International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.281.03

Supervised Learning Approaches for Early Diagnosis of Diabetes

Nidhi Malik
Department of Computer Science Engineering
The North Cap University, Gurugram, Haryana, India.

Anjali Bansal
Department of Computer Science Engineering
The North Cap University, Gurugram, Haryana, India.

Mansi
Department of Computer Science Engineering
The North Cap University, Gurugram, Haryana, India.

Abstract - Diabates, a chronic disease, deteriorates a person's health significantly. Stalled detection can lead to various long-term complications like nerve diseases and vision impairment. Early detection conduces prompt intervention and lifestyle changes that result in an upliftment of the overall health. In this work, we have explored 7 supervised machine learning algorithms, namely, Logistic Regression, Decision Trees, SVM, KNN, Random Forest, AdaBoost and XGBoost for timely and effective diagnosis of diabetes. Along with them, we also deploy an ensemble Voting Classifier to escalate the accuracy. The Pima Indian Diabates dataset has been used for training our model.

Keywords – ROC, xGBoost, AdaBoost, SVM, Logistic Regression

I. INTRODUCTION

Diabetes mellitus is a metabolic condition where there is persistently high blood sugar levels, often occurs from insufficient insulin production or the body's resistance to it. Pancreas produced is a type of hormone, that serves to blood sugar regulation by enabling glucose so that it enters cells, where it is used for energy [2] production is impaired or cells become resistant to its effects, glucose amasses in the blood stream, leading to various health complications.

Three types of diabetes mellitus:

1. **Type 1 Diabetes:** A disorder occurs when the immune system attacks the insulin that produces beta cells in the pancreas, making it harder for body to control blood sugar levels.

2.Type 2 Diabetes: A condition, in which the body's cells do not able to respond properly to insulin. This form of diabetes is largely linked to obesity, lack of physical activity and genetic predisposition.

3.Gestational Diabetes: A transient type of diabetes that develops during pregnancy as hormonal changes interfere with insulin sensitivity.

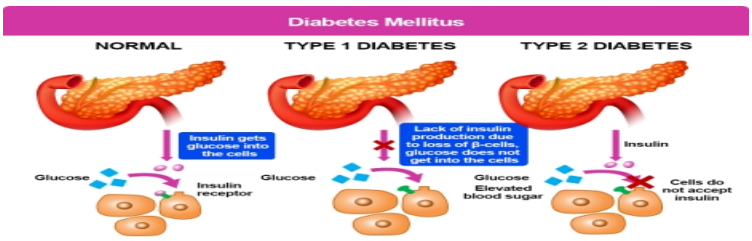

Figure 1. Comparison of Normal, Type1 and 2 diabetes mechanisms [1]

The above visualization [1] illustrates the differences between normal glucose metabolism, Type 1 diabetes and Type 2 diabetes. In a robust pancreas, insulin facilitates glucose uptake into cells. In type 1, the immune system obliterates insulin that produce beta cells [4], leads to insulin deficiency and increase in blood sugar levels. The diagram visually demonstrates the pathophysiology of diabetes mellitus, aiding in understanding disease mechanisms [5].

The rest of the paper is organized as follows. Section II contains the summary of the research conducted in the field till date. A detailed analysis of the dataset used in provided in section III. Proposed techniques are explained in section IV. Experimental results are presented in section V. Concluding remarks are given in section VI.

II. Literature Review

The rising prevalence of diabetes worldwide has intensified the need for improved diagnostic methods that enable early detection and better disease management. These limitations have prompted the integration of machine learning (ML) and AI in healthcare, particularly for enhancing diabetes detection and monitoring. Machine learning algorithms have shown significant potential in predicting diabetes, assessing risk factors, and optimizing treatment plans. Models such as Random Forests (RF), Decision Trees (DT), K-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Neural Networks, have delivered promising results in diabetes classification tasks. Decision Trees provide clear, interpretable rules for classification, while ensemble models like Random Forests improve prediction accuracy by combining multiple trees. Support Vector Machines have been effective in complex and high-dimensional data handling, making them suitable for diabetes prediction [7]. Additionally, Neural Networks, particularly Deep Learning models, excel at capturing complex data relationships, improving the accuracy of prediction outcomes. These advanced models can efficiently analyse large volumes of patient data, uncover hidden patterns, and suggest tailored treatment strategies [8].

