IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Artificial Neural Network based Algorithm for Early Prediction of Diabetes

Mandvi Srivastava[1] and Dr. Nimisha Mishra[2] and Dr. Amitabh Wahi[3]

[1]B.Sc(Hons) Mathematics, Amity University Uttar Pradesh , Lucknow, India

[2]Assistant Professor, Department of Mathematics, Amity University Uttar Pradesh, India

[3]Professor, Department of Mathematics, Amity University Uttar Pradesh, Lucknow, India

ABSTRACT

Diabetes is a chronic disease that is spreading quickly and poses a serious threat to global public health. Its distinguishing characteristics is elevated blood glucose levels caused by inadequate insulin production, impaired insulin action, or both many people are undiagnosed until major adverse effects like kidney failure, neuropathy, cardiovascular disease, and vision impairment appear since it is silent in the early stages. To reduce long term health problem early detection is essential. In recent years, machine learning and artificial intelligence have proven more effective in detecting and recognizing illnesses. When it comes to replicating the complex and non-linear interactions found in clinical datasets, ANN have shown themselves to be the most successful. The dataset used in this work is an ANN-based approach for the early detection of patient records with different independent properties. Extensive data preprocessing methods, including as normalization, scaling using standardization techniques, and dataset partitioning into training and testing sets, were employed to increase model efficiency and stability. A 10-fold cross validation process was used to verify the predictive model's efficiency and accuracy. Two optimization methods-SGD and RPROP, was used to trained the ANN model. In terms of validation accuracy , F1-score, MCC value, and classification stability, RPROP is better than SGD for both optimizers show excellent class classification ability as the ROC-AUC values above 0.99. few false positive and false negatives are confirm by confusion matrix analysis, which is crucial for medical evaluation systems. The proposed methodology can help physicians with early risk assessment, timely clinical intervention, and preventative healthcare planning to improve patient outcomes and reduce the burden of disease.

Keyword: Diabetes, ANN, Machine learning, Early detection, Classification algorithm.

INTRODUCTION

Diabetes mellitus is one of the most prevalent chronic metabolic illnesses worldwide, and its rapidly increasing prevalence has made it a serious public health concern . it is characterized by elevated blood glucose levels due to either insufficient insulin action, or both. Type 1 and Type 2 are two main types of diabetes, with type 2 accounting for about 90% of all occurrences. Insufficient care of diabetes and delayed diagnosis can lead to serious issues including cardiovascular diseases, renal failure, vision loss, and an increased chance of mortality[1]. HbA1c measurements, fasting plasma glucose testing, and oral glucose tolerance esting are widely used in clinical practice. Machine learning techniques such as logistic regression, decision trees, k-nearest neighbors, support vector machines are widely used for diabetes prediction as traditional methods are

