IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14820

Analysis of Machine Learning and Deep Learning Methods for Early Diabetes Prediction

Ms. B. MADHUVANTHI[1], Dr. T.S. BASKARAN[2]

Research Scholar, PG & Research Department of Computer Science, A. V. V. M. Sri Pushpam College (Autonomous), Poondi-613503, Thanjavur, "Affiliated to Bharathidasan University, Tiruchirappalli - 620024", TamilNadu, India.[1]

Associate Professor& Research Supervisor, PG & Research Department of Computer Science, PG & Research Department of Computer Science, A. V. V. M. Sri Pushpam College (Autonomous), Poondi-613503, Thanjavur, "Affiliated to Bharathidasan University, Tiruchirappalli - 620024", TamilNadu, India.[2]

Abstract: Diabetes, commonly known as diabetes mellitus, is a condition that affects how the body processes blood sugar. It occurs when the pancreas either cannot produce enough insulin or the body is unable to effectively use the insulin that is produced. Insulin, a hormone secreted by the pancreas, facilitates the transport of glucose from food into cells, where it is used for energy. Uncontrolled diabetes often leads to hyperglycemia (high blood sugar), which, along with other health complications, can significantly damage nerves and blood vessels. According to 2014 statistics, a substantial number of individuals aged 18 and older had diabetes, and in 2019, diabetes alone was responsible for 1.5 million deaths.

However, with the rapid advancement of machine learning (ML) and deep learning (DL) classification algorithms, early detection of diabetes has become significantly more feasible across various fields, including healthcare. In this study, we conducted a comparative analysis of multiple ML and DL techniques for early diabetes prediction. We utilized a diabetes dataset from the UCI repository, comprising 17 attributes, including the target class, and evaluated the performance of all proposed algorithms using a range of performance metrics. Our experiments indicated that the XGBoost classifier outperformed all other algorithms, achieving nearly 100% accuracy, while the remaining models demonstrated accuracy levels exceeding 90%.

Keywords: Diabetes prediction; XGBoost; KNN; CNN; LSTM; Classification.

I. INTRODUCTION

Diabetes is a serious medical condition characterized by elevated blood sugar levels. It develops when the body cannot effectively utilize insulin, a hormone produced by specialized pancreatic cells called islets [1]. Diabetes is a leading cause of heart disease, stroke, amputation, kidney failure, blindness, and premature death [2]. Its prevalence is increasing due to multiple factors, including excessive salt intake, unhealthy diet, overweight or obesity, immune system disorders, insulin resistance, stress, and genetic predisposition [3]. Currently, approximately 463 million people aged 20 to 79 are affected by diabetes, and this number is projected to rise to 700 million by 2045 [4]. Diabetes caused 4.2 million deaths, with type 1 diabetes affecting over 1.1 million children and adolescents, while 374 million people are at elevated risk of developing type 2 diabetes. The most common types of diabetes include type 1, type 2, and gestational diabetes.

Although preventing diabetes is challenging, proper management and early diagnosis can significantly reduce risk. Modern medical science offers various methods for early detection, and machine learning (ML) has proven to be highly effective in identifying diabetes at initial stages. Researchers have reported promising results using ML techniques such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Logistic Regression, and XGBoost. The rapid development of deep learning (DL) techniques, combined with the availability of large datasets, has further enhanced the ability to predict diabetes-related outcomes, including diagnosis, glucose level management, and complication evaluation [5].

While accuracy is often used to assess the performance of prediction models, other metrics such as precision, recall, F1-

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14820

score, execution time, and ROC value are essential for determining the most effective algorithm. Early-stage detection can enable timely intervention and appropriate treatment for diabetic patients. To classify early-stage diabetes as either positive or negative, we implemented and evaluated several classification algorithms, including XGBoost, Decision Tree, Random Forest, SVM, Multi-Layer Perceptrons (MLP), and Logistic Regression. We used a benchmark UCI dataset [6] containing 16 attributes and 520 instances, splitting the data into 416 training and 104 testing samples. Performance was compared across multiple evaluation metrics, including precision, recall, F1-score, execution time, and ROC value. The analysis demonstrated that the XGBoost classifier achieved the best performance, with approximately 99.99% training accuracy and 99.0% testing accuracy.

