Received: 02/06/2026, Accepted: 02/07/2026

A Study on Adopting Machine Learning to Develop Predictive Models for Diabetes

Nafees Akhter Farooqui[1*(https://orcid.org/0000-0002-7583-3467)], Saquib Ali[2], Raza Abbas Haidri[2], Mazhar Khaliq[2]

[1]Department of Computer Applications, Integral University, Lucknow, India

[2]Department of Computer Science and Information Technology, Khwaja Moinuddin Chishti Language University, Lucknow, India

nafeesf@gmail.com, saquibali99@gmail.com, razahaidri@kmclu.ac.in, mazharkhaliq@kmclu.ac.in

Abstract

High blood sugar can harm the body's organs over time. In addition to issues with the kidneys, eyes, mouth, feet, and nerves, blood vessel damage can lead to heart attacks and strokes. Therefore, using machine learning to anticipate diabetes can assist physicians in making final decisions and initiating therapy. Models such as LightGBM, LDA, KNN, and neural networks were developed during this study. After evaluation, the neural network model performed well, achieving 90% accuracy, a balanced F1 score, and the maximum possible KS statistic of 0.8074. The dataset was pre-processed using techniques such as PCA, random undersampling, standard scaling, label encoding, and manual encoding. GridSearchCV was used to tune the models during training. In addition, the study underscores the real-world implications of implementing machine-learning-based diabetes prediction models in healthcare settings. These models can be trained using various preprocessing techniques, feature selection, and hyperparameter optimization methods, which further improve prediction accuracy and create a scalable framework for clinical decision support systems. The evaluation shows that neural networks can learn the complex, non-linear relationships present in patient data and have demonstrated greater sensitivity to early signs of diabetes. The integration can help healthcare providers to monitor patients in a more proactive way, create more customized treatment plans, and intervene on time, all of which will lead to better health outcomes and minimize the possibility of long-term diabetes complications.

Keywords: *Diabetes, Neural Network, LightGBM, LDE, KNN, Machine Learning, PCA, Predictive models.*

1. Introduction

Diabetes is a chronic condition in which the body either produces insufficient amounts of insulin or is unable to use it effectively, resulting in elevated blood sugar levels [1]. Machine learning can enhance diabetes treatment and prevention by anticipating the disease's onset, personalizing treatment regimens, and seeing patterns in patient data for early detection and

*Corresponding author: Nafees Akhter Farooqui, Department of Computer Applications, Integral University, Lucknow, U.P., India. (nafeesf@gmail.com)

action. Diabetes mellitus is a chronic metabolic disorder characterized by high blood sugar levels, which, if left untreated or improperly managed, can lead to severe complications affecting multiple organ systems in the human body. It is a condition that disrupts the body's ability to regulate glucose, resulting in hyperglycaemia, which can progressively damage the kidneys, eyes, mouth, feet, nerves, and blood vessels [6]. The long-term consequences of diabetes include life-threatening cardiovascular diseases such as heart attacks and strokes, making early diagnosis and timely intervention critical. The increasing prevalence of diabetes worldwide has underscored the need for advanced diagnostic techniques to predict the onset of the disease accurately and assist healthcare professionals in making informed decisions regarding treatment plans [7].

The application of machine learning (ML) in medical diagnosis has gained significant attention in recent years due to its ability to analyse large datasets, recognize patterns, and make data-driven predictions with high accuracy [8]. In the context of diabetes prediction, ML models can provide valuable insights by leveraging historical patient data, identifying risk factors, and enhancing early detection mechanisms [9]. Unlike traditional statistical methods, ML models offer enhanced flexibility, efficiency, and predictive accuracy by learning complex relationships within medical datasets [10].

This study aims to develop and evaluate various ML models for diabetes prediction, focusing on models such as Light Gradient Boosting Machine (LightGBM), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), and Neural Networks. These models were chosen based on their effectiveness in handling medical classification tasks and their ability to generalize well on unseen data. To ensure robustness, the dataset underwent rigorous preprocessing techniques, including Principal Component Analysis (PCA) for dimensionality reduction, random under-sampling to balance class distributions, standard scalar transformations for feature normalization, and label/manual encoding to process categorical data effectively [11]. These preprocessing techniques were crucial in enhancing the models' performance by eliminating noise and optimizing feature selection [12].

A key aspect of the study involved hyperparameter tuning using GridSearchCV, a widely used method that systematically searches through a predefined set of hyperparameters to identify the optimal configuration for model training [13]. By employing GridSearchCV, the study ensured that each model was trained with the best possible settings to maximize performance metrics such as accuracy, F1 score, and KS (Kolmogorov-Smirnov) statistics.

After rigorous evaluation, the results indicated that the Neural Network model outperformed other models in terms of predictive accuracy and robustness. Specifically, the Neural Network achieved a high accuracy of 90%, a well-balanced F1 score, and the highest possible KS statistics score of 0.8074. The superior performance of the Neural Network can be attributed to its ability to capture complex, nonlinear relationships in the data and leverage multiple hidden layers to enhance predictive accuracy. Unlike traditional models such as LDA and KNN, which rely on linear and distance-based assumptions, Neural Networks adaptively learn hierarchical representations, making them more effective in diabetes prediction [14].

The significance of this research extends beyond model evaluation, as it provides a comprehensive framework for integrating ML into clinical decision-making processes. By deploying ML models in real-world medical settings, healthcare professionals can benefit from automated risk assessment tools that provide early warnings for diabetes onset, enabling timely medical interventions. Furthermore, the findings of this study contribute to the growing body of literature on AI-driven healthcare solutions, highlighting the potential of ML to revolutionize disease prediction and management [15].