often costly, time-consuming, and not suitable for large population [2]. ANNs, a crucial component of deep learning, have attracted a lot of attention in medical prediction problems due to their powerful ability to represent complex and non-linear relationships in data. Because ANNs are made up of linked layers of neurons that instantly recognize patterns from incoming data, a significant degree of manual feature engineering is not required.[1],[3].in response to the growing need for accurate and efficient early diagnostic tool, this research proposes and artificial neural network based algorithm for the early prediction of diabetes. The proposed approach uses a range of clinical parameters to train an ANN model that can identify complex patterns associated with diabetic symptoms.teh developed approach aims to lower misclassification rates while improving accuracy and reliability when compared to conventional machine learning methods. By providing automated and data-driven prediction systems, this project seeks to assist medical professionals with early risk evaluation, timely action, and improved patient care outcomes. According to global assessments supported by the united nation organization(UNO) diabetes has emerged as one of the most significant non-communicable disease. Diabetes is a major cause of long-term morbidity, disability, and early mortality, according to the UNO, diabetes is increasing due to sedentary lifestyles, obesity, aging populations, and poor eating habits. One of the main concerns brought up by UNO is that many diabetics remain undiagnosed , which permits illness to deteriorate and result in major consequences. These issues, which radically lower quality of life and raise healthcare costs, include cardiovascular diseases, kidney failure, neuropathy, retinopathy, and an increased susceptibility to infections, to lessen the worldwide effects of diabetes and enhance population health outcomes, the UN accentuate early screening, preventive healthcare practices, and the use of digital and data-driven technology[4]. Diabetes is accelerating health risk in india that affects both urban and rural populations, according to the Indian council of medical research(ICMR). According to ICMR statistics, dietary modifications, lifestyle changes, and unawareness are the main source of india's rising diabetes occurence. ICMR has brought attention to the significant issue of the high rate of undetected and pre-detected diabetes individuals, particularly in rural and resource constrained areas where access to standard medical testing is restricted. In order to address this issue, ICMR strongly promotes early risk assessment, population-level screening, and the integration of the state-of-the-art computational methods into healthcare systems. [5].Furthermore, the growing incidence of diabetes poses a serious threat to healthcare systems and economic expansion, according to national research groups and international health organizations. Diabetes not only affects a persons health but also result in long-term disability, decreased productivity, and increased medical costs. Both UNO and ICMR stress that systematic screening and early detection using conventional clinical procedures are difficult since early stage diabetes and pre-diabetic illness often remain undiagnosed [4],[5].

Machine Learning Techniques for Diabetes Prediction:

Many studies have looked into the applications of machine learning algorithms for early diabetes prediction using clinical and demographic data. Kavakiotis et al. highlighted the importance of data mining and machine learning methods in diabetes research in their comprehensive review, emphasizing their role in early diagnosis and clinical decision support[6]. Traditional supervised learning methods, have been widely employed for diabetes prediction due to their computational efficiency and ease of use [7],[8].these models use basic health markers including blood pressure, body mass index, insulin level, age, and glucose concentration to identify individuals who are at risk of developing diabetes. Although these approaches provide faster traning times and a limited ability to capture complex non-linear correlations seen in diabetes related clinical datasets sometimes restrict their efficacy[9]. The limitation of traditional machine learning are being addresses by the increasing usage of ANN for diabetes prediction [10,]. according to recent research, ANN-based methods have lower misordering rates and higher sensitivity, which makes them appropriate for early-stage diabetes risk assessment [1],[3].

ANN-Based Early Diabetes Prediction Algorithm:

An ANN is utilized to forecast the early onset of diabetes using preprocessed clinical data. The dataset is cleaned and normalized before being divided into training and testing sets. An input layer, one or more hidden layers, and an output layer make up the ANN architecture that is then constructed. The model is trained using both forward and reverse propagation to minimize classification error. Finally, following testing on unknown data, the trained ANN is used to classify individuals as either having diabetes or not. Using common classification measures such as accuracy, precision, recall, and F1-score, the performance of the suggested ANN-based diabetes prediction model is assessed. In order to show how successful the ANN technique is, its performance is contrasted with that of popular machine learning algorithms including Random Forest(RF), Logistic Regression(LR), Decision Tree(DT), and Support Vector Machine(SVM). Experimental results show that the ANN model performs better in terms of prediction accuracy and generalization than traditional machine learning techniques. The improved performance is due to ANN's capacity to comprehend complex interactions and non-linear linkages among clinical parameters, both of which are crucial for early diabetes detection. The proposed ANN shows good prediction accuracy while ensuring fast processing efficiency, making it suitable for early diabetes detection.

