points and their respective categories. SVMs help handle high-dimensional feature spaces and have built-in overfitting protection. However, they are primarily used for classification problems. This machine learning technique was employed in our work to classify diabetes prediction from the numeric dataset.

D. Random Forest

The Random Forest (RF) classifier trains multiple decision trees in parallel and then aggregates the results to make predictions. This approach is also an ensemble method because it combines various models to improve performance. Each decision tree in the Random Forest is trained on a random subset of features and data points, called bootstrapping. The final output of the Random Forest is determined by taking the majority vote of all the individual decision trees. It has become a popular method for classification and regression tasks and is known for its ability to handle high-dimensional data and reduce overfitting.

E. Artificial Neural Network

The classification task in our study is achieved by utilizing an Artificial Neural Network (ANN). An ANN is composed of interconnected units called artificial neurons, which can modify their internal structure by adjusting the weight of the connections based on the input data. This complex network can capture intricate patterns or knowledge by analyzing vast amounts of data, enabling data mining and artificial intelligence applications to analyze large datasets. The ANN's initial parameter settings, such as weight, bias, and learning rate, are crucial to its performance in the classification task.

F. CatBoost

CatBoost is a machine learning algorithm specifically designed to handle categorical data effectively. Unlike other machine learning algorithms that require preprocessing of categorical data, CatBoost can directly take flat features during training. It is an efficient algorithm that provides state-of-the-art results in regression and classification tasks and can compete with other leading machine-learning techniques. It generates a robust predictive model using a greedy approach similar to gradient descent in function space. It creates an ensemble of weak predictive models, such as decision trees, to produce a robust predictive model.

G. AdaBoost

AdaBoost is one of the most widely used boosting algorithms in machine learning. The algorithm works by iteratively creating a solid classifier by combining several weak classifiers in a weighted sum. The weights are updated at each iteration based on the performance of the weak classifiers on the training data. The idea behind AdaBoost is to give more weight to misclassified data points so that the subsequent vulnerable learners focus more on these data points, eventually improving the classifier's overall performance.

H. Gradient Boosting

Gradient Boosting is a popular technique used in solving regression and classification problems. It involves creating an ensemble of weak predictive models or essential learners, typically decision trees. The method is designed in stages, with an initially chosen loss function for optimization. The

key concept is to gradually improve the model by fitting a new decision tree to the residuals of the previous one, thus minimizing the loss function. This big idea arose from the observation of Leo Breiman, and later explicit algorithms for regression gradient improvement was developed by Jerome H. Friedman, and more general perspectives on functional gradient improvement were proposed by L. Mason et. to [13].

I. XGBoost

XGBoost stands for eXtreme Gradient Boosting and is a practical implementation of gradient boosting methods developed by Tianqi Chen and later enhanced by contributions from various developers under the Distributed Machine Learning Community. This technique is well-known for its parallel tree-boosting feature, which provides accurate and efficient solutions to data science problems. It is popular in ML competitions like Higgs Machine Learning Challenge and Kaggle competitions dealing with structured or tabular data. It is also used to predict online review polarity based on customer purchase decisions, where ranking scores are extracted as critical features from the data.

IV. Discussions and Conclusions

The results of our study provide valuable insights into the prediction of diabetes occurrence using machine learning (ML) algorithms. It was found that boosting ML algorithms, such as AdaBoost, CatBoost, Gradient Boost, and XGBoost, outperformed other models when applied to the diabetes dataset from Bangladesh. The potential of these algorithms for accurately predicting diabetes in this specific population was highlighted.

The success of the boosting algorithms can be attributed to their ability to handle complex relationships and interactions within the dataset. Nonlinear patterns and dependencies were effectively captured, resulting in enhanced predictive performance. This is particularly advantageous in diabetes prediction, as the disease is influenced by multiple factors and exhibits intricate relationships.

Moreover, it was demonstrated that even basic ML models like Random Forests and Decision Trees showed promising performance in predicting diabetes. While these models may not have surpassed accuracy, their simplicity and interpretability made them valuable options in certain scenarios. These models could elucidate key predictors and their relative importance in determining diabetes occurrence.

Despite the promising results, there are limitations to be considered. Firstly, the study was limited to datasets from Bangladesh, India, and Germany, which may not be representative of other populations. In future studies, including diverse datasets from various regions will ensure the generalizability of the findings. Secondly, the availability of labeled data remains a challenge in diabetes research. Obtaining larger and more comprehensive datasets would further improve the accuracy and robustness of the predictive models.

In terms of future work, integrating additional features and data sources in the prediction models could be explored. For instance, incorporating genetic information, lifestyle factors, and biomarkers could enhance the accuracy and personalized nature of diabetes prediction. Furthermore, the development of hybrid models that combine the strengths of different ML algorithms could potentially yield even better predictive performance.

In conclusion, the effectiveness of boosting ML algorithms in predicting diabetes was highlighted, particularly in the context of the Bangladesh dataset. The contribution of this study to the growing body of knowledge on diabetes prediction and the importance of early detection and intervention were emphasized. Further research addressing the limitations and exploring new avenues will continue to advance our understanding and ultimately improve the prediction and management of diabetes.

References

[1] N. P. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," *Procedia Computer Science*, vol. 167, pp. 706–716, 2020.

[2] D. V. Gunasekeran, D. S. Ting, G. S. Tan, and T. Y. Wong, "Artificial intelligence for diabetic retinopathy screening, prediction and management," *Current opinion in ophthalmology*, vol. 31, no. 5, pp. 357–365, 2020.

[3] J. D. Bodapati, V. Naralasetti, S. N. Shareef, S. Hakak, M. Bilal, P. K. R. Maddikunta, and O. Jo, "Blended multi-modal deep convnet features for diabetic retinopathy severity prediction," *Electronics*, vol. 9, no. 6, p. 914, 2020.

[4] J. Vehí, I. Contreras, S. Oviedo, L. Biagi, and A. Bertachi, "Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning," *Health informatics journal*, vol. 26, no. 1, pp. 703–718, 2020.

[5] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[6] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76 516–76 531, 2020.

[7] T. N. Joshi, P. Chawan *et al.*, "Diabetes prediction using machine learning techniques," *Ijera*, vol. 8, no. 1, pp. 9–13, 2018.

[8] B. Pranto, S. M. Mehnaz, E. B. Mahid, I. M. Sadman, A. Rahman, and S. Momen, "Evaluating machine learning methods for predicting diabetes among female patients in bangladesh," *Information*, vol. 11, no. 8, p. 374, 2020.

[9] M. U. Emon, M. S. Keya, M. S. Kaiser, T. Tanha, M. S. Zulfiker *et al.*, "Primary stage of diabetes prediction using machine learning approaches," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. IEEE, 2021, pp. 364–367.

[10] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 1322–1328.

[11] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[13] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.