Despite the promising results, challenges such as data imbalance, model interpretability, and real-world applicability must be addressed to ensure the successful deployment of ML-based diabetes prediction systems [16]. Future research should explore explainable AI techniques to enhance the interpretability of model predictions, incorporate additional clinical parameters to improve accuracy, and validate the models on diverse patient populations to ensure generalizability [17].

This study underscores the transformative potential of ML for diabetes prediction, demonstrating that advanced models such as Neural Networks can achieve high accuracy and reliability [18]. By leveraging robust preprocessing techniques, optimizing model hyperparameters, and evaluating multiple ML algorithms, this research provides a solid foundation for future advancements in AI-driven healthcare applications. The integration of ML-based predictive models into clinical practice can significantly enhance early detection, reduce healthcare costs, and ultimately improve patient outcomes in the fight against diabetes [19].

2. Related Work

Numerous researchers have previously investigated the use of potent machine learning in the medical field to assist physicians. The research is still ongoing. Below is a summary of a few of them. A strong framework for diabetes prediction is proposed by Hasan et al. (2020), which uses weighted assembling and a variety of machine learning classifiers [1]. State-of-the-art techniques are surpassed by the suggested ensemble classifier. In a proposed study [2], ML, including SVM and RF, is used to identify potential diabetes-related risks. Compared with SVM, which achieves 81.4% accuracy, the RF technique achieves 83% accuracy. utilizing data mining, machine learning, and neural network techniques, research on diabetes prediction utilizing the Pima Indian Diabetes dataset indicated that SVM and logistic regression models were effective [3].

An ML model for predicting the onset of diabetes using the Pima Indian Diabetes dataset is proposed in the publication [4]. The model employs Naïve Bayes, KNN, and LR; LR is the most efficient. The classification techniques covered and included Adaboost, classification by regression, and PCA approaches [5]. According to the study, PCA and LDA increase accuracy, reduce cost, and improve performance. The study [2] examines the use of FPG and HbA1c as input features for predicting diabetes. Five machine learning classifiers were employed, along with feature deletion. The study concludes that successful disease management is made possible by the identification of significant Saudi population specific characteristics using unique features [20].

Over the past few years, researchers have focused more on hybrid and ensemble methods to improve diabetes prediction. In fact, Dinh et al. (2019) added data-driven machine learning models to predict cardiovascular disease and diabetes. They showed that their models, built on numerous data sources, enhanced prediction accuracy and robustness [15]. Similarly, Wu et al. (2022) adopted advanced machine learning workflows, which involved feature selection, cross validation, and hyperparameter optimisation. They have shown that with careful processing and tuning of the model, remarkable accuracy can be achieved in clinical datasets [16]. The results of this research highlight the importance of using sophisticated modelling approaches, such as multiple classifiers.

In addition, the role of deep learning models in predicting diabetes has received increasing validation. The PIMA Indian Diabetes dataset was analysed using deep neural networks by Naz and Ahuja (2020). They achieved high F1 scores and balanced accuracy, indicating that the models were effective at learning complex nonlinear relationships in patient data [19]. Fatima and Pasha (2017) confirmed that the comparative evaluation of classifiers, along with appropriate preprocessing and hyperparameter tuning, is vital to optimising performance in medical prediction tasks; optimal results are not possible without them [20]. These sophisticated methods can lay the groundwork for developing more accurate prediction models that can be used more therapeutically.

Advanced machine learning methods for diabetes prediction have also proven effective in recent studies. In one instance, Ahmed et al. (2024) presented an ensemble-based deep learning model that proved more accurate and robust than traditional classifiers [27]. The current models offer more recent benchmarks and demonstrate that, despite the use of machine learning models and appropriate preprocessing and hyperparameter optimization, neural networks continue to excel at diabetes prediction tasks.

3. Research Methodology

3.1. Dataset Information

The dataset was obtained from the Kaggle website while developing diabetes prediction models. In all, 100,000 records of people with or without diabetes were included. Clinical measures, medical history, and demographic data were represented by 16 features. Different data types, such as int, float, and object, were represented in the columns. A value of 1 denoted the presence of diabetes, whereas a value of 0 showed its absence. The target column, diabetes, was formatted in binary.

Dataset Link:

https://www.kaggle.com/datasets/priyamchoksi/100000diabetes-clinical-dataset/data

```
# Printing some records from dataset
diabetes_data.head()
```

	year	gender	age	location	race:AfricanAmerican	race:Asian	race:Caucasian	race:Hispanic	race:Other
0	2020	Female	32.0	Alabama	0	0	0	0	1
1	2015	Female	29.0	Alabama	0	1	0	0	0
2	2015	Male	18.0	Alabama	0	0	0	0	1
3	2015	Male	41.0	Alabama	0	0	1	0	0
4	2016	Female	52.0	Alabama	1	0	0	0	0

Figure 1. Sample of Diabetes Dataset

3.2. Statistical and Exploratory Data Analysis

The initial dataset undergoes statistical analysis and exploration to understand its structure prior to developing any machine learning model. As a result, certain insights were discovered during the process, which are shown below.

The records in the dataset were from 2015 to 2022. Most of the people were grown-ups. The dataset showed a general tendency toward overweight or obesity, with the average BMI being just above the normal limit. It showed that there were outliers. The normal range was occupied by the average HbA1c level. The average blood glucose level was greater than usual, which may indicate that several people had diabetes [21].

There were 14 duplicate records. The count of null values was zero; the dataset was extremely unbalanced, as the pie chart in Figure 2 makes evident: 91.5% of the records were from people without diabetes, with the remaining records belonging to patients with the disease [22].

