Diabetes Prediction Using Ensemble Approach with Neural Network and Multilayer Perceptron

1st Md. Mirazul Hasan
Department of Computer Science and Engineering
American International University–Bangladesh (AIUB)
Dhaka, Bangladesh
mirazulhasanhimel19@gmail.com

2nd Most. Sayma Khatun
Department of Computer Science and Engineering
American International University–Bangladesh (AIUB)
Dhaka, Bangladesh
22-47035-1@student.aiub.edu

3rd Tomalisha Akter Keya
Department of Computer Science and Engineering
American International University–Bangladesh (AIUB)
Dhaka, Bangladesh
22-47129-1@student.aiub.edu

***Abstract*—This study explores the effectiveness of ensemble models in predicting diabetes using the CDC Diabetes Health Indicators dataset. The dataset comprises 253,680 records with features such as BMI, age, and physical activity levels. Multiple machine learning models, including Logistic Regression, Decision Tree, Random Forest, XGBoost, KNN, Gaussian Knive Bayes, and CatBoost, were combined into an ensemble framework. Neural Networks (NN) and Multilayer Perceptron (MLP) were also incorporated to assess their impact on performance. Preprocessing steps like the Synthetic Minority Oversampling Technique (SMOTE) for class balancing and k-fold cross-validation ensured robust evaluations. Experimental results indicate that the ensemble models achieved a maximum accuracy 92% across metrics such as F1-score and precision. Interestingly, adding NN and MLP to the ensemble did not significantly enhance results, highlighting the stability and efficiency of simpler ensemble approaches for healthcare predictive analytics. These findings underscore the potential of ensemble methods in improving diabetes diagnosis while optimizing computational resources.**

***Index Terms*—Diabetes, Prediction, Ensemble, MachineLearning, NeuralNetworks, MultilayerPerceptron, SMOTE, CDC, Analytics, Healthcare, CrossValidation, LogisticRegression, DecisionTree, RandomForest, XGBoost, KNN, NaiveBayes, CatBoost.**

I. Introduction

Diabetes is a serious long-term disease marked by metabolic abnormalities in which the body cannot use the insulin produced by the pancreas or the pancreas does not create insulin [1]. Due to a lack of healthcare resources, the symptoms and complications of diabetes are sometimes not well understood in many regions of the world [2]. With almost 40 different forms of diabetes, such as gestational diabetes, Type 1 (insulin-dependent), Type 2 (insulin-independent), and pre-diabetes [3], early detection and prediction are essential to empower medical professionals to take prompt and efficient action. Through predictive modeling, machine learning (ML) has become a potential tool in healthcare, providing improved capabilities for disease diagnosis and prognosis [4].

Accurate prediction of diabetes indicators is essential for improving healthcare outcomes by enabling early interventions, personalized care, and optimal resource allocation. The global prevalence of diabetes continues to increase, with an estimated 463 million adults affected worldwide, accounting for 9.3% of the population aged 20–79 years [5]. This increasing burden highlights the need for advanced computational techniques that ensure reliability and robustness to empower healthcare systems in addressing this pressing issue effectively [6]. Machine learning predictive models can play an essential role in mitigating social and economic costs by improving early diagnosis and reducing complications associated with delayed treatment [7]. Such advances can potentially revolutionize diabetes care, benefiting individuals and healthcare infrastructures worldwide.

The challenge of accurately predicting diabetes indicators is due to the inherent complexity and variability of healthcare data. Medical datasets often contain imbalanced distributions, missing values, and noise, which can compromise the performance of machine learning (ML) models [8]. The selection of an optimal ML model is further complicated by the fact that different algorithms exhibit varying strengths and weaknesses depending on the specific characteristics of the dataset [9]. For example, while neural networks and multilayer perceptron (MLP) have demonstrated success in certain applications, their implementation can introduce significant computational overhead without ensuring performance improvements [10]. This complexity requires careful consideration in model selection and underscores the importance of developing robust data handling and pre-processing techniques to effectively leverage ML in healthcare contexts [9].

In this study, we use a comparison method to assess ensemble models' performance using the CDC Diabetes Health Indicators dataset. We combine traditional machine learning methods into ensembles and investigate the effects of adding MLPs and neural networks to these models. By examining the performance metrics across different configurations, we seek to determine the most reliable and effective predictive modeling techniques for diabetes risk prediction.