III. Detailed Description of the Diabetes Dataset

Feature	Description	Unit/Value Range
Pregnancies	Times the patient has been pregnant	Integer (0 and above)
Glucose	Plasma glucose concentration after 2 hours in an OGTT	mg/dL (0 - 199)
BloodPressure	Diastolic blood pressure	mm Hg (0 - 122)
SkinThickness	Triceps skin fold thickness	mm (0 - 99)
Insulin	2-Hour serum insulin	µU/mL (0 - 846)
BMI	Body Mass Index (weight in kg/(height in m)^2)	kg/m² (0 - 67.1)

DiabetesPedigreeFunction	A function that scores the likelihood of diabetes based on family history	Continuous (0.08 - 2.42)
Age	Age of the patient	Years (21 - 81)
Outcome	Diabetes diagnosis result	0 = No, 1 = Yes

The dataset contains the medically diagnosed measurements for the prediction of diabetes.

- It also includes the physiological attributes and personal information.
- Target variable (Outcome) refers diabetes status: 0 means no diabetes and 1 indicates diabetes present
- Some features have zero values, which represent missing data.
- Prepossessing is required to handle these missing values for accurate analysis

The dataset includes key features linked to diabetes risk. **Pregnancies** correlate with diabetes, especially in older women. **Glucose** is the strongest predictor, with levels above 140 mg/dL indicating risk. **Blood pressure** has a mild correlation, with zero values likely being errors. **Skin thickness** and **Insulin** data show inconsistencies, requiring careful handling. **BMI** strongly links to diabetes, emphasizing obesity as a major factor. **Diabetes Pedigree Function** highlights genetic risk, while **Age** shows increased diabetes risk after 45. Glucose, BMI, and age are the most significant predictors.

IV. METHEDOLOGY

A. Data Analysis

The correlation heatmap reveals key insights about the relationships between various features in the dataset as shown in figure. 2 below. Strong Correlations: Glucose, BMI, and Age show a strong positive correlation with the Outcome variable, indicating their significant role in diabetes prediction.

- Weak Correlations: Insulin and Skin Thickness exhibit weaker correlations, which may be influenced by missing or zero values in the data.
- Pregnancies and Age Relationship: Pregnancies and Age show a moderate correlation, suggesting that older individuals tend to have more pregnancies.

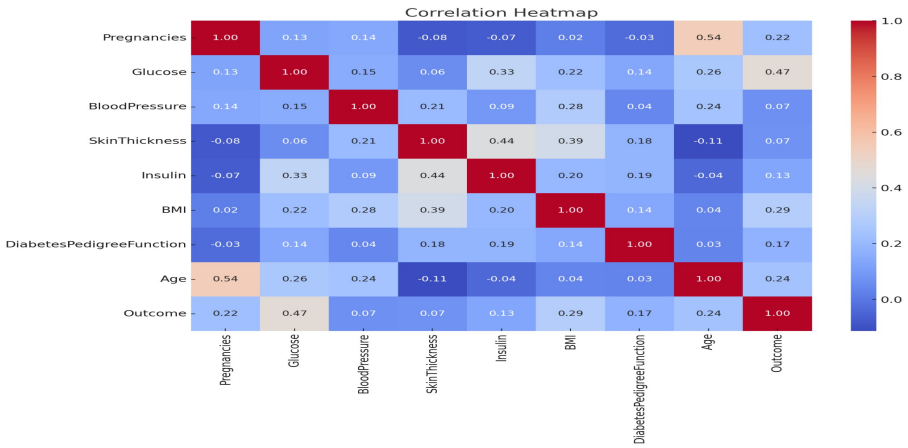

Figure 2 Correlation Heatmap for Diabetes Prediction Features

Pair Plot Observations

- Distinct clusters are noticeable for diabetic and non-diabetic patients in Glucose, BMI, and Age features, supporting their predictive importance.
- Features like SkinThickness and Insulin show less clear separation, reinforcing their lower correlation with diabetes outcomes.

These insights highlight the importance of addressing data quality issues, particularly zero values, to improve model performance.