Artificial Neural Network Applications in Disease Prediction:

ANN have become one of the most effective computer models in the healthcare sector due to their ability to extract complex, non-linear correlations from high-dimensional medical data. ANN based models have been extensively studied and shown to outperform conventional machine learning methods in the field of diabetes prediction by automatically learning interaction among clinical parameters . these abilities enable the early identifications of high-risk patients, which facilitates improved disease treatment and preventive measures[1],[3],[6],[10]. Artificial neural networks are used in cardiovascular disease prediction to evaluate patient history, lifestyle factors, and physiological signs in order to determine the risk of heart disease and stroke. By identifying minute patterns and long-term trends in cardiovascular data-which are sometimes challenging to analyze using linear or rule-based approaches-ANN models have shown increased predictive accuracy[6],[9]. Neural networks learn hierarchical feature representative from clinical records and imaging data, which helps with early-stage detection and better survival rates. Similarly, ANN-based techniques have been used in cancer diagnosis and prognosis, especially for breast and lung cancer[6],[9]. ANN-based predictive models have also been beneficial for neurological conditions including Parkinson's and Alzheimer's disease. Neural networks are utilized in these applications to analyze clinical evaluations and patient behavior patterns in order to detect early motor dysfunction and cognitive decline. ANN-based early prediction allows for prompt treatment intervention, which is essential for reducing the rate of illness development and enhancing patient quality of life[6]. In order to identify early renal impairment and lower the risk of developing end-stage renal failure, ANN models have been used to examine biochemical markers and demographic characteristics in the context of chronic kidney disease[9]. By recognizing aberrant liver function patterns from blood test results and patient history, artificial neural networks have further proven their efficacy in the prediction of liver disease. These models help physicians diagnose liver disease early and keep an eye on its condition throughout models help physicians diagnose liver disease early and keep an eye on its condition throughout time[8],[9]. ANN-based systems evaluate patient vitals, environmental exposure, and medical history to forecast disease severity and exacerbation risk in respiratory conditions like asthma and chronic obstructive lung disease, allowing for proactive therapeutic management[6]. Another important application of ANN models is the prediction of hypertension, which supports early risk assessment and the prevention of cardiovascular disease by learning long-term blood pressure trends and lifestyle variables[7,[8]. ANN-based methods have also made substantial advances to the prediction of infectious disease. Neural networks have been utilized in the early detection of illnesses like COVID-19 and pneumonia, where prompt treatment and an efficient public health response depend on quick and precise prediction. ANN models are appropriate for real-time screening and decision assistance during disease outbreaks because, once trained , they make quick inferences[6],[9]. Artificial neutral networks continuously show benefits in terms of

prediction accuracy, adaptability, and computational efficiency across these various disease domains. These features make ANN-based models dependable and scalable for use in large-scale screening, intelligence clinical decision support systems, and early disease prediction in contemporary healthcare applications[1],[6],[9],[10].

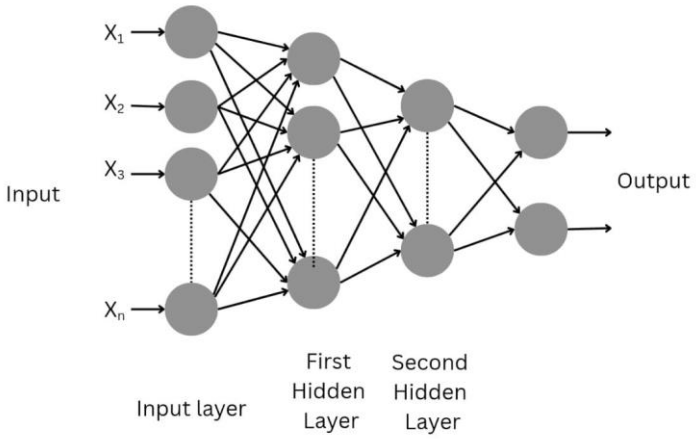


Fig 1. ANN Architecture with Input, Hidden, and Output Layers

Problem Statement

Pattern categorization is a class of tasks that are challenging for computers to solve but that people can typically do quite well. It is difficult and demanding for a machine learning classifier to categorize the patterns into the appropriate classes. If the prediction Y only contains two categories, the classification problem is called a binary classification: if Y has three or more categories, it is called a multi-class classification. The suggested work's goal is to choose a classifier model that would accurately divide the m-dimensional k-class dataset into the appropriate categories. The dataset is incorrectly classified into the appropriate categorize by the classifier. Two class pattern identification problems-diabetes and non-diabetes-make up the dataset for this work. It is an issue of binary classification. SVM-based classifier were taken into the consideration in this method as a mathematical tool to categorize the patterns into the appropriate classes. The Ml models performed the pattern classification more quickly and accurately.