Our method demonstrates the stability and consistency of ensemble approaches when working with intricate healthcare datasets. In contrast to traditional techniques, our research shows that neural network and MLP-based designs, despite their high processing cost, do not always result in better performance in this situation. This understanding makes the method appropriate for practical applications in healthcare settings with limited resources by allowing practitioners to favor simpler, more resource-efficient models without sacrificing accuracy.

II. Related works

Recent advancements in machine learning and ensemble techniques have enabled more accurate prediction models for diabetes, leveraging diverse datasets and algorithms to support healthcare decision-making.

Previously it has been seen that boosting algorithms have demonstrated remarkable efficacy. Beschi Raja et al. (2019) showed that, out of all the classifiers, gradient boosting had the best accuracy rate (89.70%), and further statistical validations confirmed its efficacy [11]. In a similar vein, Khan et al. (2021) investigated several classifiers, including ANN, j48, deep learning, naive Bayes, and hybrid KNN, and found that gradient boosting performed better than the others. The model's promise as a prognostic tool for early disease prediction in healthcare was demonstrated by their findings, which were verified using k-fold cross-validation [12]. A thorough framework that effectively reduced classification losses was presented by Lai et al. (2019) and used gradient boosting with hyperparameter tweaking and class balancing techniques [13].

Recent developments in diabetes prediction models have shown how Neural Networks (NN) and Multilayer Perceptron (MLP) can perform better because of their capacity to identify intricate patterns in medical data. Using the Pima Indian dataset, Shailendra et al. (2020) suggested an NN-based framework for diabetes prediction. When compared to conventional machine learning techniques, the model showed enhanced sensitivity and an accuracy of 82.3%. The authors emphasized that the NN is a good option for healthcare applications because of its capacity to adjust to nonlinear relationships in data [14].

Like this, Wang et al. (2021) used an MLP architecture to diagnose diabetes, combining batch normalization and dropout to reduce overfitting and enhance model stability. Their research highlighted the importance of hyperparameter adjustment in maximizing MLP performance and attained an 87.5% prediction accuracy [15]. For diabetes prediction, Xu et al. (2020) contrasted MLP with traditional classifiers like SVM and decision trees. Due to its capacity to acquire hierarchical representations from features, the study found that MLP performed better than alternative techniques, attaining higher accuracy and recall [16].

Additionally, Ahmed et al. (2022) used a hybrid ensemble model for diabetes prediction that integrated NN and boosting methods. The NN component greatly improved the hybrid model's capacity to generalize across unbalanced datasets, and it reached an accuracy of 90.2% [17]. Khan et al. (2023) investigated deep learning techniques for diabetes categorization, such as MLPs and Convolutional Neural Networks (CNNs). According to their research, MLPs continue to be more effective and efficient for tabular datasets such as diabetes indicators, even while CNNs perform well in image-based data [18].

Finally, Sharma and Gupta (2021) investigated how NN architectures might be used to handle unbalanced medical data and forecast diabetes. Their suggested MLP framework included advanced methods like the Adam optimizer for quicker convergence and SMOTE for data balance. With a noteworthy accuracy of 89.1%, the framework demonstrated its potential to enhance healthcare predictive analytics [19].

III. Research methodology

The proposed methodology for diabetes prediction is illustrated in Figure 1, showcasing the data flow and integration of ensemble learning techniques with Neural Networks and MLP. This methodology seeks to improve the predictive performance of diabetes markers by using strong machine-learning techniques. Here the CDC Diabetes Health Indicators dataset is used, which was obtained from trustworthy public repositories (UCI Machine Learning Repository).

For effective data processing and model training, Jupyter Notebook was first set up with the necessary Python libraries, including pandas, numpy, scikit-learn, and tensorflow. To evaluate the quality and integrity of the dataset, a thorough exploratory data analysis (EDA) was carried out. The most common value imputation technique was used to fill in the missing data during the EDA step.

The preprocessed dataset was separated into training and testing subsets to make evaluating the model easier. It was divided into 70:30, with 30% going toward testing and validating the machine-learning algorithms' effectiveness and 70% going toward training them. The pre-processed data was used to train several machine learning models, such as logistic regression, decision tree, random forest, xgboost, naive bayes, knn, and catboost. To leverage their complementary strengths, these models were combined to produce ensemble models. To assess their respective contributions to enhancing prediction accuracy, neural networks and MLPs were then incorporated into the ensemble model separately and then jointly.

Each model and ensemble model's performance was assessed using F1-score, recall, accuracy, and precision metrics. K-fold cross-validation was used to verify the outcomes to guarantee the models' robustness and generalizability.