The primary research objectives of this study include:
1. Analyzing publicly available datasets and their relevance in diabetes research.
2. Performing a comprehensive comparison of ML and DL techniques for early-stage diabetes detection.
3. Evaluating performance metrics to determine the most effective predictive algorithms.
4. Identifying future research directions for advancing diabetes prediction studies.

The remainder of this paper is organized as follows: Section II presents a brief review of related literature. Section III outlines the research methodology. Experimental results are discussed in Section IV. Finally, conclusions and directions for future work are provided in Section V.

II. LITERATURE REVIEW

In this section, we highlight several researchers who have made significant contributions to the prediction of diabetes mellitus by leveraging public medical datasets through machine learning (ML) and deep learning (DL) approaches.
For example, Lin et al. [7] evaluated Naive Bayes, SVM, and ANN classifiers using a diabetes dataset. They conducted a weight-adjusted voting-based analysis and concluded that combining multiple models resulted in higher classification accuracy than any single model. Kandhasamy et al. [8] proposed a predictive analytics model using J48 (C4.5), K-Nearest Neighbors (KNN), Random Forest, and SVM classifiers. Initially, J48 achieved the highest accuracy of 73.82% before data preprocessing, whereas KNN and Random Forest outperformed other algorithms after preprocessing. Agrawal et al. [9] developed a hybrid model for diabetes risk prediction using CNN, KNN, SVM, SVM+LDA, Naive Bayes, ID3, C4.5, and CART algorithms on a dataset of 738 patients, achieving a maximum accuracy of 88.10% by combining SVM and LDA.

Naiarun et al. [10] developed a web-based application for diabetes prediction and compared multiple ML and DL algorithms, including Random Forest, Decision Tree, Logistic Regression, CNN, and Naive Bayes, along with bagging and boosting methods. Their results showed that Random Forest outperformed others with an accuracy of 85.55% and an ROC value of 0.912. Kavakiotis et al. [11] evaluated Logistic Regression, Naive Bayes, and SVM using 10-fold cross-validation, with SVM achieving the highest accuracy of 84%. Similarly, Zheng et al. [12] applied several ML algorithms—including Random Forest, KNN, SVM, Naive Bayes, Decision Tree, and Logistic Regression—using 10-fold cross-validation to predict diabetes at an early stage, identifying areas for improvement in feature selection.

Ahmed et al. [13] developed a J48 Decision Tree model for managing type 2 diabetes, utilizing seven specific patient attributes (age, gender, renal problems, smoking, hypertension, cardiac problems, and diabetes status). Their model achieved an accuracy of 70.80% and an ROC value of 0.624. Oleiwi et al. [14] proposed a novel ML-based classification model for early-stage diabetes prediction, emphasizing the use of significant features to achieve results aligned with clinical outcomes. They trained Random Forest, Multi-Layer Perceptron (MLP), and Radial Basis Function (RBF) classifiers, with RBF achieving the highest accuracy of 98.80%. Bukhari et al. [15] employed an Artificial Neural Network (ANN) with varying numbers of neurons in hidden layers (5–50) to predict diabetes in females using the PIMA Indians dataset [16], achieving 93% accuracy on the validation set.

Most of these studies improved prediction performance by employing multiple ML and DL classification techniques. Common evaluation metrics include accuracy, precision, recall, F1-score, ROC score, and execution time. To conclude, our research presents a comparison of real diagnostic medical datasets using prominent ML classification algorithms for early-stage diabetes mellitus prediction, considering several key risk factors.

III. METHODOLOGY

Our proposed methodology consists of four main components, as illustrated in Figure 1, which collectively achieve the research objectives. First, we collected the diabetes dataset and performed data pre-processing to ensure quality and consistency. Next, the dataset was split into training and testing sets using a tenfold cross-validation approach. The

DOI: 10.17148/IJARCCE.2025.14820

selected machine learning and deep learning algorithms were then applied to the training set to predict early-stage diabetes mellitus. Finally, the performance of these algorithms was evaluated on the test set using a variety of evaluation metrics. Each of these phases is briefly discussed in the following sections.

