Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/86/20241610

Apply multiple machine learning models to diabetes prediction

Shitong Qin

Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan, China

20213801024@m.scnu.edu.cn

Abstract. Diabetes mellitus, a pervasive and chronic metabolic disorder, imposes a substantial burden on global health systems due to its requirement for lifelong management and the myriad of complications associated with inadequate control. The ability to accurately forecast the onset of this disease is paramount, as it enables preemptive interventions and tailored treatment strategies that can significantly mitigate its impact. This paper investigates the application of machine learning techniques and deep learning models in diabetes prediction. This paper makes use of the Pima Indian Diabetes Dataset (PIDD) from Kaggle, which has 768 data entries and eight characteristics like blood pressure, blood sugar, and body mass index (BMI). Various algorithms, including Support Vector Machine (SVM), Decision Trees(DT), Random Forest(RF) , and Fully Connected Neural Network (FCNN), are implemented and compared. Identifying the strengths and limitations of each model, the results emphasize the potential of advanced computational models in improving the accuracy and clinical usefulness of diabetes prediction. The best-performing model is the FCNN model, with a test accuracy of 78.67% and an AUC value of 83.36%.

Keywords: Diabetes prediction, Fully Connected Neural Network, Random Forest, Decision Tree, Support Vector Machine.

1. Introduction

Diabetes mellitus, a chronic metabolic condition that is becoming increasingly prevalent and posing a serious threat to global public health, is characterized by elevated blood sugar levels [1]. Globally, the prevalence of diabetes is rising, significantly taxing healthcare systems and impairing millions of people's quality of life. Early prediction of diabetes onset is crucial for preventing its progression and complications, such as cardiovascular diseases, kidney damage, and neuropathy. It is a condition that requires lifelong management and has the potential to lead to severe complications if left uncontrolled. Therefore, the ability to accurately predict the onset of diabetes is of utmost importance, offering a critical advantage in both prevention and treatment strategies for patients and healthcare providers alike.

An analysis of previous research has revealed a deficiency in studies pertaining to diabetes prediction that compare the effectiveness of various machine learning models and employ numerous models for prediction. Previous studies have focused on improving the predictive performance of one model or have not conducted performance comparisons between models such as Random Forest (RF),Decision

DOI: 10.54254/2755-2721/86/20241610

Tree (DT), Support Vector Machine (SVM) and Fully Connected Neural Network (FCNN) for diabetes prediction.

This paper delves deeply into the field of diabetes prediction analysis, focusing particularly on the application of various machine learning techniques and models in diabetes prediction, and comparing the performance of each model. The goal of this paper is to close the gap in the research on diabetes prediction, determine which diabetes prediction model is the best, and make a contribution to the field by analyzing and contrasting the performance of RT, DT, FCNN, and SVM models in diabetes prediction.In this experiment,the Pima Indian Diabetes Dataset (PIDD) from Kaggle is utilized, which includes eight features such as Body Mass Index (BMI), blood sugar, blood pressure, and a total of 768 data entries.The research is based on the implementation and comparative analysis of various algorithms, ranging from conventional machine learning techniques like SVM and DT, to more sophisticated FCNN. Through rigorous evaluation, the unique strengths and shortcomings of each model within the context of diabetes prediction have been identified. The findings of this study highlight the pivotal role that advanced computational models play in enhancing predictive accuracy and, by extension, the clinical utility of diabetes risk assessment.

2. Related work

Machine learning stands as a forefront technology within the realms of artificial intelligence and computer science, is being applied by numerous researchers to solve a variety of complex problems, including the prediction of diabetes [2]. The authors used binary Logistic regression to identify risk factors associated with the comorbidity of diabetic kidney disease (DKD) and hyperuricemia (HUA), and constructed and validated a risk prediction model. The model attained an Area Under the Curve (AUC) value of 0.821 and a sensitivity rate of 78.57%, providing an effective basis for clinical assessment of the risk of DKD patients developing HUA [3]. In [4],Three machine learning techniques, K-Nearest Neighbors (K-NN), SVM, and RF, were chosen by Madhumita Pal and colleagues and utilized on the Pima Indian Diabetes Dataset (PIDD) from the UCI database.The aim was to make early predictions for diabetes. The study results indicate that among the three algorithms, with an accuracy of 78.57%, the RF model was the most accurate in predicting the risk of diabetes.

The research employed a range of feature selection techniques to evaluate the efficacy of four algorithms—RF, Extreme Gradient Boosting (XGB), KNN, and Ensemble Learning (VC)—in diabetes analysis, coming to the conclusion that the RF model worked best [5]. A distinctive aspect of the study was the selection of a larger community dataset with more extensive data.