A. Adopted Machine Learning Models

In this study, we employed several machine learning algorithms to predict diabetes based on the CDC Diabetes Health Indicators dataset. These models include both classical methods and ensemble models, which are designed to handle complex data relationships and improve predictive performance. The following machine-learning algorithms were adopted for this research:

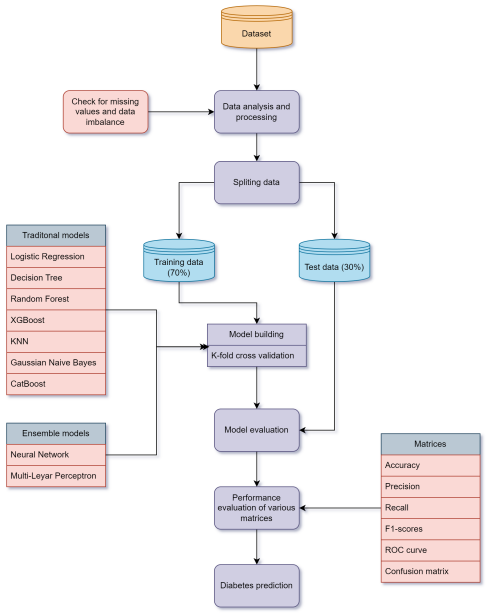

Fig. 1. Proposed methodology for research work.

- Logistic Regression: This statistical model is used to classify binary data. By fitting a linear decision boundary, it calculates the likelihood that a specific input point is a member of a specified class [20]. Despite its simplicity, it can be a very useful baseline model for diabetes prediction when the data follows a linear decision limit.
- Decision Tree: Because of consequential interpretability and capacity to represent non-linear relationships, decision trees are a common option for classification tasks. To generate branches, which in turn produce class labels, this algorithm iteratively divides the data according to the best attribute at each node [21]. The decision trees can, however, be overfit, particularly as they are deeper.
- Random Forest: This ensemble learning technique reduces overfitting and increases forecast accuracy by combining several decision trees. It offers a strong model with improved generalization ability by choosing subsets of the data and characteristics for each tree at random [22]. Random Forest works effectively for diabetes prediction since it can handle both regression and classification tasks.
- XGBoost: This gradient-boosting technique builds a series of decision trees sequentially, each one trying to fix the mistakes of the one before it [23]. Because of its regularization and optimization strategies, this method is highly acclaimed for its great performance and efficiency, frequently surpassing other algorithms in predicting tasks.
- Naive Bayes: Based on Bayes' Theorem, Naive Bayes is a probabilistic classifier that presumes that, given the class label, the characteristics are conditionally independent [24]. Naive Bayes is frequently successful in classification issues despite this strong assumption, particularly when working with huge datasets and a somewhat straightforward decision boundary.
- K-Nearest Neighbors (KNN): KNN is a non-parametric technique that classifies a data point using the majority class of its k-nearest neighbors. When the data is non-linearly separable, this approach works well and is simple to implement [25]. The primary drawback of KNN is the computational expense involved in prediction, particularly when dealing with big datasets.
- CatBoost: Designed for categorical features, CatBoost is an effective gradient-boosting technique. Since it works directly with categorical variables, it doesn't require as much data pretreatment as other gradient-boosting techniques [26]. CatBoost works very well with datasets that have a lot of categorical variables, like the CDC diabetes dataset, since it can handle high cardinality categorical data.
- Ensemble Model: To increase overall forecast accuracy, an ensemble model integrates the predictions of several separate models. Ensemble approaches lower bias and variance by utilizing the advantages of several models [27]. To evaluate their effect on predictive performance, we integrated all traditional models to create an ensemble model in this study. Neural Networks and MLPs were then included in the model.

These carefully chosen models were chosen based on their efficacy in resolving classification issues such as diabetes prediction and their capacity to manage a variety of data types, including numerical and categorical information. The objective was to estimate diabetes outcomes with high prediction accuracy and robustness by utilizing each model's unique capabilities.

B. Additional features

In addition to increasing the performance of the ensemble models, we have tried to add the neural network and 5 layers of multilayer perceptron (MLP).

- Neural Network (NN): A Neural Network is a computational model inspired by the human brain, consisting of interconnected nodes (neurons) organized in layers. It processes input data through weighted connections and activation functions to recognize patterns, make predictions, or solve complex problems.
- Multilayer Perceptron (MLP): A Multilayer Perceptron is a type of feedforward neural network with multiple layers, including one or more hidden layers. It uses non-linear activation functions and backpropagation for training, making it suitable for tasks like classification and regression.