B. Data Preprocessing

The data prepossessing workflow starts with acquiring The Pima Indians Diabetes Database taken from the National Institute of Diabetes and Digestive [14], containing 768 instances with eight relevant attributes. After uploading the dataset, missing and zero values are identified and replaced with median values to maintain data integrity. A Pearson correlation matrix reveals key predictors such as glucose levels (0.47), age, and pregnancies (0.54). Data visualization aids in understanding feature interactions before segmenting the dataset into independent features along with target variable. The data is then divided into training (80%), testing (20%) sets to ensure reliable model evaluation.

C. Machine Learning Model Selection and Training

Seven ML models were employed: Support Vector Classifier (SVC), Logistic Regression, Decision Tree (DT), AdaBoost, K - Nearest Neighbor (KNN), Random Forest, and Extreme Gradient Boosting (XGB).

1) Logistic Regression

Logistic regression is used here for classification [12]. Here we want to predict the probability of outcome. It predicts the output of a categorical dependent variable. Therefore, the outcome can be either 0 or 1[13], Yes or no, True or false but instead of getting exact value 0 or 1, we get the value between 0 and 1. Here we fit a S shaped logistic function which predicts the maximum value. We have used this model here to predict the diabetes of a person.

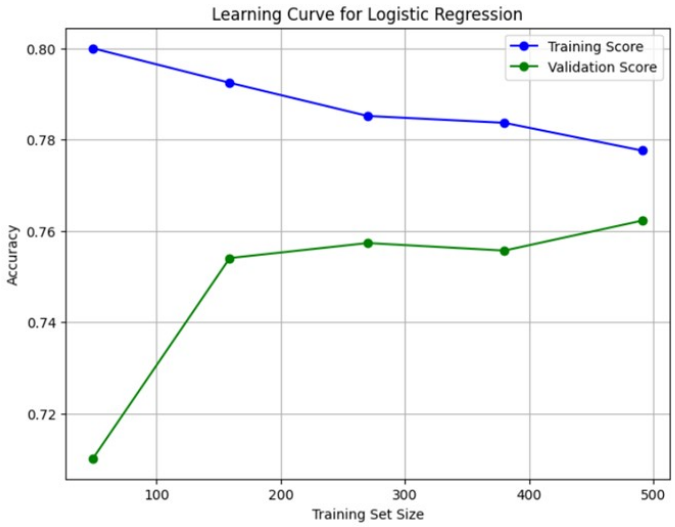

Figure 3: Learning curve for Logistic Regression

The figure [3] shows the learning curve for logistic regression in which the blue line shows the training score and the green line is used to show the validation score. The Training score starts with 80% or 0.8 which means the overfitting and slowly decreasing (which is good) as the training set increases. This is normal as the model is learning from these patterns instead of memorizing them. The validation score starts lower and increases gradually which indicates that the model is improving as more data gets. There is minimal gap between the training data and validation score which indicates the model is balanced and overfitting is minimum. Overall, this model is stable and validation performance is good.

2) Decision Tree

Decision tree basically shows the different options of choices for a solution of a problem. It also shows how different factors are related to each other. It is like a binary tree starting with a root node then divided into two nodes and the same for each node till leaf nodes. It is very highly interpretable and can handle both categorical and numerical data.[14]

