European Journal of
Technology
(EJT)

Towards Early Forecast of Diabetes Mellitus via Machine Learning Systems in Healthcare

Sri Krishna Kireeti Nandiraju, Sandeep Kumar Chundru, Srikanth Reddy Vangala, Ram Mohan Polam, Bhavana Kamarthapu, Ajay Babu Kakani

European Journal of Technology
ISSN 2520-0712 (online)
Vol.9, Issue 1, pp1 35 - 50, 2025
www.ajpojournals.org

Towards Early Forecast of Diabetes Mellitus via Machine Learning Systems in Healthcare

Sri Krishna Kireeti Nandiraju[1*], Sandeep Kumar Chundru[2], Srikanth Reddy Vangala[3], Ram Mohan Polam[4], Bhavana Kamarthapu[5], Ajay Babu Kakani[6]
[1*]University of Illinois at Springfield, [2]University of Central Missouri, [3]University of Bridgeport, [4]University of Illinois at Springfield, [5]Fairleigh Dickinson University, [6]Wright State University,

Article history
Submitted 01.07.2025 Revised Version Received 02.07.2025 Accepted 04.07.2025

Abstract

Purpose: Diabetes mellitus poses a major challenge to global health, especially in developing nations where early detection and treatment remain difficult. Using a Convolutional Neural Network (CNN) technique, this study seeks to construct an effective early prediction model for diabetes. The focus is on improving diagnostic capacities in settings with limited resources.

Materials and Methods: The proposed methodology utilises the PIMA Indians Diabetes dataset. A rigorous data preprocessing pipeline was applied, such as min-max normalization, outlier identification and elimination and organized median imputation of missing values. One possible approach to addressing the issue of class imbalance is the Synthetic Minority Oversampling Technique (SMOTE). The dataset was divided into two parts namely the training set and the testing set by using a ratio of 80:20. Some of the measures adopted to determine the efficiency of a convolutional neural network (CNN) model that was trained on diabetes prediction included accuracy, recall, precision, F1-score, AUC, and Brier score.

Findings: The CNN model results were better as compared to the baseline models. It was more accurate than the ANN-based model and the RF-based model, with respective accuracies of 90.29% and 82.35%. With an F1-score of 96.06 per cent and a recall of 96.66 per cent, the CNN model demonstrated considerable faithfulness and predictive power.

Unique Contribution to Theory Practice and Policy: The proposed CNN model, with its high accuracy and reliability, has great potential for combining it with a tele-monitoring device to support early diagnosis and routine monitoring of diabetes, especially in underserved areas. Future studies may be of interest in the deployment and validation of the model in real-time within clinics to enhance its practical value.

Keywords*: Diabetes mellitus healthcare, Convolutional Neural Network (CNN), machine learning, PIMA Dataset*

JEL Classification: *I10, I12, C45, C53*

INTRODUCTION

Diabetes mellitus is a debilitating and potentially fatal metabolic disorder that affects people on a global scale. Public health policy-statistics show that conditions for diabetes are on the rise, especially in developing countries where healthcare access and resources are limited [1]. There are two primary types of diabetes mellitus. In type 1, insulin production is impaired, and in type 2, insulin metabolism plays a role [2]. Poorly controlled diabetes mellitus can result in severe complications including cardiovascular disease, complete renal failure, neuropathy, blindness and stroke. Diabetes mellitus significantly affects both patients and the healthcare system around the world.

Even with the well-known dangers involved with early detection of diabetes, it remains a big challenge. Many traditional diagnostic methods depend heavily on medical visits and glucose testing, which can be uncomfortable and missed opportunities for individuals with limited access to healthcare or limited mobility. Available technologies, such as Continuous Glucose Monitoring (CGM) devices, are useful technological solutions. CGM devices enable remote testing and continuous/prolonged monitoring of blood sugar levels [3]. The low adoption of CGM wearable devices in low-resource settings is closely tied to affordability and accessibility [4]. In addition, people in the most remote areas often lack access to specialised health professionals who are expert enough to intervene rapidly. In response, tele-monitoring platforms have been developed that harness mobile technology, by allowing for glucose readings to be transmitted to healthcare providers via SMS or telemonitoring procedures over GSM networks [5]. However, these platforms still require accurate and early diagnosis to be effective.

Researchers are considering AI and ML methods as rapid and accurate detection becomes more important. There has been some encouraging progress in the use of ML models for diabetes classification and risk factor identification, including DT, SVM, and ANN [6][7]. These models only improve with more data, and the increased amount of data allows them to be more personalized in their assessment of risk and to assess the available options for planning clinical decisions. They also provide an adaptable, scalable, and evidence-based approach to addressing the major clinical gap left by traditional diagnostic measures.

The study addresses these issues and demonstrates the use of a Convolutional Neural Network (CNN) model to identify diabetes at an early stage, utilising the PIMA Indians Diabetes dataset. The model employs class balancing and data preprocessing to ensure high performance. According to our results, the proposed CNN model is more accurate and recalls the traditionally used methods. The strategy will lead to early treatment and ongoing management of patients in poorly supported or remote regions when combined with tele-monitoring sites. The impact of the work is its potential to promote accessible, AI-powered healthcare interventions that lower risks, improve disease trajectories, and reduce the burden on healthcare systems.