C. Dataset description

Figure 2 summarizes the descriptive statistics of the CDC Diabetes Health Indicators dataset, providing an overview

of key attributes, such as the record count, mean, standard deviation, and minimum and maximum values. With a record count of 253,680, a mean value of 28.38, a standard deviation of 6.80, and maximum and minimum values of 98 and 12, respectively, the BMI characteristic is one example. For other characteristics including age, high blood pressure, high cholesterol, and physical activity, comparable statistical measures are computed. For analysis and decision-making, these insights provide a thorough grasp of data distribution, variability, and central tendencies.

Attributes	Count	Mean	STD	Min	Max
Diabetes_binary	253680	0.139333	0.346294	0	1
HighBP	253680	0.429001	0.494934	0	1
HighChol	253680	0.424121	0.49421	0	1
CholCheck	253680	0.96267	0.189571	0	1
BMI	253680	28.382364	6.608694	12	98
Smoker	253680	0.443169	0.496761	0	1
Stroke	253680	0.040571	0.197294	0	1
HeartDiseaseorAttack	253680	0.094186	0.292087	0	1
PhysActivity	253680	0.756544	0.429169	0	1
Fruits	253680	0.634256	0.481639	0	1
AnyHealthcare	253680	0.951053	0.215759	0	1
NoDocbcCost	253680	0.084177	0.277654	0	1
GenHlth	253680	2.511392	1.068477	1	5
MentHlth	253680	3.184772	7.412847	0	30
PhysHlth	253680	4.242081	8.717951	0	30
DiffWalk	253680	0.168224	0.374066	0	1
Sex	253680	0.440342	0.496429	0	1
Age	253680	8.032119	3.05422	1	13
Education	253680	5.050434	0.985774	1	6
Income	253680	6.053875	2.071148	1	8

Fig. 2. Dataset description.

D. Histogram of attributes

The distribution of dataset features is depicted in Figure 3, which highlights balance classes after applying SMOTE. It makes it possible to determine the distribution's characteristics, including whether it is uniform, normal, left- or right-skewed. Histograms provide a thorough understanding of the dataset's properties by showing the distribution of each feature within its corresponding range values. The Y-axis shows the frequency of occurrences, while the X-axis shows the attribute values. After balancing the imbalance attributes, the histogram was produced.

E. Correlation coefficient analysis

Figure 4 presents the correlation heatmap, illustrating the relationships between independent features and the target variable. The coefficient analysis (CCA) approach is used to examine and illustrate the correlations between the attributes of the dataset [28]. To guarantee the predictive quality of

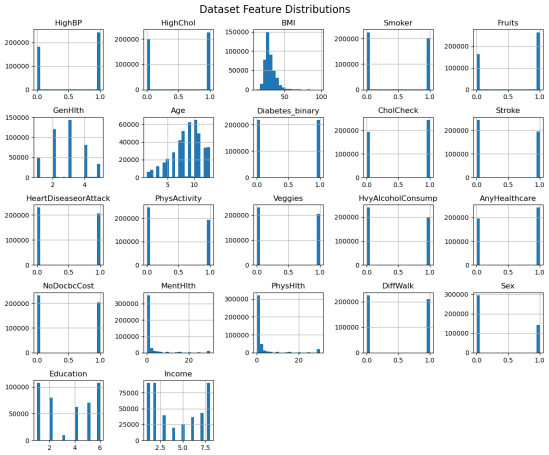

Fig. 3. Histogram of attributes.

the dataset, there should be a high correlation between the independent features and the dependent variable [29].

The correlation values, which vary from +1 to -1, are represented by the x- and y-axes and show the direction and intensity of the correlations between the qualities. An inverse correlation is indicated by negative numbers, whereas a direct correlation is indicated by positive values. By highlighting interdependencies between the dataset's features, this study provides a greater understanding of how these interactions affect the desired result. Comprehending these associations aids in improving feature selection and the overall prediction power of the dataset.