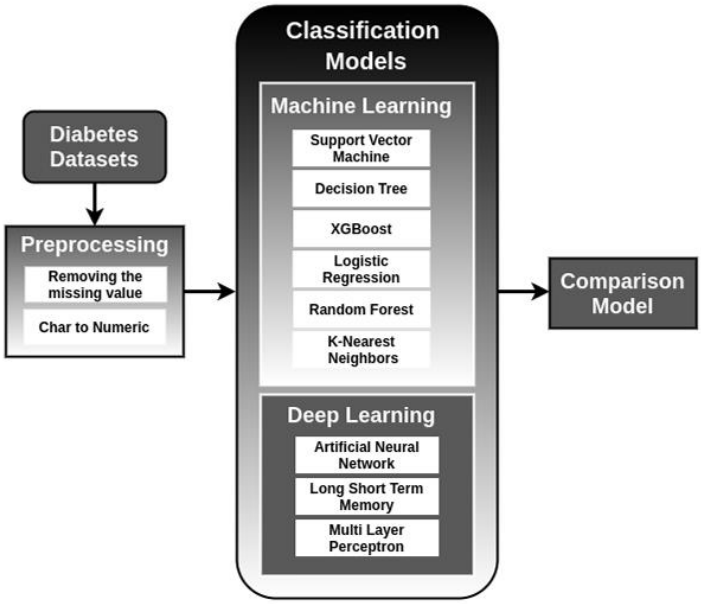

Figure 1 Graphical representation of proposed methodology

A. Dataset and Attributes

2. In this study, we used the diabetes dataset from the UCI repository [6] to evaluate the effectiveness of machine learning and deep learning algorithms for early-stage diabetes diagnosis. The dataset was collected via a structured questionnaire administered to 520 patients at Sylhet Diabetes Hospital in Bangladesh, including individuals recently diagnosed with diabetes or exhibiting diabetes-related symptoms. It contains 16 attributes, with 320 positive and 200 negative instances, where the positive and negative labels indicate whether a patient is at risk of diabetes. The attributes and their possible values are as follows:

- Age: 20 to 65 years
- Sex: 1 represents Male and 0 represents Female
- Polyuria: 1 means Yes and 0 means No
- Sudden Weight loss : 1 and 0 represent Yes and No respectively
- Weakness : 1 represents Yes and 0 represents No
- Polyphagia : 1 and 0 means Yes and No
- Genital thrush : 1 equal to Yes and 0 equal to No
- Visual blurring : 1 and 0 means Yes and NO
- Itching : 1 and 0 represents Yes and No respectively
- Irritability : 1 means Yes and 0 means No
- Delayed Healing: 1 denotes yes and 0 denotes No
- Partial Paresis : 1 and 0 denotes Yes and No
- Muscle Stiffness : Yes and No represents 1 and 0 respec tively
- Alopecia: 1 means Yes, 0 means No
- Obesity : 1 means Yes and 0 means No
- Class : 1 represents Positive and 0 represents Negative

A. Pre-Processing

Data pre-processing was performed to prepare the diabetes dataset for effective analysis and to address missing or inconsistent values. Nominal attribute values, which are not directly suitable for machine learning or deep learning algorithms, were converted into numeric form. For example, in the sex attribute, male was encoded as 1 and female as 0. Similarly, for other categorical attributes, “Yes” was converted to 1 and “No” to 0, where 1 represents a positive value and 0 represents a negative value in the class label.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14820

C. Classifier Algorithms

I. Machine Learning Classifier:

a. **XGBoost:** XGBoost (XGB) is an efficient and powerful implementation of the Gradient Boosted Trees algorithm. It is a robust distributed machine learning platform designed to scale tree boosting algorithms effectively. In a distributed environment, XGBoost supports fast parallel tree construction, is highly configurable, and offers fault tolerance. It can handle datasets ranging from a single node with tens of millions of samples to billions of samples distributed across multiple systems [17].

$$Y = \sum_{i=1}^{n} (w_i . x_i) \qquad (1)$$

b. **Random Forest:** Random Forest (RF) is a versatile classifier used for both regression and classification tasks. It consists of an ensemble of decision trees, each trained on randomly sampled, independently and identically distributed vectors, with the final prediction determined by majority voting across the trees [18]. RF is a simple yet flexible machine learning algorithm that generally delivers strong performance even without extensive hyperparameter tuning.

c. **Decision Tree:** A Decision Tree (DT) is a supervised learning method primarily used for classification tasks. It is also effective for identifying relevant features and patterns within large datasets, making it a powerful tool for predictive modeling and discrimination [19]. Each node in a decision tree—either a decision node or a leaf node—performs a test on an attribute, creating branches that split the data based on the test outcome.