Problem Statement

Despite increasing prevalence of diabetes mellitus and consequent severe health implications worldwide, early and precise diagnosis of the condition remains a very challenging undertaking; particularly in low-resource areas. Typical diagnostic tools rely on frequent clinical visits and invasive tests, neither of which can serve populations that lack access to healthcare. Although technologies such as Continuous Glucose Monitoring (CGM) and tele-monitoring platforms show promise, their performance is hindered by high prices, poor adoption, and a reliance on timely diagnosis. Additionally, the existing machine learning models are promising. Still, they lack the accuracy and robustness of their performance due to

the poor processing of the data before it is used and deficient feature interaction modelling. Thus, the necessity of a scalable and cost-efficient solution with a high level of reliability and the ability to predict diabetes with sufficient accuracy, particularly in underserved communities, is very urgent.

Motivation of the study

The increasing prevalence of diabetes and its potential extreme comorbidities emphasize the necessity to have an accurate and accessible as well and scalable, early detection system. Conventional diagnostic devices have weaknesses associated with small data sizes and the human capacity to interpret results, which can slow down response times and lead to hazardous health conditions. DL, and CNNs specifically, have demonstrated their capability to automate search and pattern identification, as well as enhance accuracy in prediction, which is promising in the context of an increasing presence of AI in healthcare.

Although CNNs have been historically considered well-suited for performing on spatially organised data, regardless of whether they are used on images, it has been recently shown that they are also useful for establishing hierarchical and non-linear relationships on structured data by encoding tabular information as a 1D feature map. In contrast to ANNs or tree-based models where features are considered either separately or in pre-determined hierarchies, CNNs allow learning local dependencies and interactions between subsets of features. This feature enables CNNs to identify minor yet perhaps essential videograms in medical information, videograms that can help enhance the effectiveness of classification and decrease FPs and FNs.

The paper discusses an experiment to determine whether a convolutional neural network (CNN) model can be used to improve the early prediction of diabetes based on the PIMA Indians Diabetes dataset. The goal is to enhance clinical decision-making and support resource-constrained settings through automation and tele-monitoring capabilities. Here are four concise bullet points summarizing the motivation and contribution of the study:

- Developed a robust CNN-based model that predicts the occurrence of diabetes with a high accuracy of 96.66% using the PIMA Indians Diabetes dataset.
- Justified the use of CNN for structured data by leveraging its ability to model local feature interactions, which traditional models may overlook.
- Implemented advanced data handling techniques, including missing value imputation, outlier treatment, and class imbalance correction, to improve model reliability.
- Conducted comparative analysis showing the CNN model outperforms traditional ML models such as ANN and Random Forest (RF) in diabetes prediction.
- Provided a scalable and automated approach suitable for integration with tele-monitoring platforms to improve early intervention and patient outcomes.

Novelty and Justification of the Study

In this work, a novel application is presented using a customized CNN architecture specifically tailored for the Indian Diabetes Prediction Model (PIMA). Traditional machine learning methods, such as Logistic Regression, Decision Trees, and Support Vector Machines, have been widely used in previous research on diabetes prediction. While some recent works have explored CNNs, they often suffer from key limitations: minimal or no data preprocessing, inadequate handling of class imbalance, and a lack of model generalizability due to overfitting on small or unbalanced datasets. Moreover, many prior approaches required extensive manual feature

engineering and failed to leverage the CNN's full potential for automatic pattern recognition in structured medical data.

This study addresses these gaps through a more comprehensive and refined approach. First, it integrates robust data pre-treatment strategies, including missing value imputation using grouped medians, outlier correction, and class imbalance resolution via Synthetic Minority Oversampling Technique (SMOTE). These preprocessing steps enhance model robustness and reduce bias. Second, the proposed CNN architecture is optimized to learn complex feature interactions directly from the raw input, minimizing the need for manual feature selection. Lastly, a detailed comparative analysis against baseline models—ANN and RF—demonstrates the superior accuracy, recall, and F1-score of the CNN approach, confirming its practicality for early diabetes detection, especially in data-limited and resource-constrained environments.

Structure of Paper

The paper is structured as follows: Section II discusses relevant research on diabetes mellitus in healthcare utilizing both conventional and ML methods. Section III explains the CNN model and the suggested approach. Section IV explains the experimental findings and suggests a baseline model for diabetes prediction. Finally, Section V provides an overview of the study's key findings, applications, and suggestions for more research.

Literature Review

This section presents research on diabetes Mellitus in Healthcare systems that utilize diverse ML techniques; the summary of these studies is provided in Table I.

Yahyaoui et al. (2019) reviewed the 768 records comprising the PIMA Indians Diabetes dataset, which boasts eight clinical metrics. When it came time for classification, their research used SVM, DL, and RF models. A total of 268 samples were found to have diabetes, while 500 samples did not. With RF outperforming the other two methods, the reported accuracies were 83.67%, 65.38%, and 76.81%, respectively. However, the underperformance of DL suggests that deep models may require more careful tuning and advanced preprocessing to work effectively on small, structured datasets[8].

Deo and Panigrahi (2019) developed a model that predicts the probability of diabetes development based on demographics and lifestyle. Training samples were 98 and testing samples 42. They tested models that consisted of bagged trees and linear SVM with holdout and five-fold cross-validation approaches. The linear SVM achieved an accuracy of 91% (AUC = 0.908) in both validation methods. Although these results may be promising, the extremely small sample size raises concerns about overfitting and the limitations of the extra palatability of the results [9].

Islam et al. (2019) established 340 samples (26 of patient features) classified in Typical and Non-Typical diabetes cases, which included 26 features based on the presence of diabetes, and signs and symptoms. Cross-validation was used with RF, LR and Bagging classifiers. RF achieved the greatest precision of 90.29%, whereas Bagging settled at 89.12% and LR at 83.24%. The research, however, paid more attention to specific questions than planned clinical variables, and systematic preprocessing procedures, such as handling outliers and missing data, were not undertaken [10].