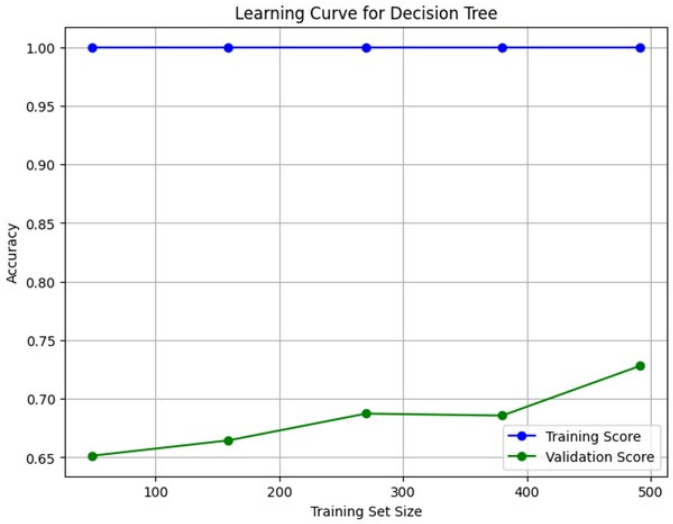

Figure 4: Learning curve for Decision Tree

Figure [4] shows the learning curve for the Decision Tree in which the blue line shows the training score and green line is used to show the validation score. The Training score starts with perfect accuracy with 1 which indicates the overfitting. The model is trying to memorize the training data instead of generalizing the training data. The validation score starts lower around 0.65 and increases gradually to 0.75 with a gap between validation score and training score. As there is high variance in the model suggests that there is slow improvement in validation accuracy.

3) Support Vector Machine

Support Vector Machine is a supervised machine learning algorithm used for regression and classification tasks. It aims to find the optimal hyperplane in an N-dimensional space to separate data points into different classes. SVM algorithm is to find the hyperplane that best separates two classes by maximizing the margin between them.

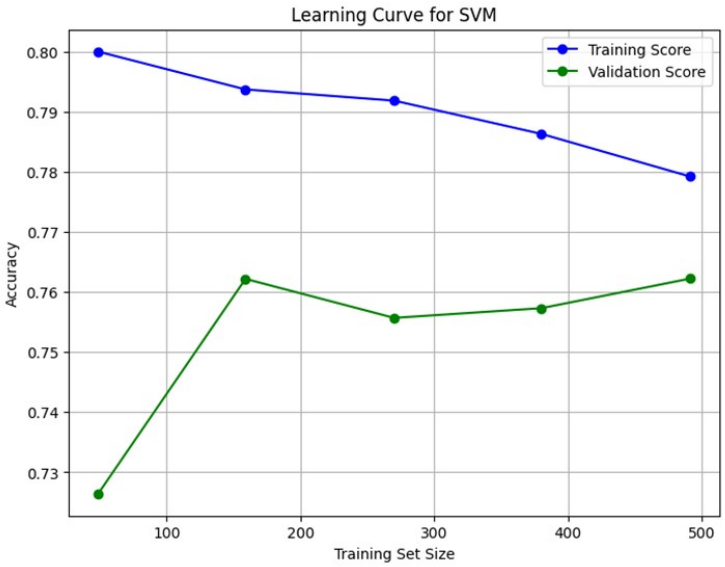

Figure 5: Learning curve for Support Vector Machine

The figure [5] shows the learning curve for Support Vector Machine in which the blue line shows the training score and the green line is used to show the validation score. The Training score starts with perfect accuracy with 0.8 and gradually decreases to 0.78 as the training set increases. The decline in training score indicates the Support Vector Machine is trying to learn more generalized patterns and avoiding the overfitting. The validation score starts at 0.73, rises and stables at 0.76, the model is underfitted as the training curve is above the validation. The model is stable but needs some improvement.

4) K-nearest Neighbor

This algorithm does not learn from the training set immediately but stores the data and performs action at the time of classification, that's why it is known as lazy learner algorithm. In KNN there is K which is just a number that tells the algorithm how many neighbors to check while making a decision. Deciding the value of K is very critical as very high value can lead to overfitting.