d. **Support Vector Machine:** Support Vector Machine (SVM) is a supervised machine learning model that is effective for capturing complex relationships between data points while performing classification. SVM is particularly well-suited for text classification tasks, as it can achieve strong performance even with a limited number of labeled samples, and it has also demonstrated promising results in various biological applications [20].

e. **K-Nearest-Neighbours:** K-Nearest Neighbors (KNN) is a supervised machine learning algorithm that is simple to implement and can be applied to both classification and regression problems. KNN classifies new instances based on their similarity to previously observed cases, assigning the new instance to the category most common among its nearest neighbors. This method is particularly useful when there is limited prior knowledge of the underlying data distribution [21].

f. **Logistic Regression:** Logistic Regression (LR), also known as a Logit Model, is a statistical method used to model binary response variables. It is a type of regression that predicts the probability of one outcome class relative to another (e.g., positive versus negative treatment outcome) by incorporating aspects of linear regression into the model [22]. The basic logistic regression model can be expressed as follows [23]:

$$\text{logit}(Y) = \text{naturallog}(\text{odds}) = \ln \frac{\pi}{\pi - 1} = \alpha + \beta X \qquad (2)$$

where the regression co-efficient β is the logit.

II. Deep Learning Classifier

a. **Multi-layer Perceptron:**

Multi-layer Perceptron (MLP) classifier is a type of feed-forward artificial neural network (FFANN) that consists of more than two layers; the f irst layer and last layer are called input layer and output layer respectively. There has also a higher one layer in the middle of the input and output layer called the hidden layer. The time complexity depends on the number of layers; if the number of the layer will increase then time complexity will also increase. For instance, every neuron accepts input like as x1, x2, x3,.....xn, and the bias and wight are denoted by (b) and (w); Input multiplied by wight and produced output which may be denoted y based on the activation function(∅) [24]

$$y = \emptyset(\sum_{i=1}^{n} ((w_i . x_i) + b)) \qquad (3)$$

DOI: 10.17148/IJARCCE.2025.14820

b. Artificial Neural Network:

Artificial neural networks (ANNs) are designed to mimic the way the human brain processes and analyzes information. Their models, created based on the brain's methods, can be used to simulate complex patterns and prediction problems. An overall architecture for building neural networks has three main components: input, hidden, and output layers. The frequency stream in feed forward networks is from internal to external components, going in a straight line [25]. We can define ANN as look like [26]

$$z = f(b + x.w) = f(b + \sum_{i=1}^{n}((x_i.w_i)) \quad (4)$$

where x d(1 n), wd(n 1), b d(1 1) as well as z d(1 1) In the above equation, b is bias, w denotes weights, the input and output node denotes x and z respectively.

c. Long Short-Term Memory(LSTM):

LSTM networks are a category of recurrent neural network that was specif ically developed to model sequences and their long-range dependencies and to accurately model sequences as a result. This model incorporates a variety of deep learning algorithms. The introduction to LSTMs has its difficulties, as well as bidirectional and sequence-to-sequence, two concepts closely related to LSTMs, which means that there is no relationship between the length of the input and the capacity of the system and then each time step, the execution time per weight is constant O(1) [27]. The input sequence x = (x1, ..., xT) and output sequence y =(y1, ..., yT) of an LSTM network by using these formulas to find the network unit activation starting from t = 1 to T: [27]

$$\begin{cases} i_t = \sigma(W(ix)x_t + \mathrm{W(im)m(t-1)} + \mathrm{W(ic)c(t-1)} + b_i) \\ f_t = \sigma(W(fx)x_t + \mathrm{W(fm)m(t-1)} + \mathrm{W(fc)c(t-1)} + b_f) \\ o_t = \sigma(W(ox)x_t + \mathrm{W(om)m(t-1)} + \mathrm{W(oc)c(t-1)} + b_o) \\ c_t = f_t \odot (c(t-1) + i_t \odot \mathrm{g(W(cx)}x_t + \mathrm{W(cm)m(t-1)} + b_c) \\ m_t = o_t \odot h(c_t) \\ y_t = \emptyset(W(ym)m_t + b_y) \end{cases} \quad (5)$$

weight matrices in which the W terms stand for W_ic, W_fc, W_oc is weight matrices have the geometric shape of a triangle with each side being the same length, and the b component is non-symmetrical. Sigma sigmoid function is represented by the σ , and I f, o and c are input gates, for example gate, exit gate and cell activation vectors, both of which have the same size as the vector m cell activation, product of the vector with elements wise denoted $\odot$, g and h are the functions for cell input and cell output, and $\emptyset$ is the function to trigger the network output.

d. Performance Evaluation Metrics

After applying cross-validation to our proposed approaches, it is essential to assess their performance using appropriate evaluation metrics. In this study, we evaluated the models using commonly employed classification metrics, including precision, recall, F1-score, ROC score, and accuracy, to measure their predictive performance [28].