Kaur et al. (2018) proposed an IoT-enabled (Internet of Things) cloud framework for predicting diabetes dates using smart wearable devices. These devices continuously recorded data on blood glucose, which was then transferred to the cloud for analysis using ensemble models. They used

five ML algorithms to create ten ensembles with the highest accuracy of 94.5 per cent achieved in a model which integrated Decision Trees and Neural Networks in the PIMA dataset. Although new, this method presupposes the already available infrastructure of real-time monitoring that is, however, inapplicable in low-income or rural areas[11].

Alassaf et al. (2018) focused on the diagnosis of diabetes by analysing clinical data from Saudi Arabia's King Fahd University Hospital (KFUH). Several ML models were used following feature selection and preprocessing; however, the Artificial Neural Network (ANN) achieved the best testing accuracy of 77.5%, surpassing SVM, Naive Bayes (NB), and K-Nearest Neighbours (KNN). However, the study's reliance on a single institution's dataset limits its applicability across diverse populations[12].

Table I provides an overview of the literature review on Diabetes Mellitus in Healthcare, including methodology, data, key findings, limitations, and future work.

Table 1: Comparative Analysis of Literature Review Based on Diabetes Mellitus in Healthcare using ML and DL Model

Author	Methodology	Data	Key Findings	Limitation	Future Work
Yahyaoui et al. (2019)	DL, SVM, and RF models	Pima Indians Diabetes dataset (768 samples, 8 features)	RF achieved the highest accuracy (83.67%) vs DL (76.81%) and SVM (65.38%)	Limited to basic models; no deep analysis of feature impact	Explore hybrid or ensemble models
Deo and Panigrahi (2019)	Bagged Trees and Linear SVM with 5-fold and holdout validation	140 samples (98 training, 42 testing); lifestyle and demographic features	Linear SVM achieved highest accuracy (91%) and AUC (0.908)	Small dataset; possible overfitting	Increase dataset size; include more features
Islam et al. (2019)	Bagging, Logistic Regression, and Random Forest	340 diabetic patients with 26 features; classified as Typical and Non-Typical	RF had highest accuracy (90.29%), followed by Bagging (89.12%) and Logistic Regression (83.24%)	Focused only on diabetic patients; lacks healthy controls	Include healthy controls for balanced dataset
Kaur et al. (2018)	Cloud-IoT based monitoring and ensemble models using Decision Tree & Neural Net	Pima Indians Diabetes dataset; wearable IoT devices for real-time glucose monitoring	The accuracy of the Decision Tree and Neural Network ensemble was 94.5%.	Depends on wearable device infrastructure	Expand to real-time deployment and testing

Alassaf et al. (2018)	ANN, SVM, Naïve Bayes, and K-NN	Clinical data from KFUH, Saudi Arabia	ANN achieved highest testing accuracy of 77.5%	Region-specific data may limit generalizability	Apply model to broader population and other regions

Research Gap

While previous studies have demonstrated the potential of ML models including SVM, RF, ANN, and ensemble techniques for diabetes prediction, several limitations persist. Many of these models rely on small or imbalanced datasets, lack robust preprocessing (e.g., handling of missing values or outliers), and exhibit limited generalizability due to overfitting or population-specific data. Additionally, the underutilization of DL, particularly CNNs, in structured/tabular clinical data highlights a missed opportunity to capture complex feature interactions. Few studies have explored CNNs with comprehensive data treatment and comparative benchmarking against traditional models. Therefore, there is a clear gap in applying optimized CNN architectures with advanced preprocessing pipelines to improve the reliability, scalability, and clinical relevance of early diabetes prediction models particularly for deployment in telehealth or resource-constrained settings.

MATERIALS AND METHODS

This study proposes a CNN-based diabetes prediction model using data from the PIMA Indians Diabetes study, which includes medical records of 768 women from the PIMA community, comprising 500 non-diabetic and 268 diabetic cases, making it an imbalanced binary classification dataset. The pre-processing phase involved handling missing values using grouped median imputation based on class labels, detecting and treating outliers with the IQR-based boxplot approach, which is followed by grouped median imputation, the SMOTE to rectify class imbalance, and min–max scaling to a [0,1] range for feature normalization. The dataset was utilised to build the testing and training sets. Layers for feature extraction, convolution, dimensionality reduction, pooling, and classification are all part of a convolutional neural network (CNN). A sigmoid activation function, well-suited to binary classification, was used to train the model. Computing time, accuracy, precision, recall, F1-score, AUC-ROC, Brier score, and precision are some of the performance evaluation measures which it was able to calculate probabilities. These metrics provide a comprehensive evaluation of the model's effectiveness, discriminatory performance, and predictive capacity. From utilising datasets to evaluating outcomes, Figure 1 demonstrates the operations in the suggested strategy.