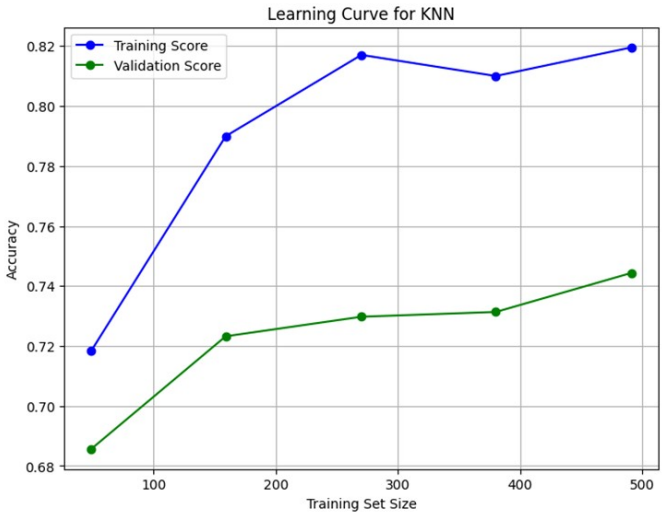

Figure 6: Learning curve for K-Nearest Neighbor

The figure [6] shows the learning curve for K-Nearest Neighbor in which the blue line shows the training score and the green line is used to show the validation score. The Training score starts at 0.72 rises to 0.82 and stables there after moving in range of 0.8-0.82. The high training accuracy in training data signifies that the model can fit the data well. The validation score starts at 0.68 then gradually rises and stables at 0.74. The KNN model suggests strong training performance, but the learning curve indicates overfitting.

5) Random Forest

Random Forest is a very powerful machine learning technique used to make predictions and uses voting to make predictions. This technique is used for classification and regression tasks. It also handles the missing data, ranks the features, scales well with large and complex data.

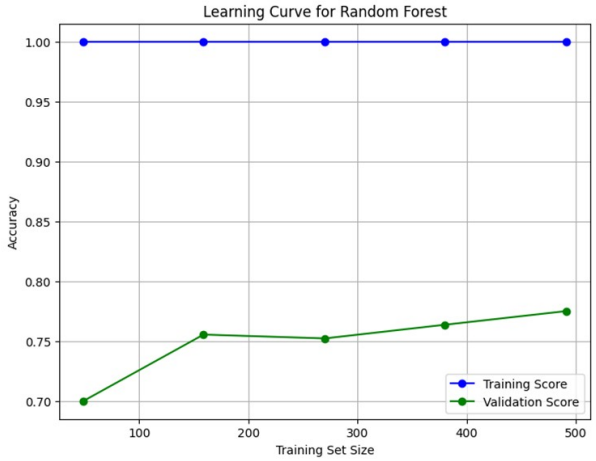

Figure 7: Learning curve for Random Forest

The figure [7] illustrates the learning curve for the Random Forest model. In this graph, the blue line represents the training score, while the green line denotes the validation score [15]. The training score starts at 1 and consistently remains at this value, indicating that the model has perfectly memorized the training data. Meanwhile, the validation score begins at approximately 0.7, gradually improving and stabilizing around 0.75 to 0.80 (approximately 0.77) as the amount of training data increases. The noticeable gap between the training and validation scores suggests that the model is experiencing overfitting.

6) XGBoost

eXtreme Gradient Boosting also known as XG Boost is an advanced machine learning algorithm used for high speed, efficiency, performance. Traditional ML algorithms like Decision Tree and Random Forest struggle with complex datasets so we use XGBoost for that. It starts with a base learner, calculates the error, trains the next tree, repeats the process and combines the all predictions.

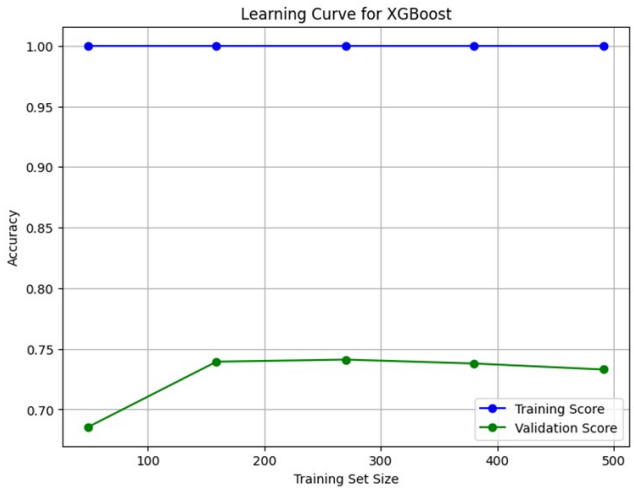

Figure 8: Learning curve for XG Boost

The figure [8] shows the learning curve for XG Boost in which the blue line is ushows the training score and green line is used to show the validation score. The Training score starts at 1 and remains consistent at 1 which indicates a perfect fit for the model. This is the strong indication of overfitting in the model as the model fails to generalize the data instead trying to memorize the data. The validation score starts below 0.7then gradually rises near 0.75 and after declining stables at 0.74 as the training data increases. The model might be too complex as adding more data does not improve the generalization.