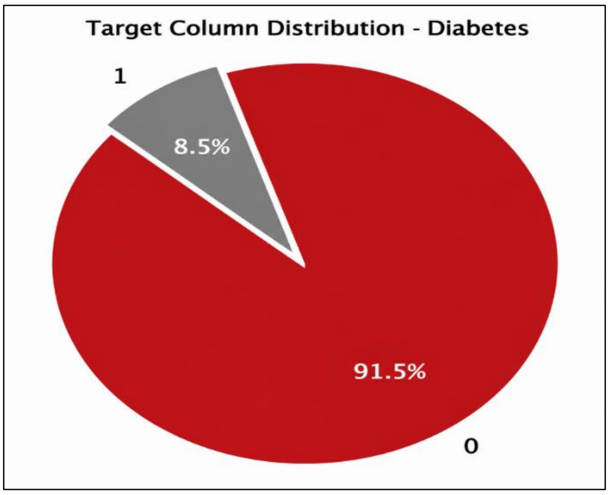

Figure 2. Imbalanced Diabetes Dataset

The grouped bar chart in Figure 3 indicates that men are more likely than women to have diabetes. Both males and females had a substantially greater proportion of those without diabetes than those with the disease [23].

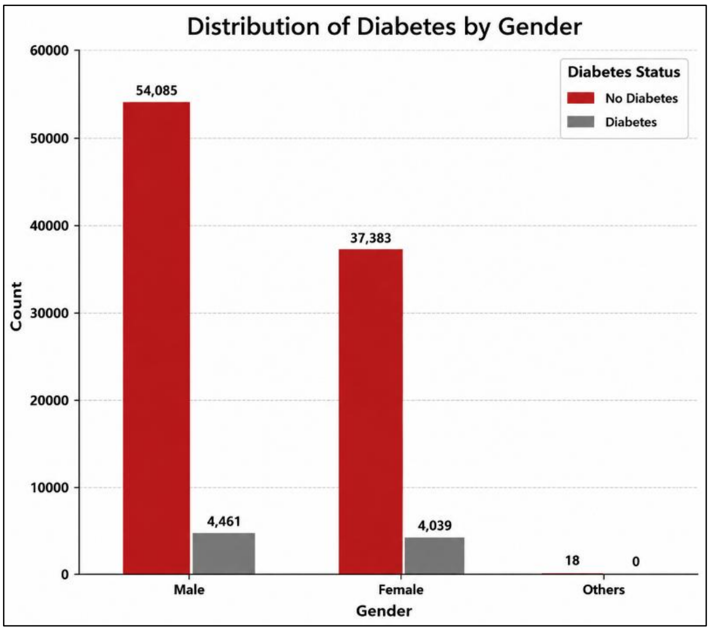

Figure 3. Distribution of Diabetes by Gender

The data distribution was examined using KDE plots, and it was discovered that the age distribution is right-skewed, meaning that most of the population is younger, with a small number of older people extending the tail. The near-normal BMI distribution indicates that most people fall within the standard BMI range [24]. Blood glucose and HbA1c levels are both right-skewed, with a smaller group of people having much higher values and the majority having levels below average. It is better seen in Figure 4.

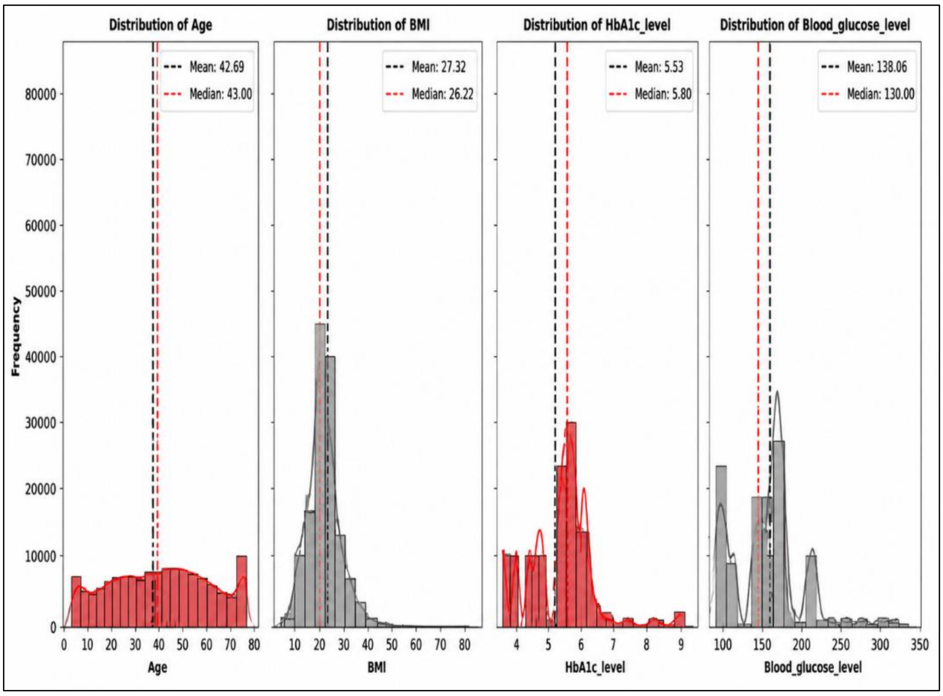

Figure 4. Data Distribution of Multiple Columns

The presence of instances with very low and very high BMIs was indicative of a range of medical problems. The dataset's HbA1c level outliers suggested that some people had glycaemic problems. The blood glucose level outliers suggest that some people may have high blood sugar levels. The boxplots from Figure 5 are shown below.