For the PIDD, the interquartile range method was employed to detect and replace outliers within the dataset. After using three classification methods (DT, Gradient Boosting, and Logistic Regression), their performances were evaluated, and it was determined that the DT model performed the best [6].In order to help the SVM understand complicated decision boundaries and adjust to the complexity of the PIDD, that research attempted to incorporate the Radial Basis Function (RBF) and the RBF block kernel as new kernels. This improved the SVM model's accuracy in predicting diabetes [7].Studies by Srishti Mahajan et al. show that the accuracy of the model is about 99% when using the random forest classification algorithm to predict kaggle's public data set on diabetes. By implementing logistic regression classification, the accuracy is approximately 94% [8]. In [9], Banibrata Paul and other authors used the PIDD in their research and applied the k-fold cross-validation method. By employing an artificial neural network combined with the scaled conjugate gradient backpropagation algorithm for diabetes prediction, they were capable of achieving a prediction accuracy rate of up to 100%. In [10], Authors Na Hu and Jiali Gao utilized the KNN algorithm, DT algorithm, and RF algorithm to predict the PIDD from the UCI database. The Random Forest model emerged as the most effective,attaining an estimated 84% accuracy rate and an approximate 0.77 F1 score.

DOI: 10.54254/2755-2721/86/20241610

3. Methodologies

3.1. Data preprocessing

Data Standardization/Normalization: In the application of models based on distance (like KNN, SVM) or those optimized for gradient descent (such as logistic regression), data normalization or standardization is frequently required. This ensures a consistent scale between different features and prevents certain features from dominating the model. A standard scalar is introduced to transform the dataset. Standard scalars function by deducting the average from each eigenvalue, followed by division by the standard deviation.This normalizes the distribution of features so that the mean is 0 and the standard deviation is 1. The following formula is used:

$$X_{new}{}' = \frac{X-\mu}{\sigma} \tag{1}$$

3.2. Data visualization analysis

From Figure 1, we can observe the following: The blood sugar levels of diabetic patients are significantly higher than those of healthy individuals, indicating that diabetic patients generally have higher blood sugar levels. The blood pressure distribution of diabetic patients is relatively higher compared to healthy individuals, suggesting a higher proportion of hypertension among diabetic patients. The insulin level distribution of diabetic patients is slightly higher than that of healthy individuals. The BMI distribution skews towards higher values, indicating a higher proportion of obesity or overweight among diabetic patients. The age distribution suggests that diabetes may be more common in certain age groups, such as the middle-aged and elderly population.