European Journal of Technology
ISSN 2520-0712 (online)
Vol.9, Issue 1, pp1 35 - 50, 2025
www.ajpojournals.org

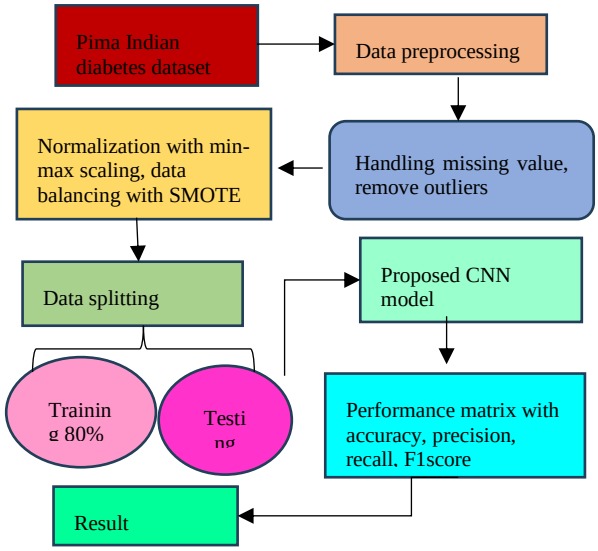

Figure 1: Flowchart of the Proposed CNN-Based Diabetes Mellitus in Healthcare

The steps are explained in the sections that follow, which also include the approach and proposed flowchart.

Data Collection

The PIMA dataset is freely accessible for research purposes and may be obtained from the UCI ML Repository. Essentially, "the majority class is nondiabetic (negative), while the minority class is diabetic (positive)." The dataset is imbalanced, consisting of medical test results from 768 individuals; 268 samples are from people with diabetes, and 500 samples are from people without the disease. The binary class label dataset known as PIMA uses a "1" for a positive result and a "0" for a negative one.

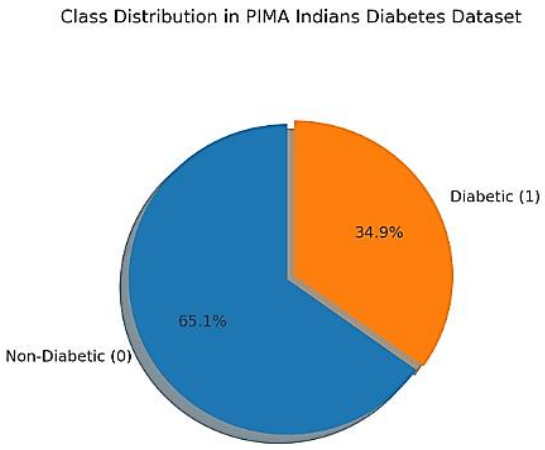

Figure 2: Unbalanced Class for PIMA Dataset.

In the PIMA Indians Diabetes dataset, class '0' (non-diabetic) accounts for 65.1% of the data, whereas class '1' (diabetic) accounts for 34.9%, as shown in Figure 2. This notable class disparity emphasizes the necessity of suitable pre-processing methods, such the SMOTE, to guarantee balanced learning and enhance the model's capacity to precisely forecast cases of diabetes.

Data Preprocessing

The data preparation phase of this effort involved several crucial steps to ensure the quality and balance of the PIMA diabetes dataset in India before training the model. The data pre-treatment methods employed in this work addressed oversampling, data normalisation, handling missing values, and identifying and correcting outliers. The purpose of the particular preparation processes is to guarantee data quality so that the trained model is free of biases:

- **Handling missing values:** To impute missing values, they use the median value of a particular group for each group. As a result, crucial data is not lost, and the overall distribution of the data is maintained.
- **Remove outliers:** Numbers of data points below. The "upper bound" (Q3 + 1.5 IQR) and the "lower bound" (Q1 − 1.5 IQR) are called outliers by people. We used grouped median imputation to handle outliers, if found, in a manner that wouldn't affect the overall data distribution.

Normalization with Min-Max Scaling

The range of all the characteristics was normalized using the min–max normalization procedure. Equation (1) specifically provides the normalization formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \ldots\ldots (1)$$

Here, X represents the original feature value, X_{min} *and* X_{max} Nd are the minimum and maximum feature values, respectively.

Data Balancing with SMOTE

The significant class disparity between patients with diabetes and those without the disease may have caused the model to become biased in favour of they see the majority class throughout training. They used SMOTE to solve this problem. Interpolating between current minority class samples and their nearest neighbours allows for the generation of synthetic minority class samples, which balance out the dataset's class distribution. This lessened bias in the model training process and the imbalance between samples with and without diabetes.

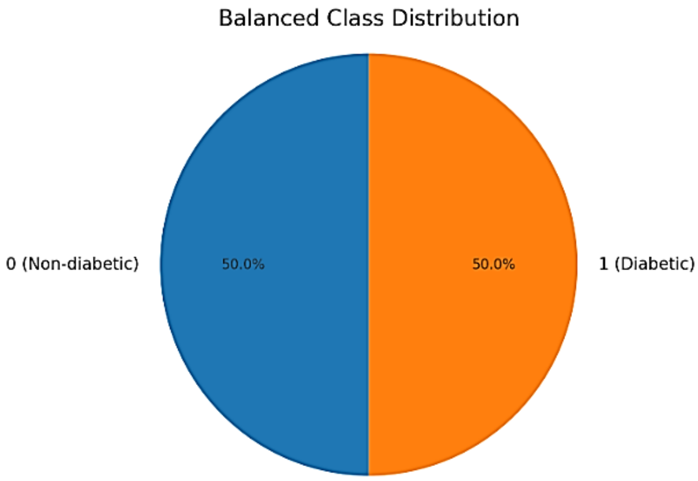

Figure 3: Balanced Class Distribution for PIMA Dataset

Figure 3 illustrates a balanced class distribution after applying the SMOTE method. In this scenario, people with diabetes (class 1) and those without (class 0) are split evenly, at fifty per

cent. In order to avoid model bias during training and guarantee equitable categorization performance, this balance is essential.