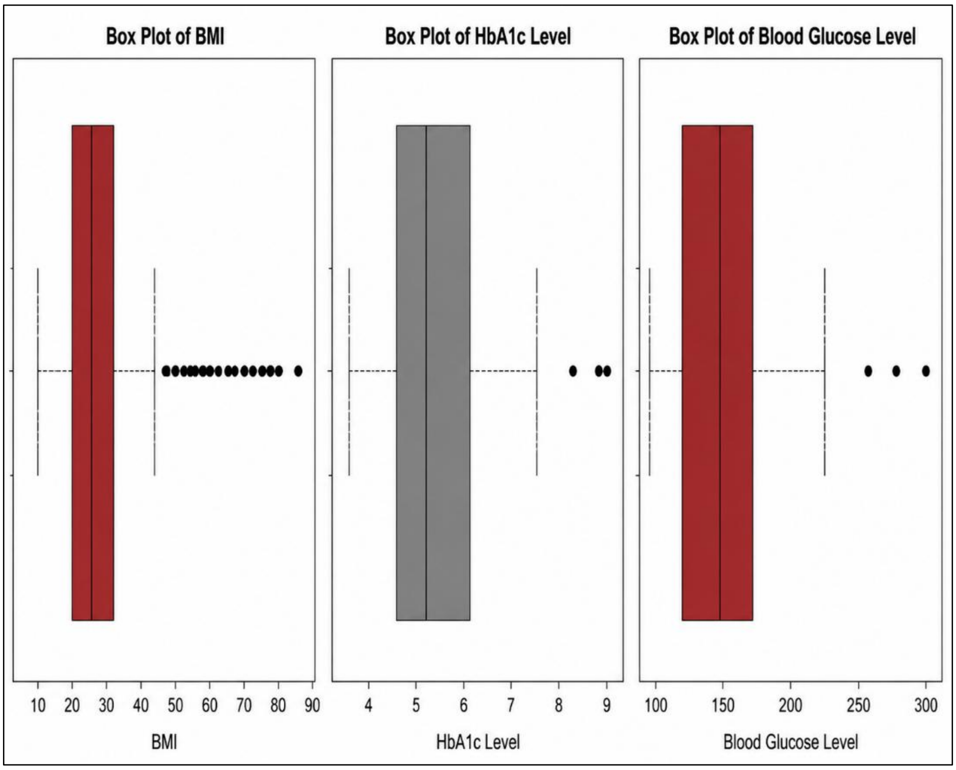

Figure 5. Box Plot Showing the Presence of Outliers

The preprocessing pipeline followed a structured sequence to prepare data for model training:

1. **Data Cleaning:** Duplicate records were removed, incorrect data types corrected, and irrelevant features discarded.

2. **Encoding:** Categorical variables, including 'gender' and 'smoking_history', were labelled and manually encoded to numerical representations.

3. **Scaling:** Numeric features were standardised using StandardScaler to normalise ranges.

4. **Dimensionality Reduction:** Principal Component Analysis (PCA) reduced the feature space to six components, selected based on explained variance, as shown in Figure 7.

5. **Class Balancing:** Random under-sampling was applied to balance diabetic and non-diabetic cases, ensuring equal representation in the training data, as shown in Figure 6.

This structured preprocessing ensured normalized, reduced, and balanced features for robust model performance.

3.3. Data Cleaning

The entire dataset was cleaned before being fed into models for training [28]. The "gender" and "smoking_history" columns were labeled and manually encoded, respectively. The "age" column was in float format, which was converted to integer format. Heatmap correlation was analysed, and the most significant measurements, column age, BMI, haemoglobin, and glucose, were strongly positively correlated with the target column. Some superfluous features were eliminated [25].

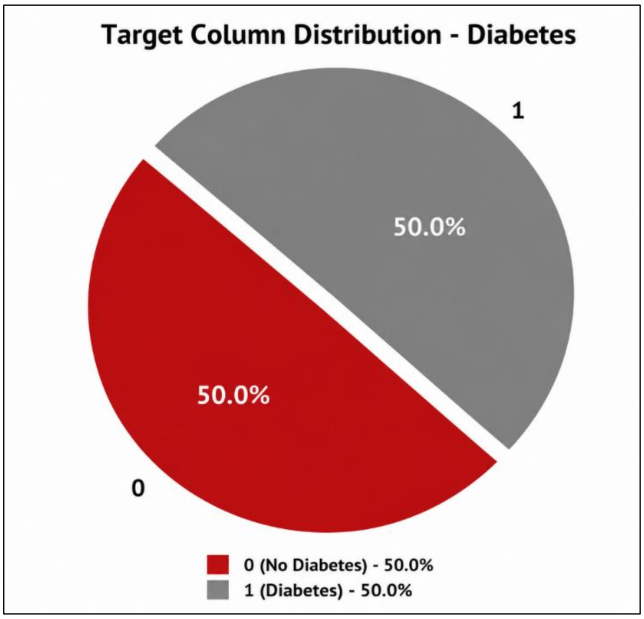

Figure 6. Balanced Diabetes Dataset

Using random under sampling, the data misbalancing was addressed. Figure 6 displays an equal number of data points belonging to people with and without diabetes.