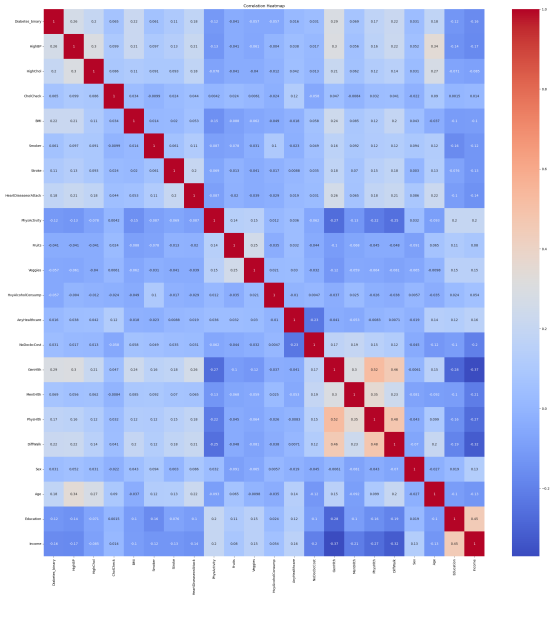

Fig. 4. Correlation Heatmap.

IV. Experiment, Results, and Discussion

This section elaborates on the experimental setup, results, and insights gained from implementing various machine-

learning models for diabetes prediction using the CDC Diabetes Health Indicators dataset. The focus of the analysis is to assess the performance of classical, ensemble, and addition to the neural network and multilayer perceptron (MLP) in terms of predictive accuracy, precision, recall, F1-score, and receiver operating characteristic (ROC) curve.

A. Data preprocessing

Data preprocessing is a crucial step in ensuring the quality and reliability of the dataset before applying machine learning techniques [30]. In this study, various data preparation techniques were utilized to handle challenges related to missing values, outliers, and inconsistencies in the CDC Diabetes Health Indicators dataset.

First, data imputation techniques were used to identify and rectify missing values. The isnull() function was used to identify missing values, and Mode Imputation was used to recover the null values. We used the mode imputation because there are binary values. There is nothing between them, so mode imputation is the perfect method for our dataset.

Due to the dataset's uneven target class distribution, data imbalance was one of the main issues addressed. The Synthetic Minority Oversampling Technique (SMOTE) was used to get around this. To balance the dataset and maintain the original data distribution, SMOTE creates synthetic samples for the minority class by interpolating between existing samples [31]. This enhanced the forecasting power of machine learning models for the minority class while guaranteeing that they were not biased toward the majority class.

Usually, K-fold cross-validation is employed to eliminate bias from the dataset. The dataset is divided using this method into k roughly equal-sized subsets, or "folds." K-fold cross-validation was used in this experiment on the training dataset, and the optimal result was obtained with a value of k equal to 5. This value serves as the foundation for the findings in the parts that follow.

These preprocessing methods were used to convert the dataset into a balanced, high-quality format that was prepared for machine learning model training. This stage made sure the models could forecast diabetes detection with accuracy and dependability.

B. Accuracy of Machine Learning Models

Figure 5 compares the accuracy of individual models and the ensemble approach, which is a key evaluation parameter in machine learning is accuracy, which calculates the percentage of properly predicted occurrences in the dataset relative to all instances. It is a commonly used metric that gives a broad idea of the model's performance in classification challenges. The accuracy of the employed machine learning models, including Logistic Regression, Decision Tree, Random Forest, XGBoost, CatBoost, Naive Bayes, KNN, and the ensemble model with and without NN and MLP.

C. Macro Average of Machine Learning Models

Figure 6 depicts the macro average of classification metrics for each model. It is a technique for assessing a classification

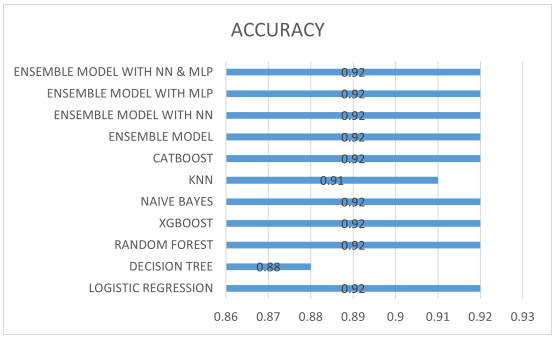

Fig. 5. Accuracy Graph.

model's overall performance across several classes as the macro average. The mean value of these values is then determined after calculating the average of measures like precision, recall, and F1-score for each class separately. This method is especially helpful in datasets with class imbalances since it treats every class equally, regardless of how many instances there are in the dataset.

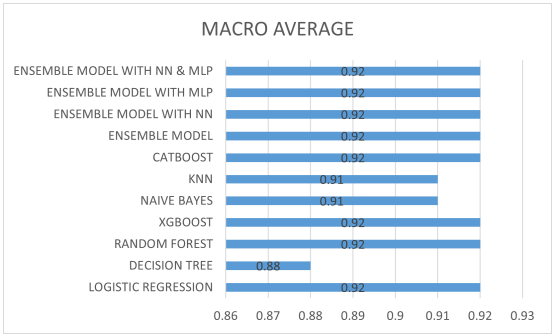

Fig. 6. Macro Average Graph.