Methodology

This section discussed the planned work's sequential work flow. Fig.2 shows the methods used in this case. Data gathering, preprocessing, data divison, model selection, training, testing, result analysis, and ultimately choosing the optimal model for the intended work are its operational components. The information was gathered from a renowned medical facility. The medical records of 768 individuals, both male and female, were taken into consideration for the inquiry. Diabetic and non-diabetic are its two classes. Ten characteristics were chosen in order to predict diabetes. AGE, RBC, PLT, GLU, CHO, TRIG, HDL, UREA, HBA1C, DIABETIC are the feature names. The charcteristics have both floating-point and integer values. The training set and test set were the two groups into which the datset was separated. ANN-based machine learning models were used to predict whether a person would have diabetes or not. The projected values were used to conduct a performnace study of the approaches. The jupyter notebook's python programming language was taken into consideration when creating the scripts.

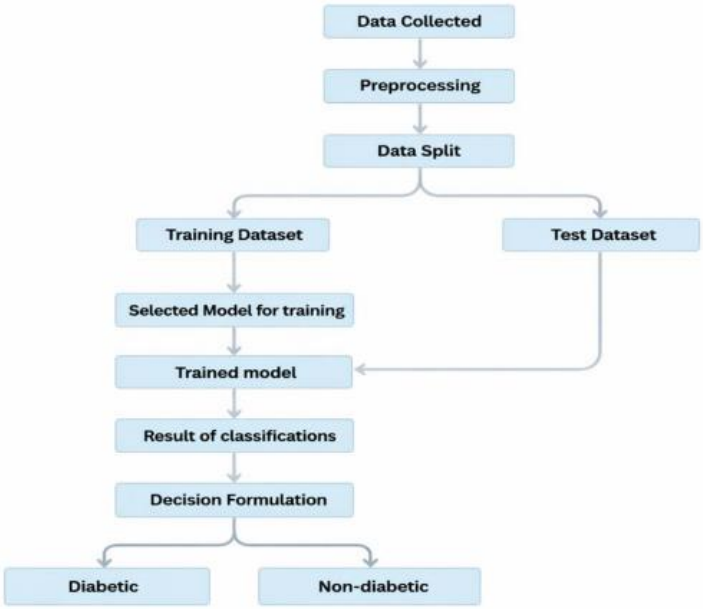


Fig 2 . Overall workflow of the proposed ANN-Based early diabetes prediction systems.

Dataset:

The Reputable hospital provided the data. Ten characteristics are included in the dataset. It is possible to anticipate the diabetic based on these characteristics. The characteristics values were gathered from 768 patients, both male and female, whose visited the hospital for medical examinations and ranged in age from 22 to 81. The dataset distribution in each class is depicted in Fig3,and the data distribution into two groups is explained in table 1. The remaining qualities are regarded as independent features, whereas the final attribute,"diabetic" is the goal or dependent variable. When the diabatic features value is 0, it indicates that the person does not have diabetes, and when it is 1, it shows that the person does. Table 2 list the name of the nine qualities that were offered.