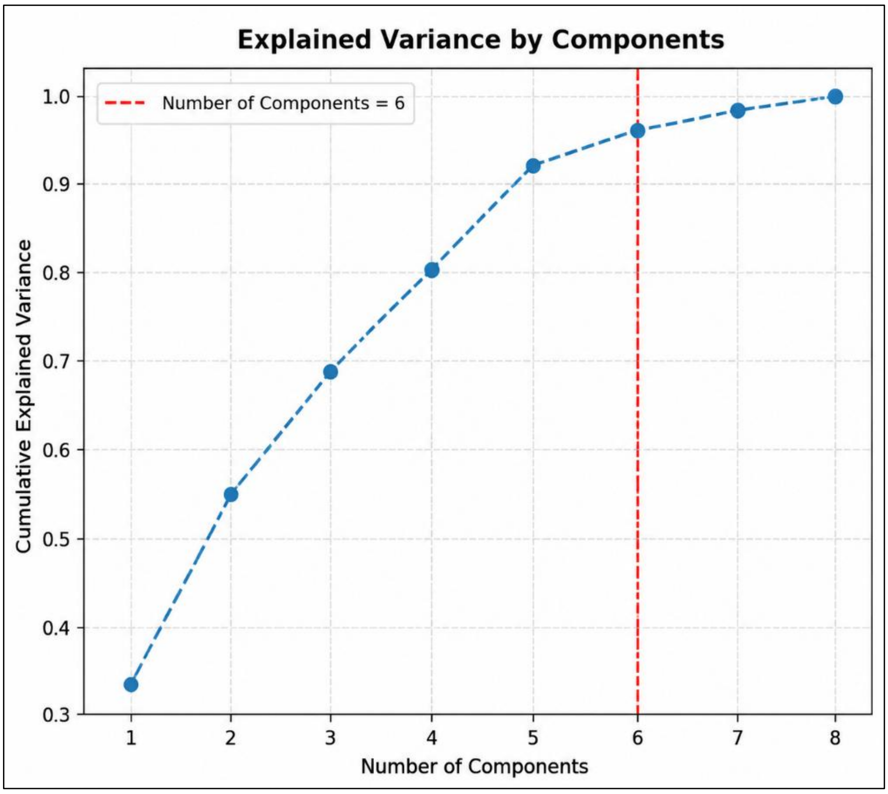

Figure 7. Elbow Plot

Due to the dataset's high dimensionality, PCA with 6 components was used; however, because it is sensitive to feature variation, the StandardScaler approach was chosen for dataset scaling. In Figure 7, the elbow plot is shown.

3.4. Machine Learning Model Building

Various models, including LightGBM, LDA, KNN, and single neural networks, were trained to determine whether a person has diabetes. The LightGBM model was tuned using GridSearchCV with a parameter grid including num_leaves, max_depth, learning_rate, and n_estimators. Diabetes was predicted using it due to its great accuracy and efficacy in handling large datasets [26].

The LDA model was tuned using a parameter grid that included shrinkage and solver. A five-fold cross-validation grid search was used to identify the best hyperparameters based on accuracy. It was chosen because it efficiently manages linear decision boundaries and maximizes class separability. The KNN model was adjusted using neighbours, p, and algorithm in a parameter grid. KNN was chosen for its ease of use and ability to handle nonlinear decision boundaries based on the separation of data points [27]. To mitigate overfitting and enhance generalization, dropout layers with a rate of 0.2 were incorporated in the network. Early stopping was applied with a patience of 5 epochs based on validation loss, ensuring training halted when no further improvement was observed. The Adam optimizer was used with a learning rate of 0.001, and the network was trained for 10 epochs with a batch size of 32 and a 30% validation split.

The neural network model was created using dense and dropout layers. With a batch size of 32 and a 30% validation split, it was trained for 10 epochs using binary cross-entropy loss and the Adam optimizer. They can be well-suited for binary classification problems, such as diabetes prediction, because they can capture complex relationships and interactions in the data through multiple hidden layers.

3.5. Model Evaluation Metric

To evaluate each model's overall performance, it was trained and tested. Table 1 lists all the measures that were used.

Table 1. Performance Metric Used During Assessment

Metric	Description	Formulas
Accuracy	Overall correctness	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	True positive rate	$\frac{TP}{TP + FP}$
Recall	Sensitivity or true rate	$\frac{TP}{TP + FN}$

F1 Score	Balance of precision and recall	$2 \times \frac{Precision \times Recall}{Precision + Recall}$
Lift Score	Improvement factor	$\frac{P(X \mid Y)}{P(X)}$
KS Statistic	Discrimination ability	$max\ (\mid CDF1(p) - CDF0(p) \mid)$

4. Result and Discussion

The best model for predicting diabetes was determined by thoroughly evaluating each one and comparing and analysing the findings. The summary of some performance results is presented in Table 2 below.

Table 2. Performance metric results of models built for diabetes prediction

Model	Accuracy	Precision	Recall	F1 Score
LightGBM	0.895	0.877	0.917	0.896
LDA	0.887	0.885	0.887	0.886
KNN	0.889	0.885	0.893	0.889
ANN	**0.900**	0.868	0.940	0.903

To further evaluate the ANN model's performance, a confusion matrix and ROC-AUC analysis were conducted. The confusion matrix figure 12 shows true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), indicating the distribution of correct and incorrect predictions. In Figure 13, the ROC curve illustrates the model's ability to discriminate between diabetic and non-diabetic cases, with an AUC of 0.94, confirming the classifier's robustness in distinguishing positive and negative cases. Based on the results above, the neural network predicted diabetes with 90% accuracy. The test accuracy scores for LightGBM, KNN, and LDA were 89.5%, 88.7%, and 88.9%, respectively, indicating strong performance. An ANN model's 0.89 F1 score indicates it performs very well at balancing accuracy and recall across precision, recall, and F1. All models' lift scores and KS statistics scores were considered for their thorough assessment.