7) AdaBoost

AdaBoost is a technique that enhances classification performance by combining multiple decision trees. It strengthens model accuracy, reduces the risk of overfitting, effectively manages imbalanced data, and offers improved interpretability.

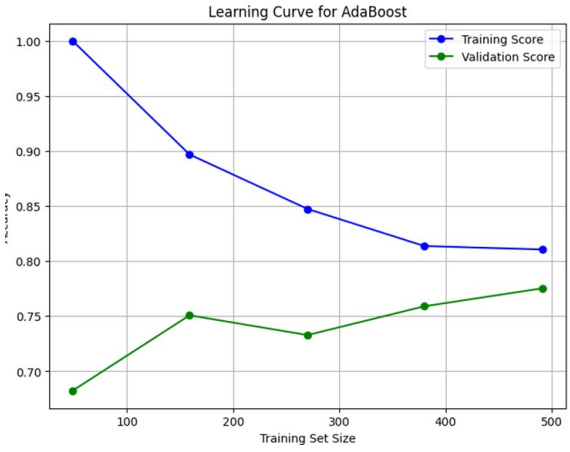

Figure 9: Learning curve for AdaBoost

The figure [9] shows the learning curve for AdaBoost in which the blue line shows the training score and the green line is used to show the validation score. The Training score starts at 1 and drops quickly to 0.8. It indicates that the model is generalizing the data well. The validation score starts below 0.7then gradually rises

near 0.75 then after declining again increases and stables near 0.77 which indicates that model's generalizing performance is improving as the data size increases.

V. MODEL EVALUATION METRICS

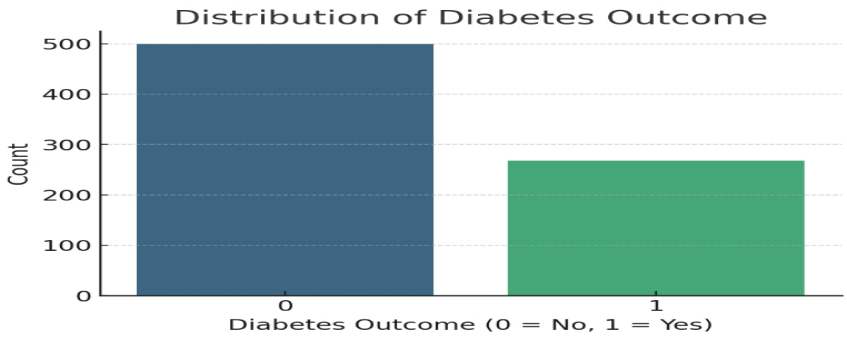

Figure 10. Distribution of Diabetes Outcome

The bar chart in figure [10] displays the distribution of diabetes outcomes within data-set that contain "0" is non-diabetic individuals and "1" is diabetic individuals. The graph reveals a noticeable class imbalance, as the count of non-diabetic cases (~500) significantly surpasses the diabetic cases (~268). This imbalance could affect the performance of predictive models, leading to a bias towards the majority class (non-diabetic).

Addressing this class imbalance is necessary when building ML models, as it ensures the model does not favor predicting the majority class. Techniques like resampling, synthetic data generation (e.g., SMOTE), or applying class weights in the model could help achieve a balanced and accurate predictive model.

The machine learning algorithm used in this work has produced some outstanding results which are here demonstrated by using the confusion matrix, Classification Report and the Receiver's Operating Characteristic (ROC) Curve [16][17].