Table 1: Attributes Description inside the dataset

S.NO	PARAMETERS	MIN VALUE	MAX VALUE
1.	AGE	18	77
2.	RBC	1.5	7.6
3.	PLT	74	690
4.	GLU	1.9	501
5.	CHO	82	345
6.	TRIG	40	718
7.	HDL	4	100
8.	UREA	12	88
9.	HBA1C	4.1	131
10.	DIABETIC	0	1

Table 2: Data distribution of diabetes v/s non diabetes

Data set	Description
No.of non diabetic cases	62.29%
No. of diabetic cases	37.71%
Total no. of samples	761
Category	2
Missing feature	0

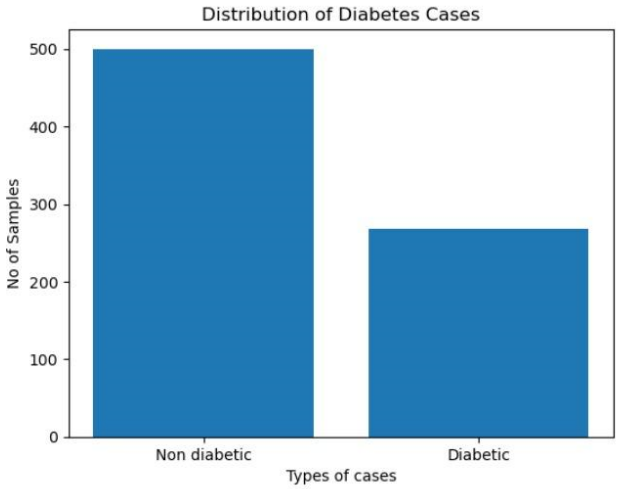


Fig 3: Distribution of data in each classes

Preprocessing

Preprocessing is an essential step in enhancing the quality of the dataset. Extreme values ad outliers were sought in the dataset. The dataset contained no extreme value or outliers. Finding outliers, identifying missing values in each row or column, and normalizing the dataset are all part of the preprocessing procedure. The dataset's columns were examined for nay missing values. The dataset contained no missing values. Prior to diving the dataset into training and test sets, the zero mean method was used for data normalization or scaling. Normalization is necessary to remove the possibility that a feature with a higher value may overpower a feature with a lower value during the model's training phase. The dataset can be scaled using a variety of techniques. Standard scaler normalization is the normalization technique used when feature values falls between[-1,1]. The algorithm was used to standardize the dataset. $Z = X - \frac{\mu}{\sigma}$ where z = normalized value , X = raw data , μ = mean value, σ = standard deviation Prior to training the model, the data was preprocessed. When the mathematical model is properly preprocessed, it learns from clean, scaled data and performs better, increasing prediction accuracy and stability.

PERFORMANCE EVALUATION METRICS FOR CLASSIFICATION MODEL

Several performance evaluation criteria are used to evaluate a categorization model's efficacy. These metrics offer a of the numerical assessment of the model's ability to forecast the desired result. While precision and recall gauge the quality of positive prediction, accuracy displays total correctness. The F1-score provides a balanced assessment of the model by combining precision and recall

Accuracy = $\frac{TP+TN}{TN+FN+FP+TP}$

Recall= $\frac{TP}{TP+FN}$

Precision= $\frac{TP}{TP+FP}$

F1-Score= $\frac{2\times precison\times recall}{precision+recall}$

Experiment

Python program was consider for development of codes of ANN trained with different optimizer, 10-cross fold validation was applied inside the program code and the average values of 10-cross fold validation result were obtained such as training accuracy, validation accuracy, F1 –score, precision and recall score respectively.

Result and Discussion:

Stochastic gradient descent(SGD) and Resilient Backpropagation (RPROP) are two optimization methods that were used to assess the performance of the suggested ANN-based early diabetes prediction model. Ten-fold cross validation was used for the evaluation, and performance indicators such as accuracy, precision, recall, F1-score, ROC-AUC were analyzed.

Prediction accuracy and classification performance:

According to the experimental findings, SGD and RPROP both perform well in terms of prediction for the early identification of diabetes. On the majority of evaluation metrics, however, RPROP consistently performs better than SGD. In terms of classification accuracy, RPROP performs better than SGD based on the average validation accuracy across folds. Moreover, RPROP's F1-score and precision are higher, indicating a superior trade-off between sensitivity and specificity. This shows that RPROP is excellent in reducing misclassification errors, in the early identification of diabetes patients. The ROC curves gives additional evidence for this observation. The ROC curve of RPROP is somewhat higher than SGD, this shows a superior ability to differentiate between diabetic and non-diabetic classes.