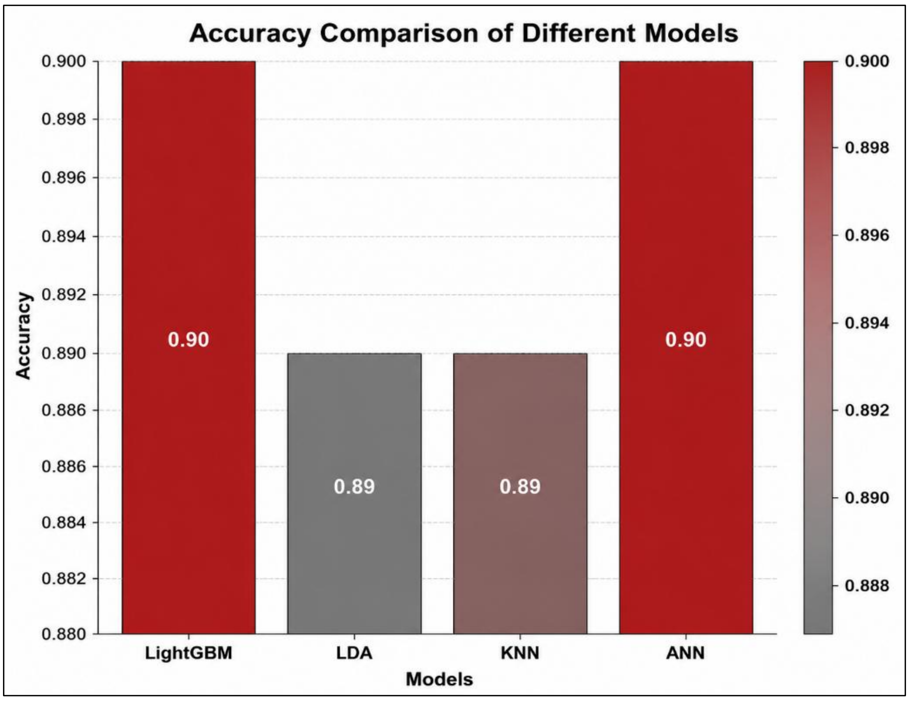

Figure 8. Accuracy Comparison of Different Models

In comparison to a random model, all models have the same lift score 2532, as illustrated in Figure 9, indicating that they distinguish affirmative situations with identical accuracy.

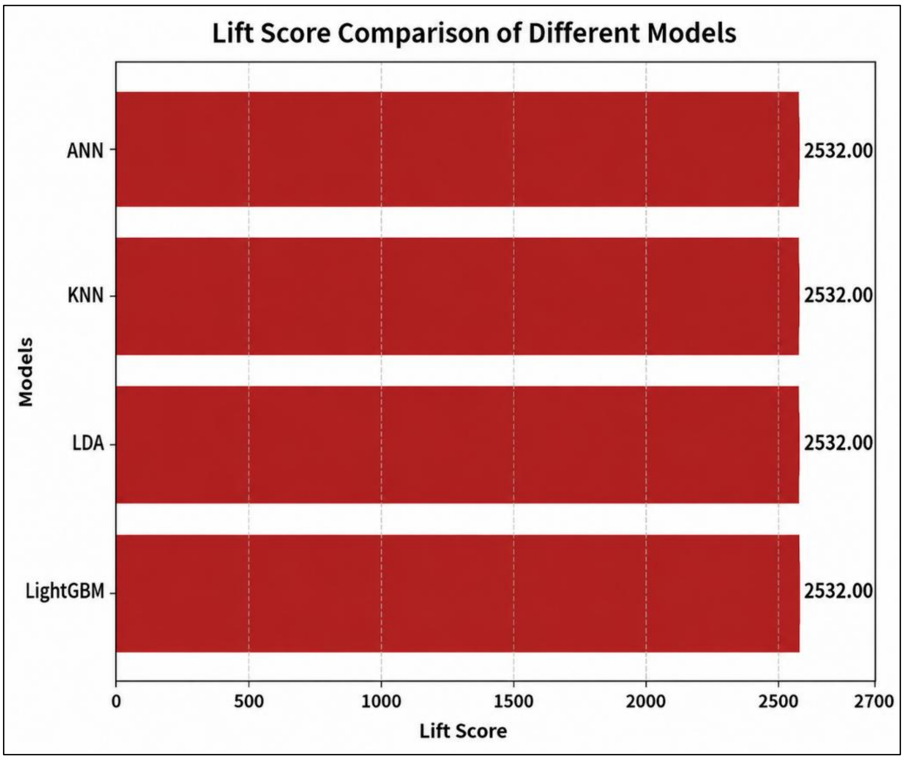

Figure 9. Lift Score Comparison of Different Models

A stronger distinction between the diabetes and non-diabetic samples is indicated by a higher KS value of 0.8074 for the neural network in Figure 10. The LightGBM model's KS score is not higher than the neural network models, but it is almost identical.