Precision: The number of true positives divided by the sum of true positives and false positives is how precision is calculated:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

Recall: The number of true positives divided by the sum of true positives and false negatives is known as recall, and it is calculated as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

F1-score The geometric average of precision and recall is defined as the f1-score. Mathematically:

$$\text{F1 score} = 2\,X\frac{\text{Precision X Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

Accuracy: The number of right predictions divided by the total number of predictions stated below is used to measure accuracy:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14820

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \qquad (9)$$

ROC-AUC Score. An ROC curve, formed by expressing the relationship between the true positive rate (sensitivity or recall) and the false positive rate (false positive rate), is called the receiver operating characteristic (ROC) curve (1-specificity). The ROC-AUC measure is commonly used to evaluate the accuracy of models that assign positive and negative class labels to binary classification problems.

IV. EXPERIMENTAL RESULT

Table 1 COMPARED TO NINE MODELS TO REVEAL WHICH ONE IS THE MOST ACCURATE.

Classification Model	Train Accuracy	Test Accuracy	Train Loss	Test Loss	Precision	Recall	F1-Score	Roc-Auc Score	Execution Time
XGB	1.000	1.000	0.000	0.000	1.000	1.000	1.000	1.000	46.074
RF	0.969	0.923	1.079	2.657	1.000	0.882	0.938	0.941	95.798
DT	0.966	0.923	1.162	2.657	0.984	0.897	0.938	0.935	31.116
KNN	0.954	0.846	1.578	5.314	1.000	0.765	0.867	0.882	43.837
SVM	0.942	0.913	1.993	2.989	0.968	0.897	0.931	0.921	51.088
LR	0.947	0.894	1.827	3.653	0.983	0.853	0.913	0.913	65.865
ANN	0.950	0.885	1.744	3.985	0.983	0.838	0.905	0.905	47.197
MLP	0.913	0.885	2.989	3.985	0.924	0.897	0.910	0.879	22.376
LSTM	0.935	0.923	1.744	3.985	0.983	0.838	0.905	0.905	146.61

In this study, machine learning (ML) and deep learning (DL) algorithms were employed to develop a computer-assisted diabetes diagnosis system using the dataset. We implemented six ML models—XGBoost, Random Forest, Decision Tree, Support Vector Machine, Logistic Regression, and K-Nearest Neighbors—as well as deep learning models including LSTM, ANN, and MLP. Prior to applying these classification algorithms, all data points in the dataset were pre-processed.

Table 1 presents the performance metrics for all ML and DL models, including training accuracy, testing accuracy, training loss, testing loss, precision, recall, F1-score, ROC-AUC score, and execution time, allowing for a comprehensive comparison of model performance. Testing accuracy reflects a model's ability to predict outcomes on unseen data, whereas training accuracy indicates performance on the training dataset. Loss measures the error in predictions, with higher loss corresponding to more incorrect predictions. Training loss is calculated during model training, while test loss evaluates performance on unseen data.

Figure 2 illustrates the F1-score, ROC-AUC score, and test accuracy across nine models. Figure 3 depicts underfitting and overfitting trends, represented by the training and test loss curves. These visualizations help assess model reliability and generalization capability.