Data Splitting

Parts of the dataset were split into training and testing sets to train the models and assess their performance with new data. Eighty per cent of the data points in the dataset were used for training the model, while twenty per cent were used for testing each iteration.

Classification of Proposed CNN Model

A subtype of deep learning models, called CNN, was developed specifically to process and analyse grid-like input, including image data. They have transformed picture classification and several other computer vision applications due to their ability to automatically recognise hierarchical structures in raw image data. A CNN consists of three layers: a fully connected layer, a pooling layer, and a convolutional layer. Figure. 4 shows all of the layers combined.

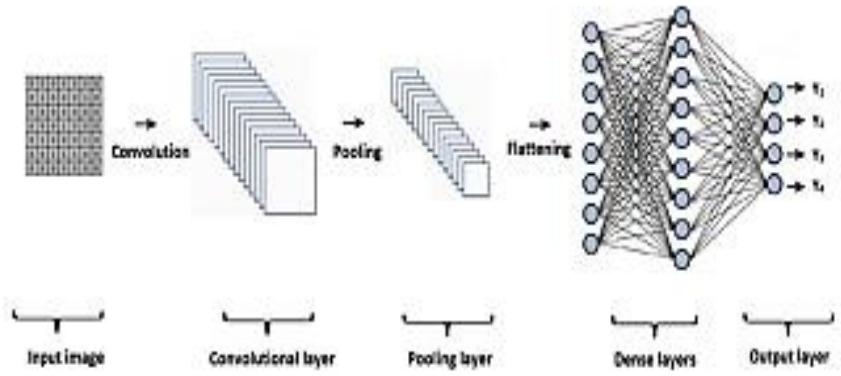

Figure 4: CNN Architecture

A CNN architecture is illustrated in Figure 4, where the input picture is flattened, processed by dense layers, and then passed through convolution and pooling layers to obtain the final output prediction. The CNN layers listed below are:

- Convolutional layer: applies a filter to the pictures and simultaneously scans a large number of pixels to produce an activation map.
- Pooling layer: Reduces the amount of data generated by the convolutional layer to increase storage efficiency.
- Fully Connected Layer: An input layer that is fully connected; the input for the next stage is a single vector that has been "flattened" from the output of the preceding layers. The first fully connected layer uses feature analysis inputs and assigns weights to them in order to anticipate the proper label. For each label, the ultimate probability is provided by the fully connected output layer.

Performance Matrix

The effectiveness of categorisation methods for predicting diabetes accuracy is tested using the Pima Indian Diabetes dataset. The results are compared with the predicted and actual clinical labelling. In the end, several evaluation measures and computation time are employed to fully assess the model's efficacy, including accuracy, sensitivity TPR, specificity TNR, precision, recall, F1score, AUC, and ROC curve. Other attributes of the model's performance are measures that quantify additional aspects, such as forecast accuracy, classification ability, and general

dependability. The evaluation framework's four fundamental components TP, FP, TN, and FN form its backbone and serve as a basis for quantifying the precision of in other words, it can help the model to distinguish between cases with and without diabetes. The below covers some metrics:

Accuracy: Finding the answer to Equation (2) is as simple as subtracting the percentage of correct guesses from the total number of predictions:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{..... (2)}$$

Precision: The positive case percentage, or the proportion of diabetic patients whose diagnoses are accurate, is determined by dividing the ratio of TP by the sum of TP and FP in Equation (3).

$$\text{Precision} = \text{TP} / \text{TP} + \text{FP} \quad \text{..... (3)}$$

Recall: A patient's proportion of diabetes is determined by dividing their TP by their total TP and FN, or positive cases, who are accurately diagnosed as having the disease. Equation (4) presents it numerically:

$$\text{Recall} = \text{TP} / \text{TP} + \text{FN} \quad \text{..... (4)}$$

F1_Score: The total number of recall and accuracy, weighted together, is this. For this reason, this score considers both incorrect negatives and erroneous positives, as indicated in Equation (5):

$$F1 = 2 * \frac{(Precision*Recall)}{(Precision+Recall)} \quad \text{..... (5)}$$

FINDINGS

In the sequence of the experimental circumstances, the experimental results and the accompanying comments will be provided. The system consisted of an Intel Core i9-13900K CPU with a base frequency of 3.0 GHz, an NVIDIA RTX 4090 GPU with 24GB of VRAM, and 64 of DDR5 RAM. Windows 11 Pro was the operating system that was used. The following results of the proposed models, discussed below in terms of performance measures — loss and accuracy— for diabetes prediction are presented.

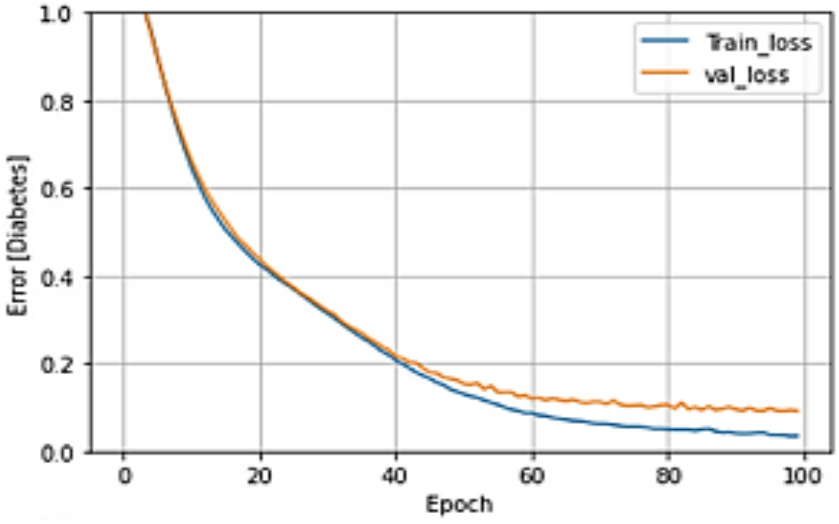

Figure 5: Proposed CNN Performance On Loss Graph

Figure 5 displays the loss curves for training and validating a diabetes prediction model across 100 epochs. The model is learning and generalising well, as evidenced by the declining trend in