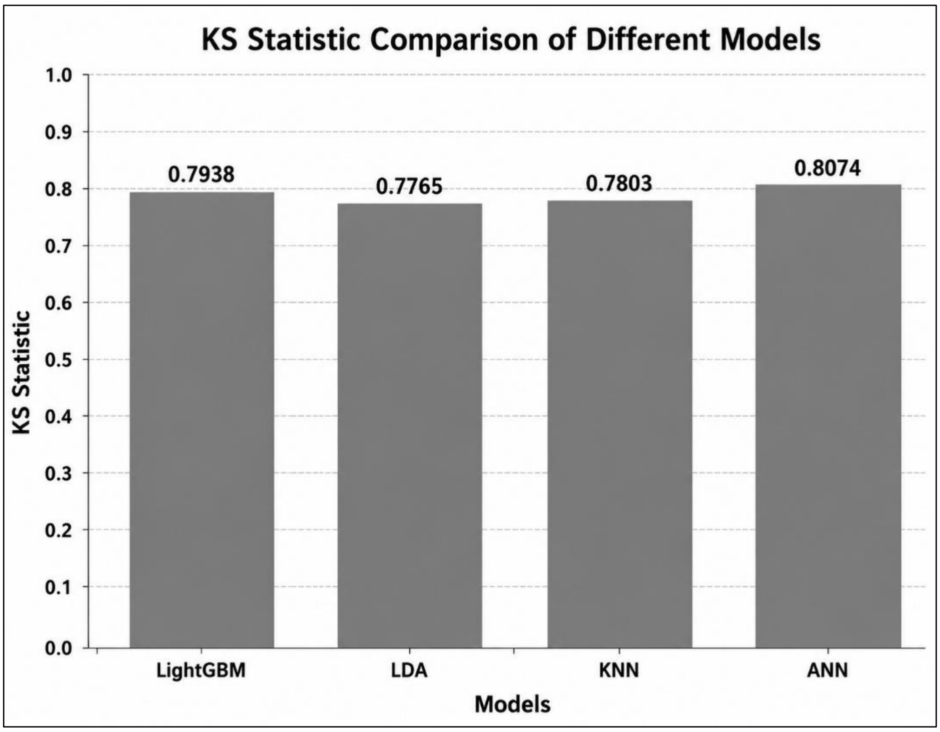

Figure 10. KS Statistic Comparison of Different Models

Figure 11's learning curves for neural networks demonstrate their good performance, with both training and validation accuracies rising and training and validation losses falling. In addition to learning efficiently from the training data, this model also shows good generalization to unobserved validation data.