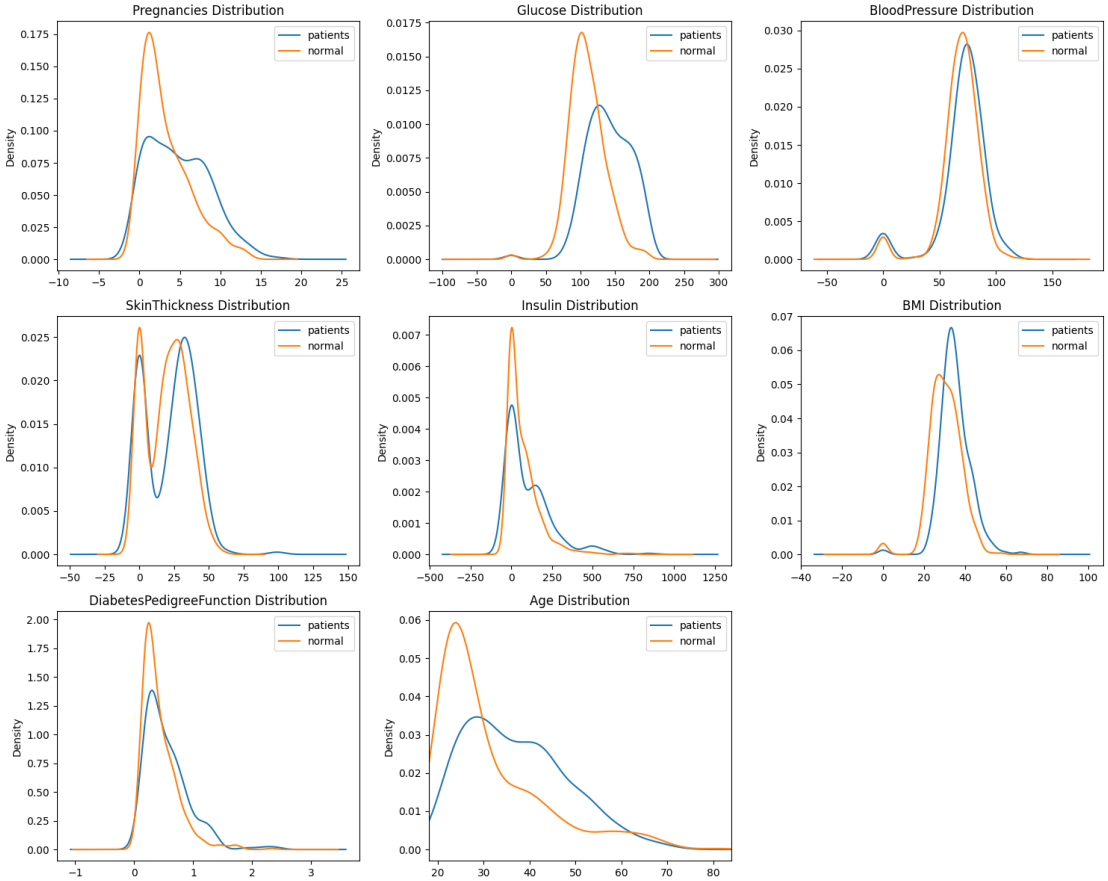

Figure 1. The distribution of each feature (Photo/Picture credit : Original)

DOI: 10.54254/2755-2721/86/20241610

From Figure 2, we can observe that blood sugar has the highest correlation with the final outcome. Following closely are the BMI index and age.

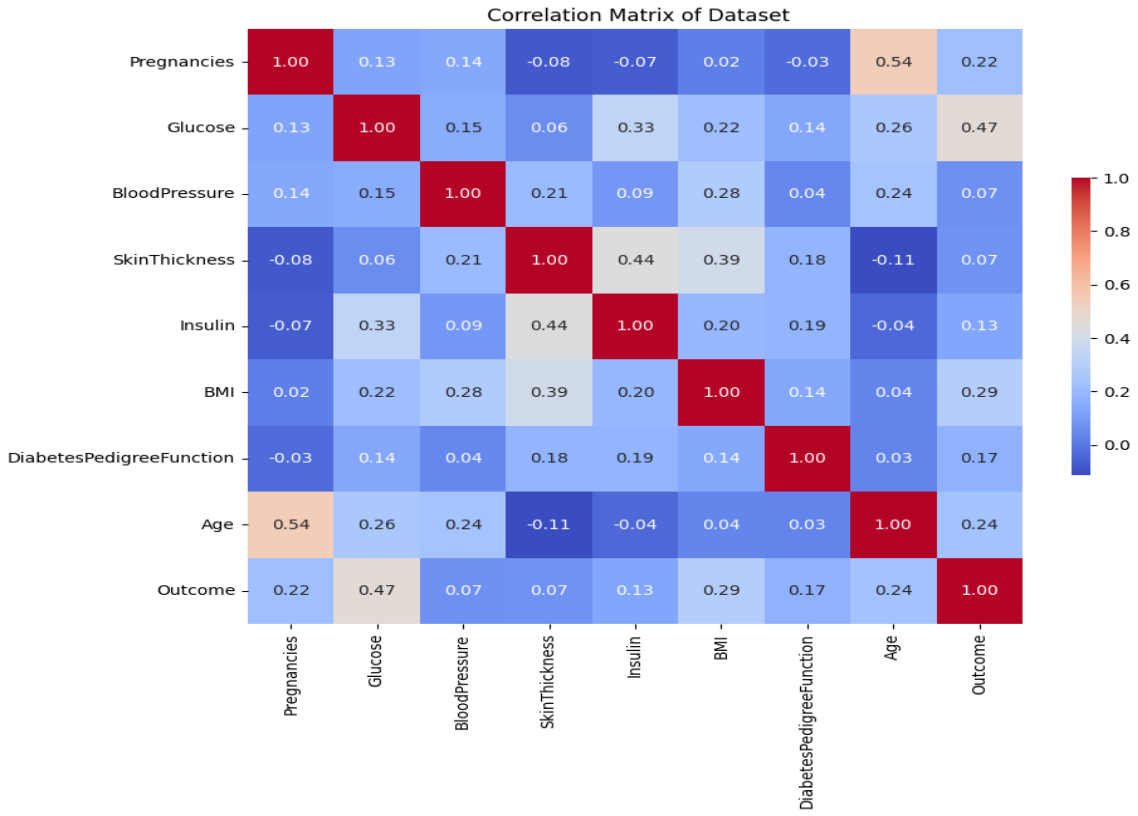

Figure 2. Correlation Matrix of Dataset (Photo/Picture credit : Original)