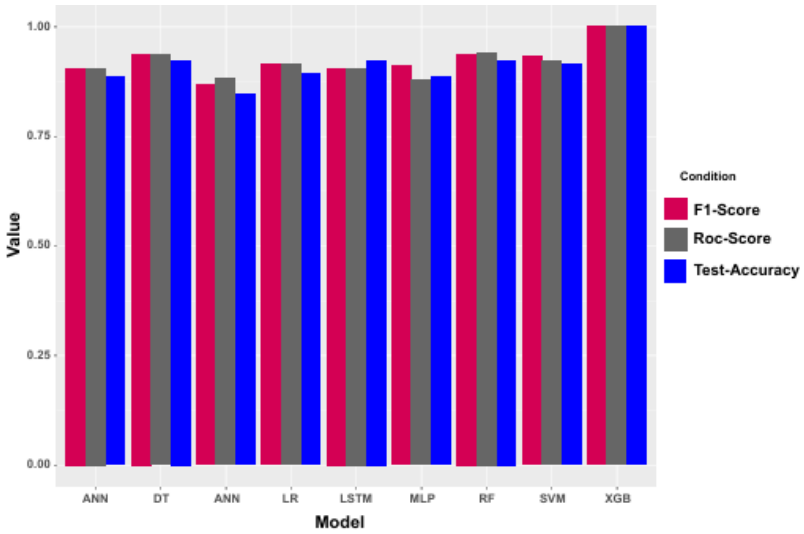

Figure 2 Bar plot that illustrated to visualize the Evaluation metrics of different models. Here, Test accuracy, F1-score and Roc-score are mapped in order to represent the comparison of different models.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14820

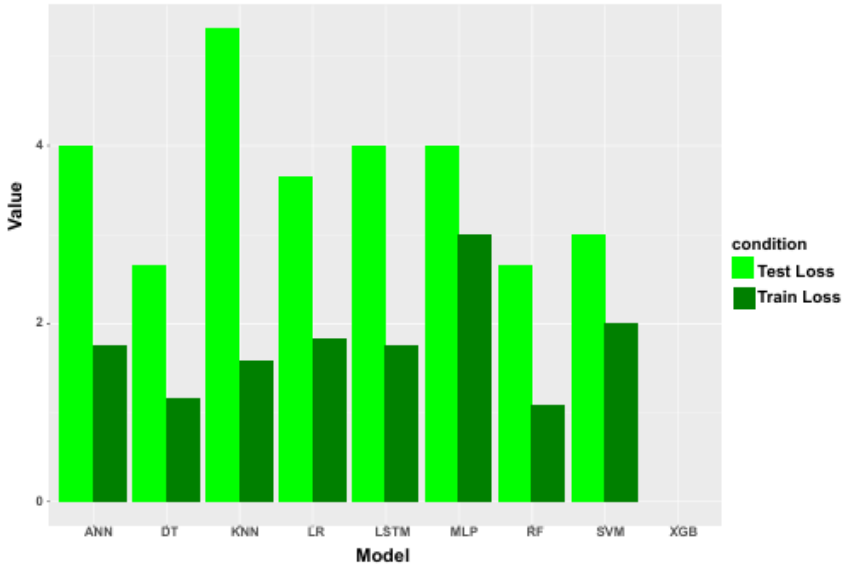

Figure 3 Bar plot illustrated to visualize the loss of different models. Both train and test loss represented to compare different models error.

XGBoost emerged as the most accurate model for this classification task, demonstrating exceptional predictive performance with 100.0% training and testing accuracy, 0.00 training and testing loss, and perfect precision, recall, F1-score, and ROC-AUC, with an execution time of 46.074 seconds. The Random Forest algorithm also performed well, achieving a testing accuracy of 92.3%.

In comparison, the deep learning models in this study were slightly less effective. Three different deep learning models, including an artificial neural network, achieved testing accuracies ranging from 88.5% to 92.3%, showing satisfactory but lower performance compared to XGBoost. Both Random Forest and LSTM-based deep learning models reached a maximum test accuracy of 92.3%, indicating their potential for effective use. However, other deep learning models did not perform as well as the machine learning models.

While our results for diabetes classification are impressive, a major limitation of this study is the relatively small size of the dataset. Despite this, the trained models are robust and are expected to generalize well to similar unseen data.

V. CONCLUSION AND FUTURE WORK

Although there is no conclusive evidence linking age directly to diabetes, early detection of the disease is critical for effective treatment. Machine learning and deep learning have significantly advanced research in risk prediction for early-stage diabetes. In this study, we explored early-stage diabetes prediction using various machine learning and deep learning classification algorithms, incorporating multiple diabetes risk factors.

We evaluated the diabetes dataset using nine classification algorithms, including XGBoost (XGB), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM). Among these, XGBoost achieved nearly 100% accuracy and outperformed all other approaches, demonstrating its superior capability for early-stage diabetes detection.