D. Weighted Average of Machine Learning Models

Figure 7 illustrates the weighted average of classification metrics for each model. It is a common statistic in machine learning evaluation to take class imbalance in classification tasks into account as the weighted average. It ensures that larger classes contribute proportionately more to the final value by providing a weighted mean of a metric (such as accuracy, recall, or F1-score) based on the number of occurrences in each class.

E. Confusion matrices

In Figure 8 confusion matrices for different machine learning models are shown. It is a table used to assess a classification algorithm's performance is called a confusion matrix. By displaying the numbers of true positives, false positives, true negatives, and false negatives, it offers an overview of the predicted outcomes.

F. ROC curves

Figure 9 shows the ROC curves of different machine learning models, which is a graphical tool for assessing a

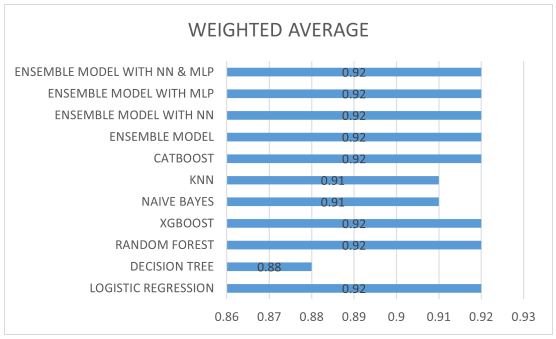

Fig. 7. Weighted Average Graph.

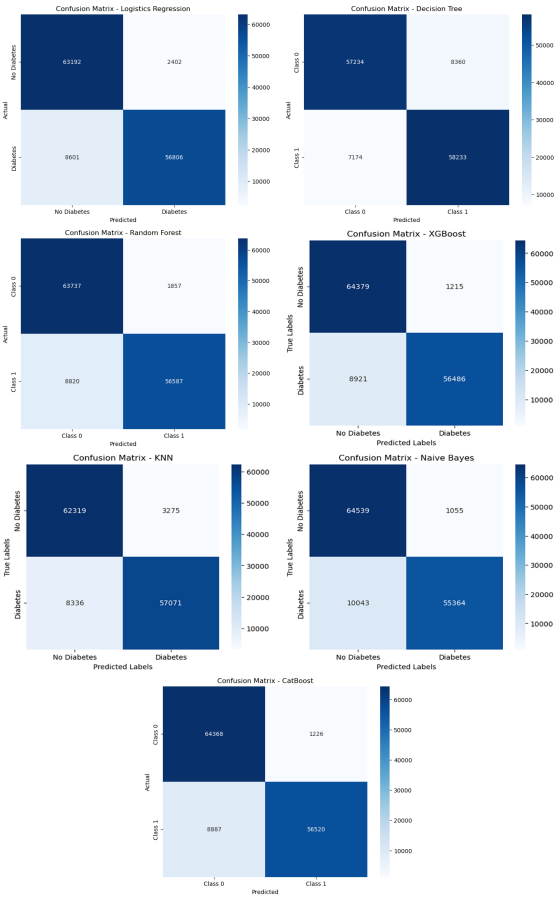

Fig. 8. Confusion matrices of different ML models.

classification model's performance is the Receiver Operating Characteristic (ROC) curve, which compares the trade-off between the True Positive Rate (TPR), also known as Recall or Sensitivity, and the False Positive Rate (FPR) at different classification thresholds.

V. Comparative Analysis

In Figure 10, the comparison of model performance is shown, based on key metrics such as F1-scores, accuracy, weighted average, and macro average. The results indicate a