3.3. Choosing models

3.3.1. Fully Connected Neural Network (FCNN) Multiple fully linked layers, each with a number of neurons, make up a FCNN. Each neuron in the layer above it is connected to all the other neurons and performs a weighted sum of its outputs. Next, a nonlinear activation function is used to determine the activity status of the neuron. A FCNN typically consists of an input layer, one or more hidden layers, and an output layer. To lower prediction errors, the network uses backpropagation to modify weights and forward propagation to compute outputs. FCNN can automatically extract features and learn complex nonlinear relationships, making them suitable for predicting diabetes, where features may have intricate nonlinear connections with the outcome. Well-trained neural networks have good generalization capabilities, useful for handling new, unseen data. In this experiment, a FCNN model was utilized, and the model was optimized using the mini-batch stochastic gradient descent algorithm (Figure 3).

Cross-entropy loss: This type of loss function is frequently used, particularly in classification-related issues. It calculates the discrepancy between the actual probability distribution and the forecasted one.

$$Loss = -y\log\hat{y} - (1-y)\log(1-\hat{y}) \tag{2}$$

Mini-batch Stochastic Gradient Descent: Instead of using one sample at a time, this method selects a random subset of samples to compute the gradient. The parameter θ represents the weights and biases of the FCNN.

$I_t$is a subset of{1,2,…,N} with size d

DOI: 10.54254/2755-2721/86/20241610

$$\theta^{t+1} = \theta^t - \alpha_t \frac{1}{d} \sum_{i_t \in I_t} \nabla_\theta f\left(\theta, x_{i_t}\right) \tag{3}$$

Mini-batch SGD is unbiased

$$E\left\{\frac{1}{d} \sum_{i_t \in I_t} \nabla_\theta f\left(\theta, x_{i_t}\right)\right\} = \nabla_\theta f(\theta, x) \approx \frac{1}{N} \sum_{n=1}^{N} \nabla_\theta f(\theta, x_i) \tag{4}$$

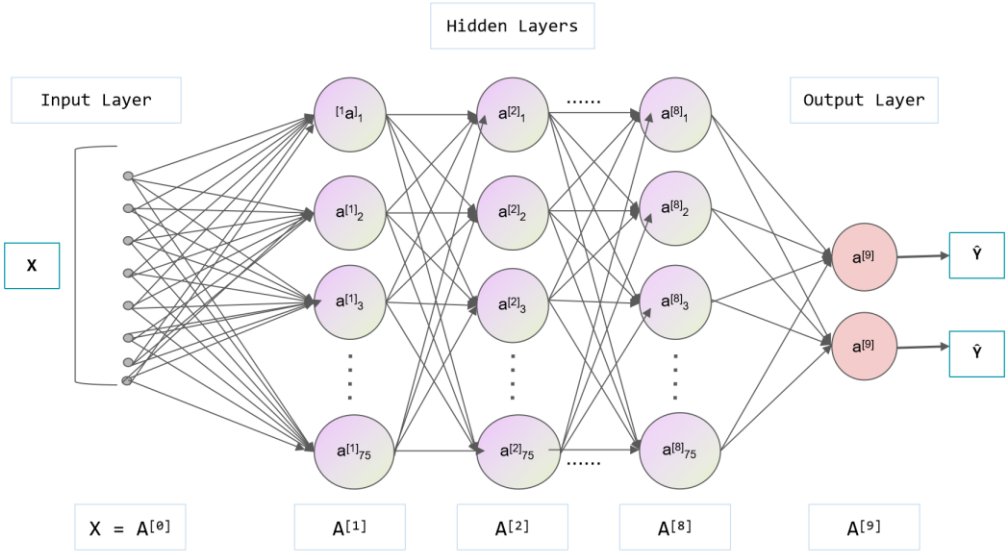

Figure 3. Construction diagram of neural network in experiment (Photo/Picture credit : Original).

3.3.2. Decision Tree The structure mimic a tree where each internal node represents a feature, every branch indicates a feature value, and every leaf node denotes a category. The tree grows by recursively selecting the best feature and split point at each step, using criteria such as information gain or Gini impurity to evaluate feature importance. The model is easy to understand and interpret, providing a clear view of feature importance through its structure.