the validation loss (orange line) and training loss (blue line). Both losses are considerable at start, but they decrease significantly within the first 20 epochs before gradually declining. The training loss continues to decline, while the validation loss peaks at epoch 50, indicating mild overfitting. However, there is not much of a difference between the two, indicating that the model is well-trained and has strong generalisation ability.

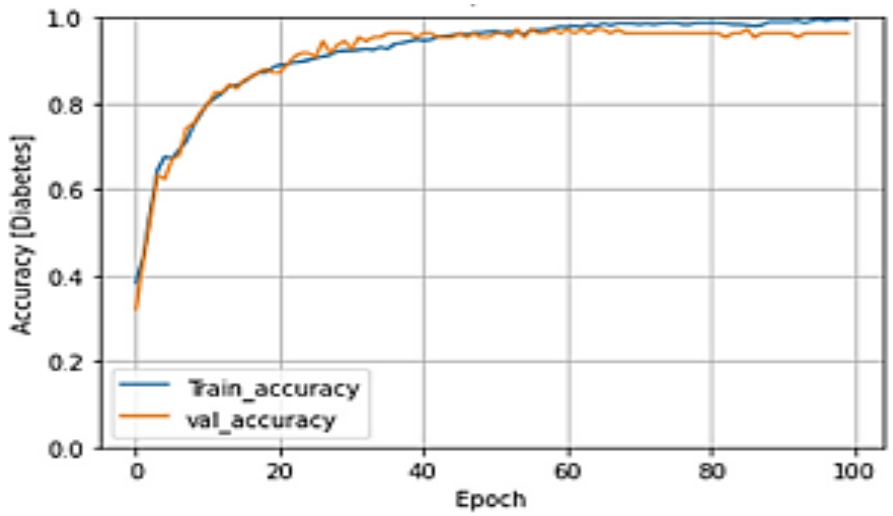

Figure 6: Proposed CNN Performance on Accuracy Graph

The training and validation accuracy curves of a diabetes prediction model across 100 evolutions are shown in Figure 6. Training accuracy (blue line) and validation accuracy (orange line) both exhibit fast progress up to around epoch 20. Following that, their accuracy remains at or over 90%. As training continues, the training accuracy has also improved, approaching 100%, and the validation accuracy remains at a similar level. The curves remain close together, indicating high generalisation performance and little overfitting. In this case, the model can predict data that is not immediately apparent with high accuracy, as it excels at identifying the underlying patterns in the data.

Table 2: Performance Matrix for Proposed CNN model

Matrix	CNN
accuracy	96.66
precision	97.34
recall	96.66
F1 score	96.96

Table II demonstrates that the suggested CNN model performs well in diabetes prediction, with an accuracy of 96.66%. Additionally, the precision was achieved to 97.34%, recall to 96.66% and F1 to 96.96%. These findings demonstrate the model's effectiveness in enabling early diabetes diagnosis by accurately detecting diabetic patients, achieving a balanced performance in terms of accuracy and recall.

Comparative Analysis

In this part, we examine using the Pima Indian Diabetes dataset, and assess and contrast several ML approaches to diabetes prognosis. The proposed CNN model is unique in that it learns and identifies complex data representations with extreme accuracy, thereby distinguishing key risk factors. On the other hand, models such as, ANN [13] and RF [10] exhibited comparatively lower performance. As summarised in Table III, the CNN model demonstrated superiority in all

key performance metrics, including F1 score, memory, accuracy, and precision, which are strong and may be helpful in a trustworthy clinical setting for diabetes diagnosis.

Table 3: Comparison between Base and Proposed Model for Diabetes Prediction

Model	Accuracy
ANN[13]	82.35
Random Forest [10]	90.29
CNN	96.66

The proposed CNN model achieves superior performance, with the best accuracy of 96.66% in Table III, for diabetes prediction. We observed that the RF model performed equally well as the CNN, with an accuracy of 90.29%, although it still lagged behind the CNN. Among the three ANN models, the one recorded the lowest accuracy at 82.35%. The results reveal that CNN model is more suited to accurately detect diabetes than both ANN and RF. The proposed CNN model for diabetes prediction has several benefits. It demonstrates outstanding overall performance and dependability, characterised by high accuracy and a great F1 score. With a high recall, the model accurately detects the majority of real instances of diabetes, and its high accuracy reduces the likelihood of FP, which would classify people with diabetes as healthy. Additionally, without requiring extensive human feature engineering, many medical diagnosis occupations benefit greatly from CNNs' ability to automatically recognise and understand complex patterns in the data. Additionally, it scales effectively to large healthcare datasets.