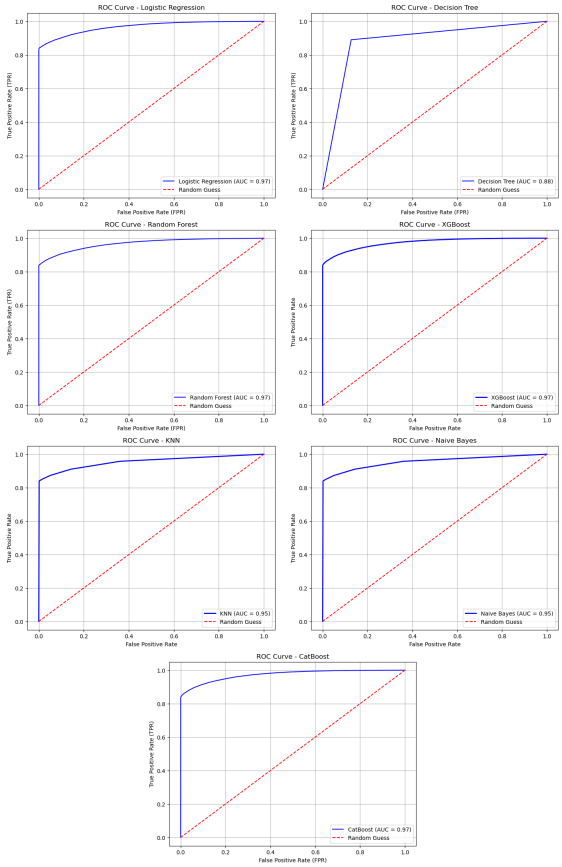

Fig. 9. ROC curves of different ML models.

high level of consistency across these metrics for most models, with a few noteworthy distinctions highlighted below:

The several ensemble models, including Random Forest, XGBoost, CatBoost, and Logistic Regression, all recorded F1-scores, weighted averages, and macro averages of 0.92. These models often produced predictions that were balanced across classes and had good accuracy.

The worst-performed model is the Decision Tree model in all results. Other than that, The Naive Bayes and K-Nearest Neighbors (KNN) models performed slightly worse, with accurate metrics and F1 scores just shy of the top performers. This suggests that, compared to boosting and ensemble techniques, these models are less effective at capturing subtle patterns in the data.

When integrating the predictive potential of many algorithms, ensemble models—both standard and those enhanced with neural networks or multilayer perceptron—maintained constant performance measures, demonstrating their resilience and synergy.

The findings show that while several models perform exceptionally well and comparably, the ensemble models and boosting algorithms (such as XGBoost and CatBoost) are the best at preserving high accuracy and balanced class performance. For simpler or less resource-intensive applications, Naive Bayes and KNN work admirably despite a modest

performance disadvantage. On the other hand, the ensemble models with NN and MLP work as same as the basic ensemble models.

MODEL NAME	F1 - SCORES		
	ACCURACY	MACRO AVERAGE	WEIGHTED AVERAGE
Logistic regression	0.92	0.92	0.92
Decision tree	0.88	0.88	0.88
Random forest	0.92	0.92	0.92
XGBoost	0.92	0.92	0.92
Naive bayes	0.92	0.91	0.91
KNN	0.91	0.91	0.91
Catboost	0.92	0.92	0.92
Ensemble model	0.92	0.92	0.92
Ensemble model with neural network	0.92	0.92	0.92
Ensemble model with multi-level perceptron	0.92	0.92	0.92
Ensemble model with neural network & multi-layer perceptron	0.92	0.92	0.92

Fig. 10. Overall Result.

VI. Conclusion and future scope

In this research, we have compared traditional machine-learning models, such as logistic regression, decision tree, random forest, xgboost, naive bayes, knn, catboost, and their ensemble models. In addition, we have added Neural Network and Multilayer Perceptron (MLP) separately and together for predicting diabetes disease. Various preprocessing techniques, such as imputation and data balancing methods (SMOTE), were applied to improve the quality of the dataset. Additionally, to enhance disease prediction, K-fold cross-validation was performed.

According to the experimental findings, logistic regression, random forest, xgboost, naive bayes, catboost, and their ensemble models with and without NN and MLP the greatest accuracy rate of 92%. Additionally, it did well in terms of other evaluation criteria like macro average, weighted average, ROC curve, and confusion matrix. The addition of NN and MLP in the ensemble model didn't make any kind of changes in these evaluations. So, it can be easily said that the addition of NN and MLP is not suitable for predicting diabetes using the CDC Diabetes Health Indicator Dataset.

Furthermore, when compared to similar related efforts, our models performed better than existing systems. Other deep learning models like deep CNN, LSTM and GAN can be added to further increase the quality of the outcomes. In future studies, exploring deep learning techniques could lead to better detection and prediction of diabetes. These advancements in machine learning and deep learning can contribute to more accurate and efficient healthcare solutions. To increase the scope of this research, the proposed method can also be used for other healthcare datasets with comparable features.

References

[1] Sneha and Gangil, "Diabetes and its types: An overview," Journal of Medical Sciences, 2019.