Training and validation behavior:

The train and validation accuracy trends across folds indicate that both optimizers have high generalization capabilities. On the other hand, RPROP shows dependable convergence characteristics. Whereas SGD is showing a bit of a large fluctuation, RPROP is consistently maintaining high validation accuracy across 10-folds. This stability shows that the learning rate and gradient magnitude choices have a lower impact on RPROP.

Computational Efficiency and Time Analysis:

The time analysis reveals:

- SGD has a reduced wall-clock time and average CPU.
- RPROP takes a little longer to calculate.
- RPROP's increased precision and stability make the time increase worthwhile.

Both optimizers are computationally viable for real-time early diabetes prediction systems. However, if a little higher computing cost is acceptable RPROP provides better predictive performance.

Implication for Early Diabetes Prediction:

Early diabetes prediction requires high sensitivity, strong classification stability, and low false negatives. Taking into account the experiment's findings

- Higher accuracy and F1-score are attained via RPROP.
- RPORP performs better in terms of ROC.
- Results from cross-validation are more consistent when RPROP is used.
- Although SGD provides quicker training, its prediction performance is marginally worse.

Because of its greater accuracy and superior generalization capacity, RPROP is therefore more dependable for early though SGD offers computational efficiency.

SGD and RPROP, were two optimization methods used in comprehensive tests to evaluate the effectiveness of the proposed ANN model for early diabetes prediction. Cross validation was used to evaluate the model's performance to guarantee its dependability. To understand the model's prediction performance, a variety of assessment criteria were taken into account. Training accuracy, validation accuracy, ROC-AUC, precision, recall. F1-score, MCC are some of these measurements. When combined, these metrics give a thorough insight of categorization performance, particularly in clinical reporting where reducing misordering is pivotal

Optimizers	Train accuracy	Validation accuracy	ROC-AUC	PR-AUC	Precision	Recall	F1-score
SGD	0.9898	0.9764	0.9953	0.9930	0.9669	0.9708	0.9688
RPROP	0.9907	0.9869	0.9955	0.9905	0.9889	0.9764	0.9825