CONCLUSION AND RECOMMENDATIONS

Conclusion

A CNN-based model, as presented in this study, demonstrates that the PIMA Indian diabetes dataset can accurately predict the onset of diabetes mellitus at an early stage with a prediction rate of 96.66%. Diabetes is a chronic illness caused by either a deficiency in insulin synthesis and activity (Type 2 Diabetes, T2D) or an inability to produce insulin (Type 1 Diabetes, T1D). We employed robust pre-processing techniques to address challenges such as handling missing values, removing outliers, addressing class imbalance using SMOTE, and applying min–max normalisation. The model's precision (97.34%), recall (96.66%), and F1 score (96.96%) all demonstrated strong prediction performance. According to the results, this strategy has the potential to be included in telemonitoring healthcare systems that offer prompt and remote diabetes diagnosis.

Recommendations

The study is limited by its reliance on a single dataset (PIMA Indians Diabetes), which may impact the effectiveness of the approach with different demographics. Additionally, we still need to assess how well the model performs in real-world clinical settings. To increase precision and clinical applicability, the model may be extended to incorporate real-time data from wearable medical devices, evaluated on bigger and more varied datasets, and coupled with other DL architectures in further research.

REFERENCES

[1] M. Fogelholm *et al.*, "PREVIEW: Prevention of diabetes through lifestyle intervention and population studies in Europe and around the world. Design, methods, and baseline participant description of an adult cohort enrolled into a three-year randomised clinical trial," *Nutrients*, 2017, doi: 10.3390/nu9060632.

[2] V. Kolluri, "An In-Depth Exploration of Unveiling Vulnerabilities: Exploring Risks in AI Models And Algorithms," *Int. J. Res. Anal. Rev.*, 2014.

[3] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Current techniques for diabetes prediction: Review and case study," *Applied Sciences (Switzerland)*. 2019. doi: 10.3390/app9214604.

[4] V. KOLLURI, "Machine Learning in Managing Healthcare Supply Chains: How Machine Learning Optimizes Supply Chains, Ensuring the Timely Availability of Medical Supplies," *Int. J. Emerg. Technol. Innov. Res.*, pp. 2349–5162, 2016.

[5] M. N. Sohail, R. Jiadong, M. U. Muhammad, S. T. Chauhdary, J. Arshad, and A. J. Verghese, "An accurate clinical implication assessment for diabetes mellitus prevalence based on a study from Nigeria," *Processes*, 2019, doi: 10.3390/pr7050289.

[6] V. Kolluri, "A Pioneering Approach to Forensic Insights: Utilization AI For Cybersecurity Incident Investigations," *Int. J. Res. Anal. Rev.*, vol. 3, no. 3, pp. 2348–1269, 2016.

[7] G. Alfian, M. Syafrudin, M. F. Ijaz, M. A. Syaekhoni, N. L. Fitriyani, and J. Rhee, "A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing," *Sensors (Switzerland)*, 2018, doi: 10.3390/s18072183.

[8] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," in *1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings*, 2019. doi: 10.1109/UBMYK48245.2019.8965556.

[9] R. Deo and S. Panigrahi, "Performance Assessment of Machine Learning Based Models for Diabetes Prediction," in *2019 IEEE Healthcare Innovations and Point of Care Technologies, HI-POCT 2019*, 2019. doi: 10.1109/HI-POCT45284.2019.8962811.

[10] M. T. Islam, M. Raihan, F. Farzana, M. G. M. Raju, and M. B. Hossain, "An Empirical Study on Diabetes Mellitus Prediction for Typical and Non-Typical Cases using Machine Learning Approaches," in *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*, 2019. doi: 10.1109/ICCCNT45670.2019.8944528.

[11] P. Kaur, N. Sharma, A. Singh, and B. Gill, "CI-DPF: A Cloud IoT-based Framework for Diabetes Prediction," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018*, 2018. doi: 10.1109/IEMCON.2018.8614775.

[12] R. A. Alassaf *et al.*, "Preemptive Diagnosis of Diabetes Mellitus Using Machine Learning," in *21st Saudi Computer Society National Computer Conference, NCC 2018*, 2018. doi: 10.1109/NCG.2018.8593201.

[13] S. K. Dey, A. Hossain, and M. M. Rahman, "Implementation of a Web Application to

Predict Diabetes Disease: An Approach Using Machine Learning Algorithm," in *2018 21st International Conference of Computer and Information Technology, ICCIT 2018*, 2018. doi: 10.1109/ICCITECHN.2018.8631968.

[14] Polu, A. R., Buddula, D. V. K. R., Narra, B., Gupta, A., Vattikonda, N., & Patchipulusu, H. (2021). Evolution of AI in Software Development and Cybersecurity: Unifying Automation, Innovation, and Protection in the Digital Age. *Available at SSRN 5266517*.

[15] Polu, A. R., Vattikonda, N., Buddula, D. V. K. R., Narra, B., Patchipulusu, H., & Gupta, A. (2021). Integrating AI-Based Sentiment Analysis With Social Media Data For Enhanced Marketing Insights. *Available at SSRN 5266555*.

[16] Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., Polu, A. R., Vattikonda, N., & Gupta, A. K. Advanced Edge Computing Frameworks for Optimizing Data Processing and Latency in IoT Networks.

[17] Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., Polu, A. R., Narra, B., & Gupta, A. K. Predictive Analytics in E-Commerce: Effective Business Analysis through Machine Learning.

[18] Kalla, D. (2022). AI-Powered Driver Behavior Analysis and Accident Prevention Systems for Advanced Driver Assistance. *International Journal of Scientific Research and Modern Technology (IJSRMT) Volume*, *1*.