Recursive procedure:

- Choose an attribute for the root node.
- For every possible value of the chosen property, create a branch.
- Using only the instances that reach each branch, repeat the previous two steps recursively for each branch.
- If a branch includes instances of the same class, stop working on it.
- Depending on the order in which attributes are selected, it is possible to create multiple decision trees.

3.3.3. Random Forest Model Made up of several decision trees, a random forest is a supervised learning model. Because every tree in a random forest is derived from a random sample, the generalization and stability of the model are improved. The random forest aggregates results from several decision trees by voting or averaging to get the final classification or regression output. It is appropriate for small datasets and can handle skewed datasets to some extent. It excels in problems involving anomaly detection, regression, and classification.Utilizing its collective traits, random forest is capable of making highly precise forecasts with a small set of training samples and aids in pinpointing key diabetes risk elements via evaluating the significance of features.

3.3.4. Support Vector Machine (SVM) Model Applied to both classification and regression applications, SVM is a potent machine learning technique. Its foundation is statistical learning theory, specifically

DOI: 10.54254/2755-2721/86/20241610

the maximum margin concept, which seeks to identify the ideal hyperplane for classifying data or predicting continuous values. An optimization issue must be solved in order to determine the ideal hyperplane for cases that are linearly separable. SVM provides kernel functions that translate data into a higher-dimensional space where it becomes linearly separable in circumstances where it is not linearly separable.When using a linear kernel, SVM's forecasting method is straightforward and effective, leading to rapid and precise outcomes.

4. Experiments

4.1. Model performance metrics

The model evaluation metrics used in this study are Accuracy, Precision, Recall, F1 Score, and AUC. AUC (Area Under the Curve) refers to the area beneath the ROC (Receiver Operating Characteristic) curve.

4.2. Hyperparameter setting

Fully Connected Neural Network Model:

- Learning rate: 0.001
- Batch size in Mini-batch Stochastic Descent: 50

Decision Tree Model: class_weight='balanced': This hyperparameter is used to automatically adjust weights to address class imbalance issues, where some classes have significantly more samples than others.

criterion='gini': Used to calculate the impurity at each node during splitting. The ‘gini’ criterion computes the impurity of each node, with a lower value indicating a purer node.

min_samples_split:2

max_depth:15

Support Vector Machine(SVM):

kernel='linear': Indicates that SVM will employ a linear kernel function and will look for a linear hyperplane in the feature space to help distinguish between classes.

Random Forest Model:

n_estimators:500

min_samples_split: 2

max_depth: 25

4.3. Visual analysis of results

The trade-off between each model's true positive rate and false positive rate is shown by the ROC Curve (Figure 4). In comparison to SVM, RF, and DT, the FCNN model's ROC curve displays a higher AUC, indicating superior performance.The AUC value of the DT model is the lowest.This indicates that the FCNN model performs best in the field of diabetes prediction. On the other hand,when it comes to correctly identifying instances as either non-diabetic or diabetic, the DT model lags behind.

Figure 5 and figure 6 depict the confusion matrix representations for a range of methodologies employed in our study. These matrices provide a comprehensive visualization that facilitates an in-depth analysis of the comparative performance among the distinct approaches under consideration. By examining these results, one can discern the effectiveness and limitations of each method in terms of classification accuracy, precision, recall, and F1 score, thereby offering valuable insights into their relative merits and applicability within the context of our research.

DOI: 10.54254/2755-2721/86/20241610

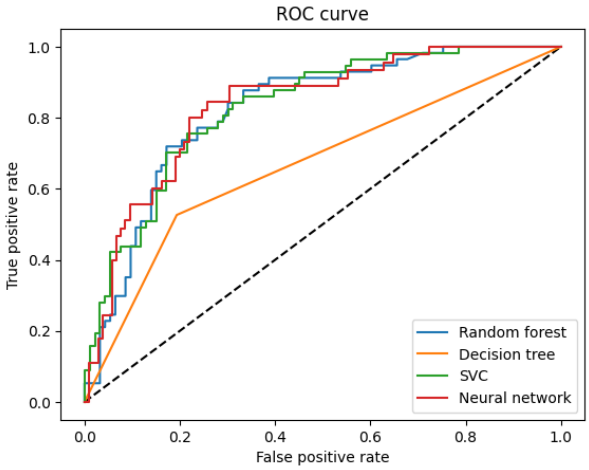

Figure 4. ROC curve of 4 machine learning models (Photo/Picture credit : Original)