Table 3: Performance comparison of optimizers

Graphical Comparison of Model Performance Metrics

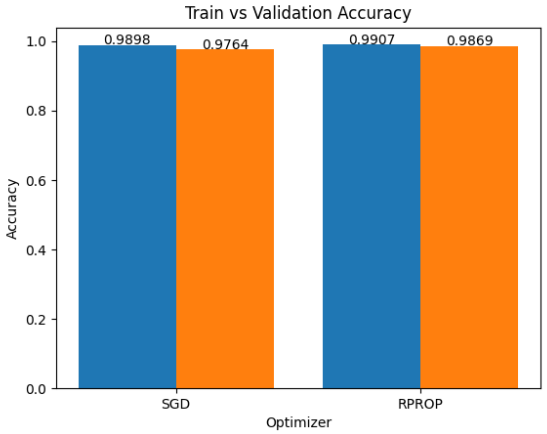


Fig 4:Train vs Validation accuracy

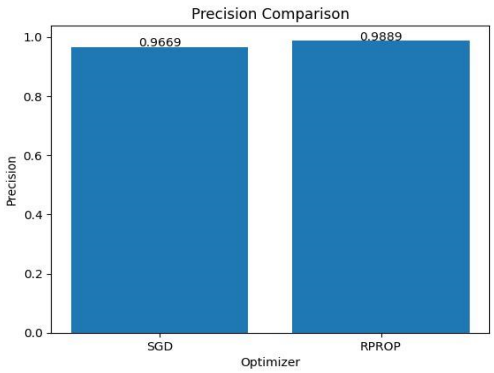


Fig 5 : Comparison of precison

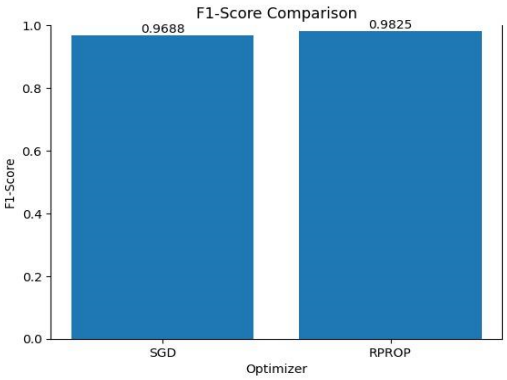


Fig 6 : Comparison of F1-score

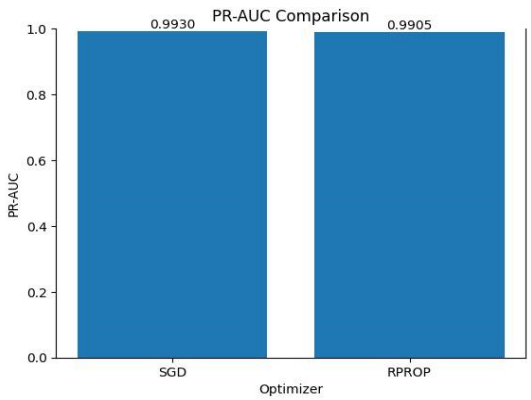


Fig 7 : Compariosn of PR-AUC

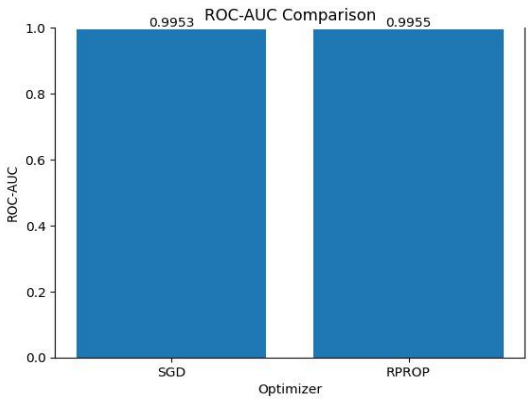


Fig 8 : Comparison of ROC-AUC

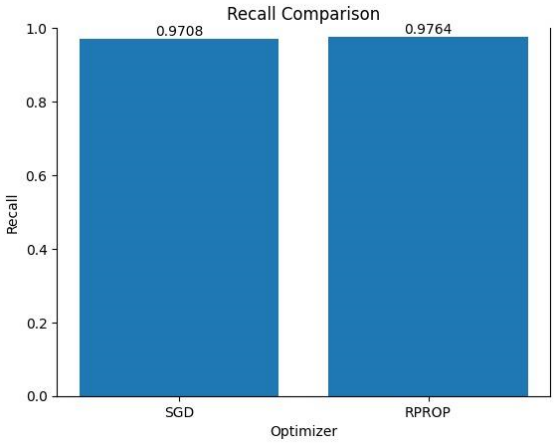


Fig 9 : Comparison of recall

Mean confusion matrices

Confusion Matrix-SGD

[117 2 2 70]

Where:

- True Negatives-117
- False Positives-2
- False Negatives-2
- True Positives-70

Confusion Matrix – RPROP

[118 1 2 70]

Where:

- True Negatives-118
- False Positives-1
- False Negatives-2
- True Positives-70

The experimental result indicate that SGD and RPROP both had good discrimination abilities and high accuracy for early prediction of diabetes. Better generalization performance is indicated by the greater validation accuracy of RPROP(98.69%) compared to SGD(97.64%). Both optimizers work exceptionally well in terms of recall, precision, and F1-score. However, RPROP work somewhat better than SGD, achieving a higher F1-score, suggesting a better trade-off between sensitivity and specificity. There are very few false and false negatives in both the models, the confusion matrix study shows. For the both the optimizers ROC-AUC values are greater than 0.99 , this indicate better class separability between cases with or without diabetes. This results support the robustness and dependability of the proposed ANN model for early prediction of diabetes.