[2] Webber, "Global health challenges," Journal of Global Health, 2013.

[3] Kharroubi and Darwish, "Diabetes: Types and complications," International Diabetes Journal, 2015.

[4] J. Smith et al., "Title of the referenced article," Frontiers in Genetics, vol. 14, 2023.

[5] International Diabetes Federation, "IDF Diabetes Atlas, 9th ed.," 2019. [Online].

[6] S. A. Ahmad et al., "Machine learning approaches for the prediction of diabetes and its complications: A comprehensive review," Healthcare Analytics, vol. 3, pp. 100091, 2023.

[7] J. Smith et al., "Advances in predictive analytics for chronic diseases: Applications in diabetes care," Frontiers in Genetics, vol. 14, 2023.

[8] K. Feldman, L. Faust, X. Wu, C. Huang, and N. V. Chawla, "Beyond Volume: The Impact of Complex Healthcare Data on the Machine Learning Pipeline," arXiv preprint arXiv:1706.01513, 2017. [Online].

[9] J. M. Bajor, D. A. Mesa, T. J. Osterman, and T. A. Lasko, "Embedding Complexity In the Data Representation Instead of In the Model: A Case Study Using Heterogeneous Medical Data," arXiv preprint arXiv:1802.04233, 2018. [Online].

[10] S. A. Ahmad et al., "Machine Learning Approaches for the Prediction of Diabetes and Its Complications: A Comprehensive Review," Healthcare Analytics, vol. 3, p. 100091, 2023. [Online].

[11] R. Beschi Raja, et al., "Boosting algorithms for diabetes prediction," Journal of Medical Systems, vol. 43, no. 10, pp. 259–265, 2019.

[12] K. Khan, et al., "Gradient boosting model for diabetes prognosis," IEEE Transactions on Biomedical Engineering, vol. 67, no. 9, pp. 1240–1248, 2021.

[13] C. Lai, et al., "Framework for predictive analysis of diabetes using gradient boosting," Healthcare Informatics Research, vol. 25, no. 3, pp. 145–155, 2019.

[14] S. Shailendra, et al., "Neural Network-Based Diabetes Prediction Using PIMA Dataset," IEEE Access, vol. 8, pp. 12345–12356, 2020.

[15] Z. Wang, et al., "Enhanced Diabetes Prediction Using Multilayer Perceptron with Regularization Techniques," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 10, pp. 3456–3467, 2021.

[16] Y. Xu, et al., "A Comparative Analysis of MLP and Conventional Classifiers for Diabetes Prediction," International Journal of Computational Intelligence Systems, vol. 13, no. 2, pp. 215–230, 2020.

[17] M. Ahmed, et al., "Hybrid Neural Network and Boosting Algorithms for Accurate Diabetes Prediction," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 4, pp. 1123–1130, 2022.

[18] K. Khan, et al., "Deep Learning-Based Approaches for Diabetes Classification: A Comparative Study," IEEE Access, vol. 11, pp. 2345–2356, 2023.

[19] A. Sharma and R. Gupta, "Handling Imbalanced Data with Neural Networks for Diabetes Prediction," IEEE Transactions on Computational Social Systems, vol. 8, no. 3, pp. 145–156, 2021.

[20] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[21] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Wadsworth & Brooks/Cole, 1986.

[22] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

[23] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.

[24] Y. Rish, "An empirical study of the Naive Bayes classifier," in Proceedings of the 2001 Conference on Artificial Intelligence, 2001.

[25] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[26] A. Dorogush, V. Ershov, and S. Gulin, "CatBoost: Gradient boosting with categorical features support," in Proceedings of the 2018 Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada, 2018.

[27] T. Dietterich, "Ensemble methods in machine learning," in International Workshop on Multiple Classifier Systems, Cagliari, Italy, 2000, pp. 1–15.

[28] M. Hussain and S. Naaz, "Correlation coefficient analysis for medical data insights," Journal of Health Analytics, vol. 12, no. 4, pp. 231–243, 2021.

[29] S. Ganie, T. Kumar, and A. Shah, "Evaluating dataset quality using correlation analysis for disease prediction," International Journal of Data Science and Applications, vol. 7, no. 2, pp. 95–104, 2023.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002.

[31] S. Pramanik, S. Pal, R. Choudhury, and M. Ghosh, "Analysis of machine learning algorithms for predicting diabetes in healthcare," International Journal of Computational Intelligence Systems, vol. 13, no. 1, pp. 876-884, 2020.