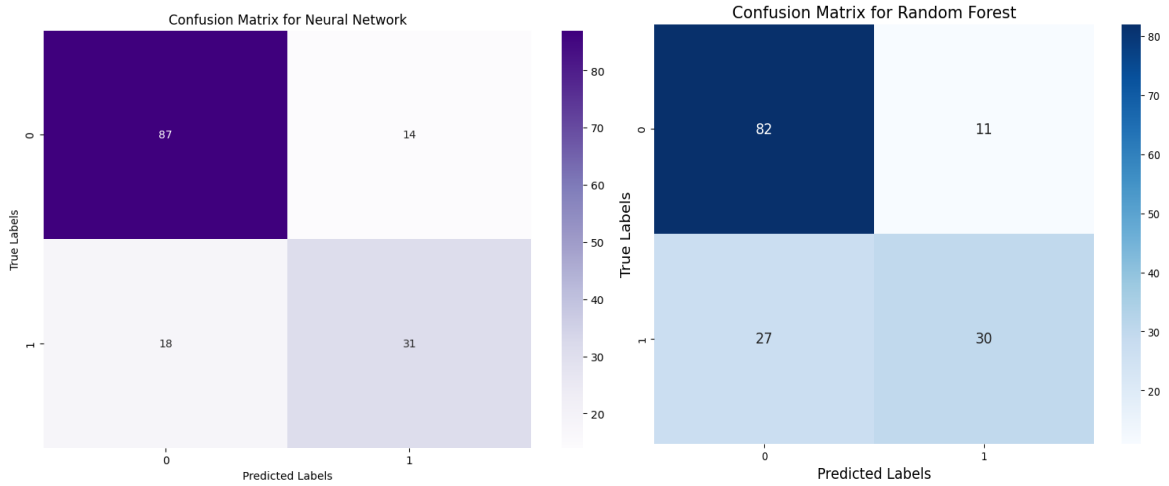

Figure 5. Confusion Matrix for Neural Network and Random Forest (Photo/Picture credit : Original)

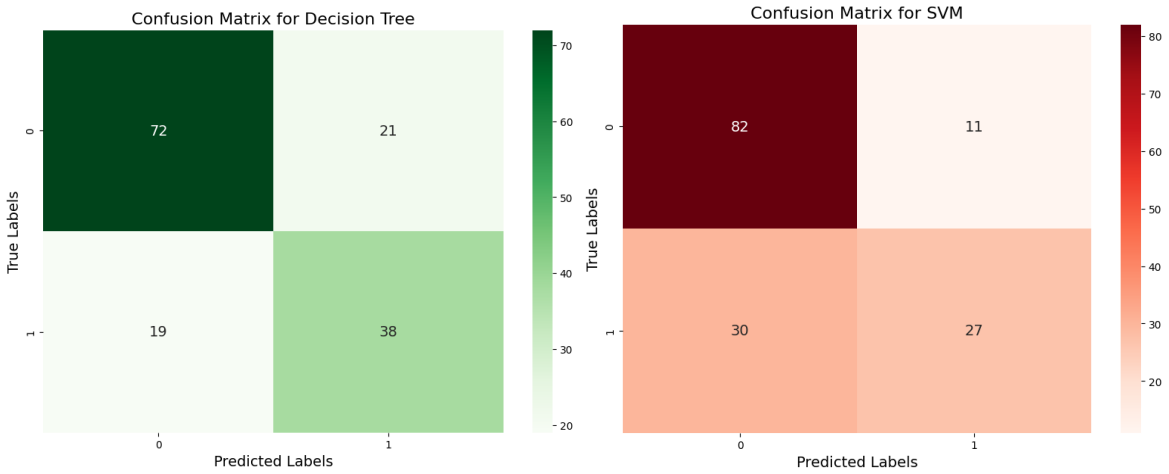

Figure 6. Confusion Matrix for Decision Tree and SVM (Photo/Picture credit : Original)

DOI: 10.54254/2755-2721/86/20241610

By utilizing the confusion matrix, one may determine the F1 Score, Accuracy, Precision, and Recall for each model.