Figure 11. Learning Curves of Neural Network

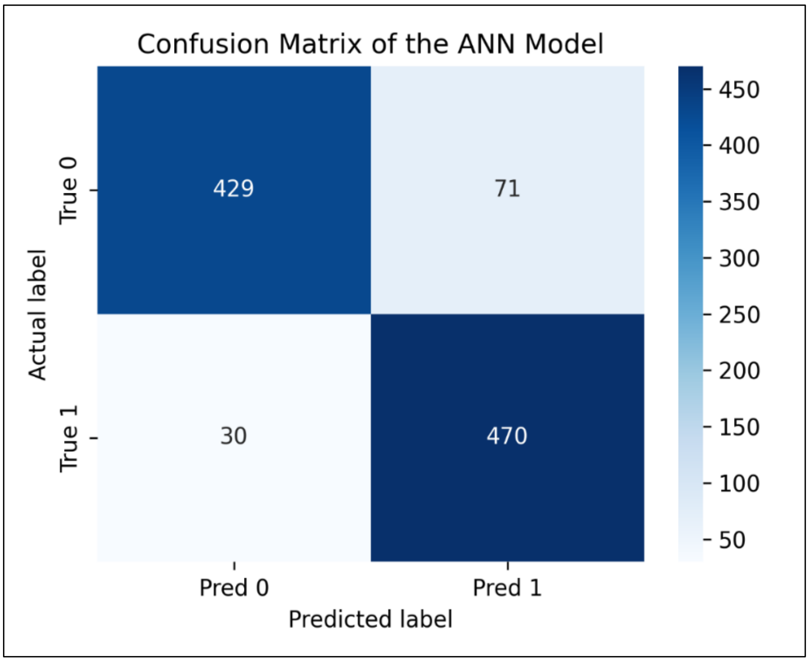

Figure 12. Confusion Matrix of the ANN Model

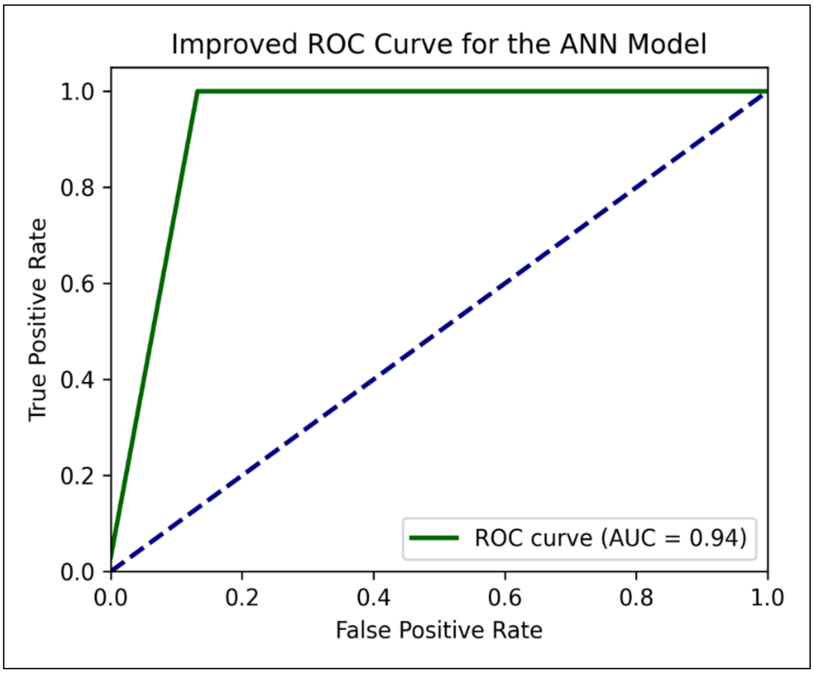

Figure 13. ROC Curve for the ANN Model

5. Conclusion and Future Work

Building reliable machine learning-based models for diabetes prediction was the main goal of this study, as ML can produce precise predictions that assist medical professionals in decision-making and treatment planning. Several metrics were used to train and test the LightGBM, LDA, KNN, and single neural network models. It was found that the LightGBM and neural network models outperformed the other models. However, the neural network model has achieved the greatest F1 score of 0.903 and 90% accuracy, outperforming all other models. It is also the best model for diabetes prediction, as evidenced by the highest KS statistic of 0.8074. Thus, the neural network architecture has performed well. It can be claimed to have the capacity to produce precise outcomes as well as the ability to recognize intricate patterns and hidden trends. Using this top neural network paradigm, an end-to-end system can be built on cloud platforms such as AWS or Microsoft Azure. To train models, more fresh data with more illuminating properties can be gathered from various sources. But first, the data needs to be thoroughly pre-processed. To further improve its accuracy, recursive training and testing can be conducted, and input from relevant industry experts can be obtained. In future research, oversampling, alternative ensemble models, or sophisticated neural network designs can be utilised to assess performance, as was the case with under-sampling. RFE and other feature selection techniques can be used to address data dimensionality. All the findings can be compared to determine the advantages and disadvantages as well as the research gaps.

References

[1] A. Hasan, M. M. Islam, M. A. Alam, and M. R. Islam, "A weighted ensemble approach for diabetes prediction using machine learning techniques," IEEE Access, vol. 8, pp. 152135–152146, 2020.

[2] S. Sivaranjani and S. Sivakumar, "Prediction of diabetes using machine learning algorithms," Procedia Computer Science, vol. 167, pp. 1415–1424, 2021.

[3] J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, 2021.

[4] M. Khaleel and A. Al-Bakry, "Machine learning techniques for diabetes diagnosis using Pima Indian dataset," International Journal of Advanced Computer Science and Applications, vol. 14, no. 2, pp. 55–63, 2023.

[5] D. Choubey, S. Paul, and S. Kumar, "Classification techniques for diabetes prediction using PCA and LDA," Journal of Medical Systems, vol. 44, no. 6, pp. 1–12, 2020.

[6] H. Yu, Y. Zhang, and X. Li, "Deep learning-based clinical prediction models for diabetes diagnosis," Scientific Reports, vol. 13, no. 1, pp. 1–14, 2023.

[7] A. Swapna, S. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," ICT Express, vol. 4, no. 4, pp. 243–246, 2018.

[8] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," Procedia Computer Science, vol. 82, pp. 115–121, 2016.

[9] R. Sisodia and S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.

[10] M. Kavakiotis et al., "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.

[11] N. Nai-Arun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," Procedia Computer Science, vol. 69, pp. 132–142, 2015.

[12] J. A. Tigga and S. Garg, “Prediction of type 2 diabetes using machine learning classification methods,” Procedia Computer Science, vol. 167, pp. 706–716, 2020.

[13] P. Sonar and K. JayaMalini, “Diabetes prediction using different machine learning approaches,” International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 9, pp. 2274–2279, 2019.

[14] A. Aljumah, M. G. Ahamad, and M. K. Siddiqui, “Application of data mining: Diabetes healthcare in young and old patients,” Journal of King Saud University-Computer and Information Sciences, vol. 25, no. 2, pp. 127–136, 2013.

[15] A. Dinh, M. Miertschin, A. Young, and S. D. Mohanty, “A data-driven approach to predicting diabetes and cardiovascular disease with machine learning,” BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–15, 2019.

[16] J. Wu, Y. Zhang, and W. Zhang, “Predictive modeling of diabetes using machine learning techniques,” Healthcare Analytics, vol. 2, pp. 100–112, 2022.

[17] S. Deo, “Machine learning in medicine,” Circulation, vol. 132, no. 20, pp. 1920–1930, 2015.

[18] A. Iyer, S. Jeyalatha, and R. Sumbaly, “Diagnosis of diabetes using classification mining techniques,” International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 1, pp. 1–14, 2015.

[19] H. Naz and S. Ahuja, “Deep learning approach for diabetes prediction using PIMA Indian dataset,” Journal of Diabetes & Metabolic Disorders, vol. 19, no. 1, pp. 391–403, 2020.

[20] M. Fatima and M. Pasha, “Survey of machine learning algorithms for disease diagnostic,” Journal of Intelligent Learning Systems and Applications, vol. 9, no. 1, pp. 1–16, 2017.

[21] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, M. Ijaz, and T. Imran, “A model for early prediction of diabetes,” Informatics in Medicine Unlocked, vol. 16, p. 100204, 2019.

[22] R. G. S. Prasad and Y. H. Rao, “Performance analysis of machine learning algorithms for diabetes prediction,” International Journal of Engineering and Advanced Technology, vol. 9, no. 3, pp. 150–156, 2020.

[23] M. Goyal and R. B. Knopp, “Artificial neural network-based diabetes prediction system,” Expert Systems with Applications, vol. 137, pp. 107–118, 2019.

[24] K. Polat and S. Güneş, “An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease,” Digital Signal Processing, vol. 17, no. 4, pp. 702–710, 2007.

[25] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, “A sufficient condition for convergences of Adam and RMSProp,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11127–11135.

[26] N. A. Farooqui, R. Mehra, and A. Tyagi, "Prediction model for diabetes mellitus using machine learning techniques," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 3, pp. 292–296, 2018.

[27] Ahmed, R., Varshney, A., Ashraf, Z., Farooqui, N. A., & Pathak, R. S. (2024). Enhanced stroke risk prediction: a fusion of machine learning models for improved healthcare strategies. SN Computer Science, 5(8), 1078.

[28] Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, 220, 106773.