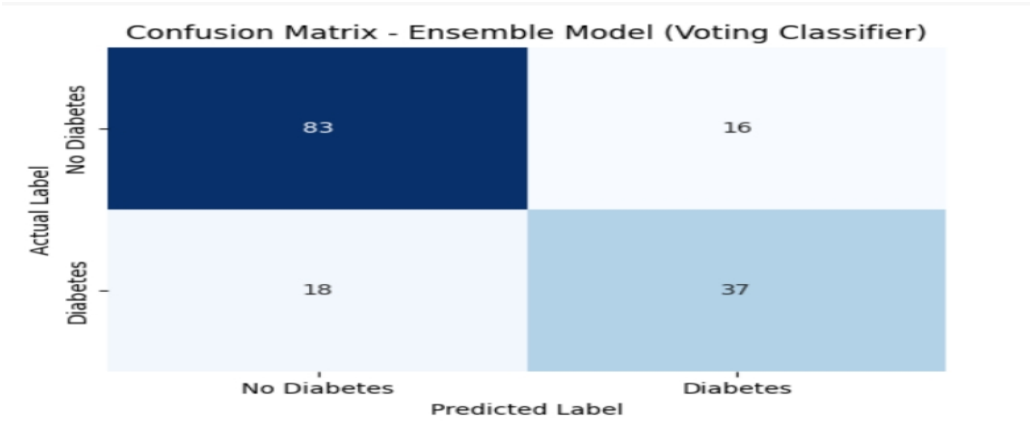

Figure 11. Confusion Matrix

Above confusion matrix in figure [12] indicates the accuracy of model which is 77.9 that is good performance as most of the predictions are correct.

```
Ensemble Model Accuracy: 0.7857

Confusion Matrix:
[[81 18]
 [15 40]]

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.82      0.83        99
           1       0.69      0.73      0.71        55

    accuracy                           0.79       154
   macro avg       0.77      0.77      0.77       154
weighted avg       0.79      0.79      0.79       154
```

Figure 12. Classification Report

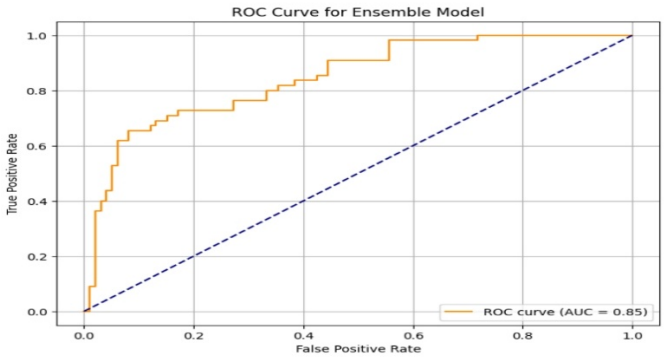

Figure 13. ROC Curve

The ROC curve figure [13] is well above the diagonal and the area under curve is 0.85.

VI. CONCLUSION

Machine learning is being widely used in number of tasks and domains [20]. This research presents a approach to diabetes diagnosis by using the machine learning models. Disease which affects thousands of people annually by making the toll that forecasts the likelihood of diabetes using the inputs given by the user. This study employed seven machine learning models suck as random forest, k-nearest neighbor, logistic regression, decision tree, XGBoost, support vector machine, Adaboost, an ensemble model which uses a voting classifier (using 5 models of 7 models) to achieve the highest accuracy. It classifies diabetes based on a publicly available dataset (Pima Indian Diabetes Database). Among all models, the ensemble model demonstrated the performance, by achieving maximum accuracy and sensitivity in both training and testing phases.

REFERENCES

[1] Ojo, O. A., Ibrahim, H. S., Rotimi, D. E., Ogunlakin, A. D., & Ojo, A. B. (2023). *Diabetes Mellitus: Molecular Mechanisms, Pathophysiology, and Pharmacological Perspectives*. Medical Novel Technologies and Development, 19, Article 100247.
[2] Terzo, S., Amato, A., & Mulè, F. (2021). *From Obesity to Alzheimer's Disease Through Insulin Resistance*. Journal of Diabetes and its Complications, 35(11), 108026.
[3] Chang, M. (2007). *Hepatitis B Virus Infection and Its Impact on Neonatal Health*. Seminars in Fetal and Neonatal Medicine.
[4] Central Department of Statistics. (2003). *Annual Report on Health and Demographics*. Riyadh, Saudi Arabia.
[5] Spencer, L. (2023). *Enhancing System Security Through Fault-Tolerant OS Design*. University of Derby (United Kingdom), ProQuest Dissertations & Theses, 30723528.