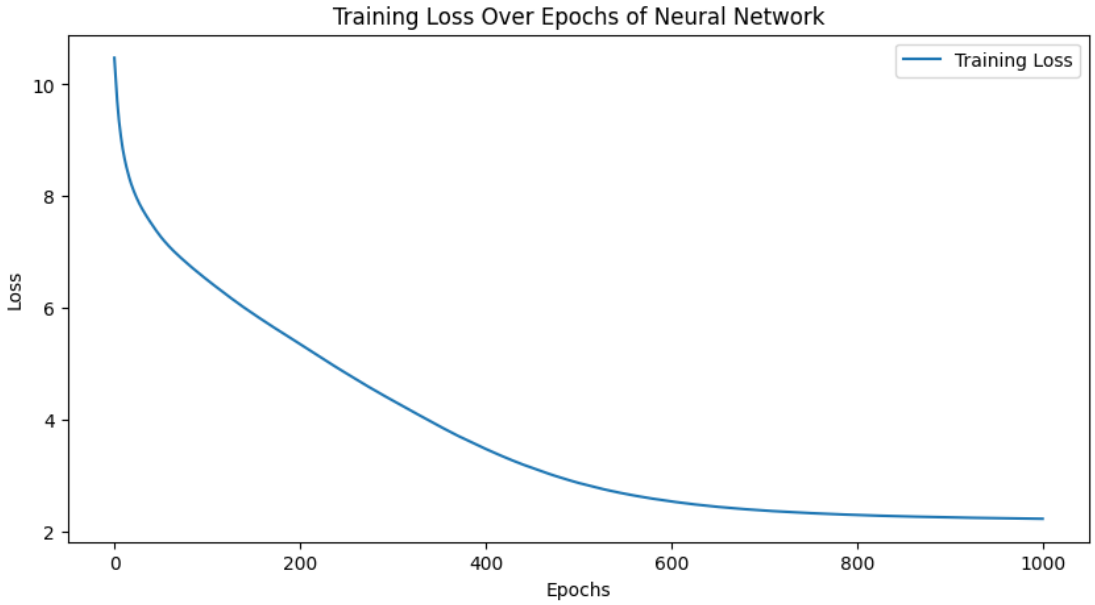

Figure 7. Training Loss Over Epochs of Neural Network (Photo/Picture credit : Original) (Photo/Picture credit : Original)

Loss Distribution displays the FCNN model's convergence and training process (Figure 7). Over epochs,the loss progressively drops, suggesting efficient learning and convergence. Figure 7 illustrates the importance of various features in the diabetes prediction model, where blood sugar levels are considered the most critical predictive factor, followed by BMI and age, as well as diabetes pedigree function. The importance of skin thickness is the lowest (Figure 8).

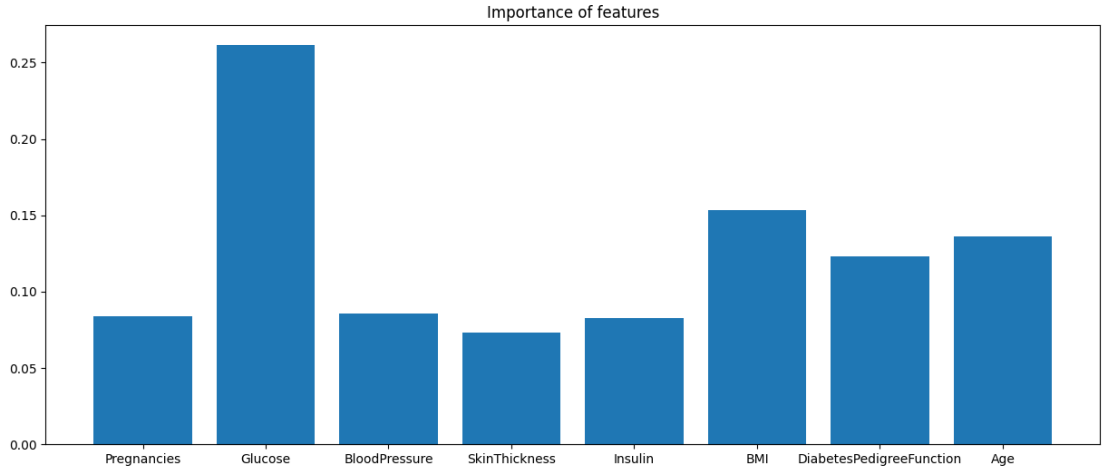

Figure 8. The importance ranking obtained through the random forest model. (Photo/Picture credit : Original)

4.4. Performance Comparison

In terms of accuracy, F1 score, and AUC, the FCNN model performed the best, showing that it can successfully identify latent patterns in the data and generalize well to unseen data (Table 1). The second best-performing model is the Random Forest model, which has slightly higher accuracy, precision, and AUC compared to SVM and Decision Tree. SVM has a moderate performance across all metrics but

DOI: 10.54254/2755-2721/86/20241610

with a relatively high AUC, which demonstrates its potential in diabetes prediction. However, the recall rate of SVM is relatively low, only 47.37%, which may result in the model missing a significant number of diabetic patients. Further optimization is needed to improve its identification capability.The Decision Tree model has a test accuracy of 73.33%, but its AUC value drops to 71.64%, indicating limited ability to distinguish positive and negative samples.