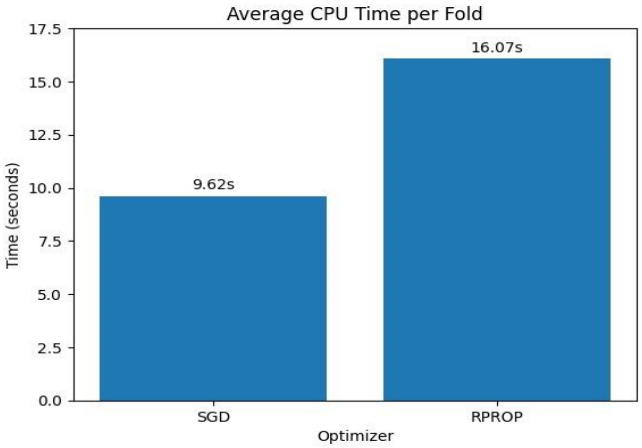


Fig 10 : Average CPU time per fold

Conclusion

Using organized medical data, the study properly created and assessed an Artificial Neural Network(ANN) based model for the early prediction of diabetes. Diabetes should be detected early to avoid serious issues and save long term health expenditures. High level of accuracy was achieved by the suggested ANN model, which successfully represented linear and non-linear interactions among medical characteristics. Optimization techniques- SGD and RPROP, demonstrate excellent prediction abilities during 10-fold cross-validation. On the contrary to SGD, RPROP show better F1-score, higher validation accuracy, and more steady performance across folds. RPROP was more reliable for accurate early-stage diabetes prediction, although SGD took low time complexcity. The model's significant diagnostic capability was further validated by ROC-AUC values which is greater than 0.99. The results show that ANN-based prediction systems can be helpful tools for clinical decision support for early risk evaluation. In the future, the model can be tested on large datasets and put into practice in actual clinical settings.

Reference

1. A.H. Abed," Leveraging Diabetes Prediction using the Deep Learning based Hybrid ANN-CNN Architecture", International Journal of Advanced Networking and Applications, vol. 16, no.6, pp.6683-6690, 2025.
2. V. Vaishnavi, R. Jha, and S. Mitra, " Comparative Study on Prediction of Diabetes Disease using Various Machine learning Models," in Proc. Int. Conf. on Recent Advances in Artificial Intelligence for Sustainable Development (RAISD 2025), Advances in Intelligent Systems Research, vol. 196, pp. 286-297, 2025.
3. K.M. Jyoti Rani, " Diabetes Prediction using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and information Technology, vol. 6, no. 4, pp. 294-305, 2020.
4. United Nations, Political Declaration of the High Level Meeting on the Prevention and Control of Non-communicable Disease, United Nations General Assembly, New York, NY, USA,2011.
5. Indian Council of Medical Research (ICMR), Guidelines for management of Type 2 diabetes, ICMR, New Delhi, India, 2018.
6. J.W. 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.
7.S. Sneha and D. Gangil, " Analysis of Diabetes Mellitus for Early Prediction Using Machine Learning Techniques," International Journal of Computational Intelligence Studies, vol.3, no.3, pp.1-12,2019.
8. P.Singh, S.Verma, and S.K. Dwivedi, " Diabetes Disease Prediction Using Machine Learning Algorithms," International Journal of Computer Applications, vol.178, no.8. pp.19-23, 2019.
9. A. Choudhury and D. Gupta, " A survey on Medical Diagnosis of Diabetes using Machine LearningTechniques," Journal of Medical Systems, vol.43, no.2, pp. 1-14, 2019.
10. T.Kavitha and R. Kannan, " Early Detection of Diabetes Mellitus using ANN " International Journal of Advanced Research in Computer Sciences, vol.9,no.1, pp.224-228, 2018.