[19] Kuraku, D. S., Kalla, D., & Samaah, F. (2022). Navigating the link between internet user attitudes and cybersecurity awareness in the era of phishing challenges. *International Advanced Research Journal in Science, Engineering and Technology*, *9*(12).

[20] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2022). Enhancing Early Diagnosis: Machine Learning Applications in Diabetes Prediction. *Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-205. DOI: doi. org/10.47363/JAICC/2022 (1)*, *191*, 2-7.

[21] Kalla, D., Kuraku, D. S., & Samaah, F. (2021). Enhancing cyber security by predicting malwares using supervised machine learning models. *International Journal of Computing and Artificial Intelligence*, *2*(2), 55-62.

[22] Katari, A., & Kalla, D. (2021). Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies. *ESP Journal of Engineering & Technology Advancements (ESP-JETA)*, *1*(1), 150-157.

[23] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2021). Facial Emotion and Sentiment Detection Using Convolutional Neural Network. *Indian Journal of Artificial Intelligence Research (INDJAIR)*, *1*(1), 1-13.

[24] Chinta, P. C. R., Katnapally, N., Ja, K., Bodepudi, V., Babu, S., & Boppana, M. S. (2022). Exploring the role of neural networks in big data-driven ERP systems for proactive cybersecurity management. *Kurdish Studies*.

[25] Routhu, K., Bodepudi, V., Jha, K. M., & Chinta, P. C. R. (2020). A Deep Learning Architectures for Enhancing Cyber Security Protocols in Big Data Integrated ERP Systems. *Available at SSRN 5102662*.

[26] Chinta, P. C. R., & Katnapally, N. (2021). Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures. *Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures*.

[27] Katnapally, N., Chinta, P. C. R., Routhu, K. K., Velaga, V., Bodepudi, V., & Karaka, L. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. *American Journal of Computing and Engineering*, *4*(2), 35-51.

[28] Chinta, P. C. R. (2022). Enhancing Supply Chain Efficiency and Performance Through ERP Optimisation Strategies. *Journal of Artificial Intelligence & Cloud Computing*, *1*(4), 10-47363.

[29] Sadaram, G., Sakuru, M., Karaka, L. M., Reddy, M. S., Bodepudi, V., Boppana, S. B., & Maka, S. R. (2022). Internet of Things (IoT) Cybersecurity Enhancement through Artificial Intelligence: A Study on Intrusion Detection Systems. *Universal Library of Engineering Technology*, (2022).

[30] Karaka, L. M. (2021). Optimising Product Enhancements Strategic Approaches to Managing Complexity. *Available at SSRN 5147875*.

[31] Jha, K. M., Bodepudi, V., Boppana, S. B., Katnapally, N., Maka, S. R., & Sakuru, M. Deep Learning-Enabled Big Data Analytics for Cybersecurity Threat Detection in ERP Ecosystems.

[32] Chandrasekaran, A., & Kalla, D. (2023). Heart disease prediction using chi-square test and linear regression. *Computer Science & Information Technology*, *13*, 135-146.

[33] Kalla, D., & Kuraku, S. (2023). Phishing website url's detection using nlp and machine learning techniques. *Journal of Artificial Intelligence*, *5*, 145.

[34] Kuraku, D. S., & Kalla, D. (2023). Impact of phishing on users with different online browsing hours and spending habits. *International Journal of Advanced Research in Computer and Communication Engineering*, *12*(10).

[35] Kuraku, S., Kalla, D., Samaah, F., & Smith, N. (2023). Cultivating proactive cybersecurity culture among IT professional to combat evolving threats. *International Journal of Electrical, Electronics and Computers*, *8*(6).

[36] Kuraku, D. S., Kalla, D., Smith, N., & Samaah, F. (2023). Exploring How User Behavior Shapes Cybersecurity Awareness in the Face of Phishing Attacks. *International Journal of Computer Trends and Technology*.

[37] Chinta, P. C. R. (2023). Leveraging Machine Learning Techniques for Predictive Analysis in Merger and Acquisition (M&A). *Journal of Artificial Intelligence and Big Data*, *3*(1), 10-31586.

[38] Kuraku, D. S., Kalla, D., Smith, N., & Samaah, F. (2023). Safeguarding FinTech: elevating employee cybersecurity awareness in financial sector. *International Journal of Applied Information Systems (IJAIS)*, *12*(42).

[39] Moore, C. (2023). AI-powered big data and ERP systems for autonomous detection of cybersecurity vulnerabilities. *Nanotechnology Perceptions*, *19*, 46-64.

[40] Chinta, P. C. R. (2023). The Art of Business Analysis in Information Management Projects: Best Practices and Insights. *DOI*, *10*.

[41] Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. *J Contemp Edu Theo Artific Intel: JCETAI-104*.

[42] Maka, S. R. (2023). Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. *Available at SSRN 5116707*.

[43] Routhu, KishanKumar & Katnapally, Niharika & Sakuru, Manikanth. (2023). Machine Learning for Cyber Defense: A Comparative Analysis of Supervised and Unsupervised Learning Approaches. Journal for ReAttach Therapy and Developmental Diversities. 6. 10.53555/jrtdd.v6i10s(2).3481.

[44] Chinta, Purna Chandra Rao & Moore, Chethan Sriharsha. (2023). Cloud-Based AI and Big Data Analytics for Real-Time Business Decision-Making. 36. 96-123. 10.47363/JAICC/2023.

[45] Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. *J Contemp Edu Theo Artific Intel: JCETAI-104*.

[46] Bodepudi, V. (2023). Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. *Journal of Artificial Intelligence and Big Data*, *3*(1), 10-31586.

License