The reasons for the FCNN's good performance are that it can capture and learn intricate patterns and nonlinear relationships in the data, which is highly effective for diabetes prediction. Moreover, in the experiment, an efficient optimization algorithm such as mini-batch stochastic gradient descent was used to optimize the model.

Table1. Model metrics analysis

Model	Test Accuracy	Precision	Recall	F1-score	AUC
SVM	72.67%	71.05%	47.37%	56.84%	82.96%
Decison Tree	73.33%	64.41%	66.67%	65.52%	71.64%
Random Forest	74.67%	73.17%	52.63%	61.22%	82.27%
FCNN	78.67%	68.89%	63.27%	65.85%	83.36%

5. Conclusion

This study has successfully demonstrated the practicality of machine learning models in diabetes prediction. These advanced models have shown significant potential in improving the accuracy of diabetes risk assessment. In particular, the FCNN model was the most effective in this research, attaining the best AUC, F1 score, and accuracy on the dataset. This highlights the ability of FCNN model to identify intricate nonlinear connections within data, which is crucial for accurately predicting the onset of diabetes. The high performance of the FCNN model can be attributed to its sophisticated architecture and the optimization techniques used, such as mini-batch stochastic gradient descent and cross-entropy loss. The accuracy of the FCNN model reached 78.67%. In addition,accuracy, precision, and AUC of the RF model were marginally greater than those of the SVM and DT, indicating strong performance as well. The SVM showed a moderate balance and has potential in terms of diabetes prediction. Although the DT model is simple and easy to understand, its lower performance metrics compared to other models indicate that its predictive power may be limited in the context of diabetes prediction. Besides, the experiment conducted a feature importance analysis and found that the main factors contributing to diabetes are glucose, age, diabetes pedigree function, and BMI, among others. This finding provides valuable insights for medical practice, suggesting that physicians should pay more attention to these features when assessing patients' risk of diabetes. By prioritizing these key indicators, potential diabetic patients can be identified more accurately.The results of this investigation advance the field of diabetes prediction by identifying the FCNN as the best model for predicting diabetes among all tested models. Nonetheless, it's critical to recognize that additional study is required to refine these models and evaluate their performance in more diverse datasets to adapt to a wider range of applications.

References

[1] Alazwari A, Johnstone A, Tafakori L, et al., Alshamrani M.A., 2023. Predicting the development of T1D and identifying its Key Performance Indicators in children. PLoS ONE, 18(3): e0282426.

[2] Mangal A, Jain V, 2022. Performance analysis of machine learning models for prediction of diabetes. Conf Inf Sci Comput Technol, Dehradun, India: 1-4.

[3] Pan L, 2024. Construction and validation of a risk prediction model for diabetic nephropathy complicated by hyperuricemia. Med Theory Pract, (09): 1559-1561.

[4] Pal M, Parija S, Panda G, 2021. Improved Prediction of Diabetes Mellitus using Machine Learning Based Approach. Int Conf on Robot Technol, Chandipur, Balasore, India: 1-6.

DOI: 10.54254/2755-2721/86/20241610

[5] Jiang L, Xia Z, Zhu R, et al., 2023. Diabetes risk prediction model based on community follow-up data using machine learning. Phys Med Rehabil, Volume 35: 102358.
[6] Bhat S, Banu M, Ansari G, et al., 2023. A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. Healthcare, Volume 4: 100273.
[7] Reza M, Hafsha U, Amin R, et al., 2023. Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset. Comput Methods Programs Biomed Update, Volume 4: 100118.
[8] Mahajan S, Sarangi P, Sahoo A, Rohra M, 2023. Diabetes Mellitus Prediction using Supervised Machine Learning Techniques. Int Conf Adv Comput Comput Technol, Gharuan, India: 587-592.
[9] Paul B, Karn B, 2021. Diabetes Mellitus Prediction using Hybrid Artificial Neural Network. Int Bus Soc Sci Conf, Gwalior, India: 1-5.
[10] Hu N, Gao J, 2023. Research on Diabetes Prediction Model Based on Machine Learning Algorithms. Conf Int Prod Autom Eng, Ottawa, ON, Canada: 200-203.