The findings of this study could support healthcare providers in detecting diabetes earlier, enabling more informed clinical decisions and potentially saving lives. However, the study has limitations, primarily the small sample size, which restricts the statistical significance of the results. In the future, we aim to collect larger, more diverse datasets to improve the accuracy and reliability of disease classification. We also plan to identify additional predictive factors that may facilitate early-stage detection. Furthermore, the proposed approach can be generalized for predicting other diseases, not just diabetes.

REFERENCES

[1]. A. Kharroubi and H. Darwish, “Diabetes mellitus: The epidemic of the century,” World journal of diabetes, vol. 6, pp. 850–67, 06 2015.

DOI: 10.17148/IJARCCE.2025.14820

[2]. D. Gahlan, R. Rajput, and V. Singh, "Metabolic syndrome in north indian type 2 diabetes mellitus patients: A comparison of four different diag nostic criteria of metabolic syndrome." Diabetes & metabolic syndrome, vol. 13, no. 1, pp. 356–362, 2018.
[3]. A. Ahmad, A. Khan, and S. Khan, "Causes, complications and man agement of diabetes mellitus," Chronicle Journal of Food and Nutrition, vol. 1, pp. 1–3, 08 2017.
[4]. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and Ioanna, "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.
[5]. T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: a systematic review," IEEE Journal of Biomedical and Health Informatics, 2020.
[6]. M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in Computer Vision and Machine Intelligence in Medical Image Analysis, vol. 992. Springer, Singapore, 2020, pp. 113–125.
[7]. L. Li, "Diagnosis of diabetes using a weight-adjusted voting approach," in 2014 IEEE International Conference on Bioinformatics and Bioengi neering, 2014, pp. 320–324.
[8]. J. P. Kandhasamy and S. Balamurali, "Performance analysis of classi f ier models to predict diabetes mellitus," Procedia Computer Science, vol. 47, pp. 45–51, 2015.
[9]. P. Agrawal and A. kumar Dewangan, "A brief survey on the techniques used for the diagnosis of diabetes-mellitus." International Research Journal of Engineering and Technology, vol. 02(03), 2015.
[10]. N. Naiarun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," Procedia Computer Science, vol. 69, pp. 132 142, 2015, the 7th International Conference on Advances in Information Technology.
[11]. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.
[12]. T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," International Journal of Medical Informatics, vol. 97, 2016.
[13]. T. Ahmed, "Developing a predicted model for diabetes type 2 treatment plans by using data mining," Journal of Theoretical and Applied Information Technology, vol. 90, pp. 181–187, 2016.
[14]. A. Oleiwi, L. Shi, Y. Tao, and L. Wei, "A comparative analysis and risk prediction of diabetes at early stage using machine learning approach," International Journal of Future Generation Communication and Networking, pp. 4151–4163, 2020.
[15]. M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, "An improved artificial neural network model for effective diabetes prediction," Complex., vol. 2021, pp. 5525271:1–5525271:10, 2021.
[16]. "Pima. university of california, irvine, school of information and computer sciences," Oct 6, 2020. [Online]. Available: https://www.kaggle.com/ uciml/pima-indians-diabetes-database
[17]. T. Chen and C. Guestrin, "Xgboost: Reliable large-scale tree boosting system," in Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2015, pp. 13–17.
[18]. E. Goel, E. Abhilasha, E. Goel, and E. Abhilasha, "Random forest: A review," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 1, 2017.
[19]. A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 18, no. 6, pp. 275–285, 2004.
[20]. W. S. Noble, "What is a support vector machine?" Nature biotechnology, vol. 24, no. 12, pp. 1565–1567, 2006.
[21]. L. E. Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no. 2, p. 1883, 2009.
[22]. J. M. Hilbe, Logistic regression models. Chapman and hall/CRC, 2009.
[23]. C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," The journal of educational research, vol. 96, no. 1, pp. 3–14, 2002.
[24]. A. Odeh, I. Keshta, and E. Abdelfattah, "Efficient detection of phishing websites using multilayer perceptron," 2020.
[25]. A. Abraham, "Artificial neural networks," Handbook of measuring system design, 2005.
[26]. P. Marius, V. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," WSEAS Transactions on Circuits and Systems, vol. 8, 07 2009.
[27]. H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.
[28]. D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.