[6] Helmy, E., Elnakib, A., & ElNakieb, Y. (2023). *The Role of Artificial Intelligence in Autism Diagnosis Using DTI and fMRI: A Comprehensive Survey*. Biomedicines, 11(7).
[7] Zheng, Y., Ley, S. H., & Hu, F. B. (2018). *Global Causes and Epidemiology of Type 2 Diabetes Mellitus and Its Complications*. Nature Reviews Endocrinology, 14(2), 88-98.
[8] Barbedo, J. G. A. (2019). *Machine Learning in Agriculture: Detecting Nutrition Deficiencies Using Proximal Images*. Computers and Electronics in Agriculture, 162, 482-492. https://doi.org/10.1016/j.compag.2019.04.037
[9] Exploiting Machine Learning Techniques in Human Resource Management:January 26, 2025 A Descriptive Research Mohammed Fadhl Abdullah 1 , Nabil Mohammed Ali Munassar 1 , Ryad A. Gbr 1 1 IT Department, Faculty of Engineering and Computing, University of Science and Technology, Aden, Yemen *Corresponding author: n.munassar@ust.edu
[10] Cai, Y., Chen, R., Gao, S., Li, W., Liu, Y., Su, G., et al. (2023). *Artificial Intelligence in Neoantigen Identification for Personalized Cancer Immunotherapy*. Frontiers in Oncology, 12, 1054231. doi: 10.3389/fonc.2022.1054231
[11] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., & Ogurtsova, K. (2019). *Global and Regional Diabetes Prevalence Estimates for 2019 and Workions for 2030 and 2045*. Diabetes Research and Clinical Practice, 157, 107843.
[12] Baader, G., & Krcmar, H. (2018). *Reducing False Positives in Fraud Detection: Combining Red Flag Approach with Process Mining*. International Journal of Accounting Information Systems, 31, 1-16.
[13] Tiong, L. C. O., & Lee, H. J. J. (2021). *E-Cheating Prevention Measures: Deep Learning Approach for Detecting Online Examination Fraud*. Journal of Latex Class Files.
[14] C. S. Lee and M. H. Wang, ―Ontology-based intelligent healthcare agent and its application to respiratory waveform recognition,‖ Expert Syst.Appl., vol. 33, no. 3, pp. 606–619, Oct. 2007
[15] Baker, M. J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H. J., Dorling, K. M., Fielden, P. R., Fogarty, S. W., Fullwood, N. J., Heys, K. A., Hughes, C., Lasch, P., Martin-Hirsch, P. L., Obinaju, B., Sockalingum, G. D., Sulé-Suso, J., Strong, R. J., Walsh, M. J., Wood, B. R., Gardner, P., & Martin, F. L. (2014). *Analyzing Biological Materials Using Fourier Transform IR Spectroscopy*. Nature Protocols, 9(8), 1771-1791. DOI: 10.1038/nprot.2014.110
[16] Malik, Nidhi and Srivastava, Yash and Aggarwal, Priyansh and Sharma, Dev and Mehra, Ujjwal, Identifying User Intent in Social Media Comments using BERT (July 26, 2024). Proceedings of the International Conference on Innovative Computing & Communication (ICICC 2024), Available at SSRN: https://ssrn.com/abstract=4906463
[17] R. Saxena, K. Jindal, N. Malik and A. Bhatia, "Air-Quality Index Prediction Using Auto Ml Library, TPOT," *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2023, pp. 1-4, doi: 10.1109/ICCCNT56998.2023.10307166.

[18] Morshed, I. B., Harmon, B., Zaman, M. S., Rahman, M. J., Afroz, S., & Rahman, M. (2017). *Inkjet Printed Fully-Passive Body-Worn Wireless Sensors for Smart and Connected Community (SCC)*. Journal of Low Power Electronics and Applications, 7, 26.
[19] Elhag, S., Fernández, A., Bawakid, A., Alshomrani, S., & Herrera, F. (2015). *Combining Genetic Fuzzy Systems and Pairwise Learning to Improve Intrusion Detection Rates*. Expert Systems with Applications, 42, 193-202.
[20] Malik, N., Jain, S. (2020). Comparative Study of Machine Learning Algorithms for Social Media Text Analysis. In: Batra, U., Roy, N., Panda, B. (eds) Data Science and Analytics. REDSET 2019. Communications in Computer and Information Science, vol 1230. Springer, Singapore. https://doi.org/10.1007/978-981-15-5830-6_19