OPEN ACCESS

Ghazala HS (2023) 135
J Contemp Edu Theo Artific Intel: 2023(3):100103-1
DOI: 10.71010/2023/JCETAI-e103-1

Research Article

Journal of Contemporary Education Theory & Artificial Intelligence

Artificial Intelligence-Based ML Techniques and Big Data Analytics to Precision Diabetes Diagnosis in Healthcare Systems

Rahul Vadisetty1*, Purna Chandra Rao Chinta2, Chethan Sriharsha Moore3, Laxmana Murthy Karaka4, Manikanth Sakuru5, Varun Bodepudi6, Srinivasa Rao Maka7, Srikanth Reddy Vangala8

[1]Wayne State University, Master of Science, rahulvy91@gmail.com
[2]Microsoft, Support Escalation Engineer, chpurnachandrarao@gmail.com
[3]Microsoft, Support Escalation Engineer, moore.chethan1@outlook.com
[4]Code Ace Solutions Inc, Software Engineer, laxmanakaraka18@gmail.com
[5]JP Morgan Chase, Lead Software Engineer, manikanth.sakuru@gmail.com
[6]Applab System Inc, Computer Programmer, bvarunklu@gmail.com
[7]North Star Group Inc, Software Engineer, makasrinu@gmail.com
[8]University of Bridgeport, Computer Science Dept, srikanthreddy1043@microsoft.com

[*]**Corresponding author:** Rahul Vadisetty, Wayne State University, Master of Science, rahulvy91@gmail.com

Citation: Rahul V, Purna Chandra Rao C, Chethan Sriharsha M, Laxmana Murthy K, Manikanth S, et al. (2023) Artificial Intelligence-Based ML Techniques and Big Data Analytics to Precision Diabetes Diagnosis in Healthcare Systems. J Contemp Edu Theo Artific Intel: JCETAI-103-1.

Received Date: November 05, 2023; **Accepted Date:** November 15, 2023; **Published Date:** November 22, 2023

Abstract

Diabetes is one of the most common and deadly metabolic illnesses, affecting millions of people around the world. It is very important to get a correct diagnosis as soon as possible to escape serious problems like heart disease, kidney failure, and neuropathy. Even though a lot of studies have been done and diagnostic methods have gotten better, it still need for more accurate and useful ways to find people with diabetes. This study tests how well AI-driven models can diagnose diabetes using the PIMA Dataset. Many different types of classification models were tested using F1-score, recall, accuracy, and precision to measure their success. These models include Bayesian networks, Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), and Logistic Regression (LR). 97.49% accuracy, 96.71% precision, 95.59% recall, and an F1-score of 95.42% show that the CNN model outperforms the other models in diabetes case categorization, with low false positives and negatives. In comparison, the DNN achieved an accuracy of 83.41%, LR attained 77.25%, and Bayesian Networks scored the lowest with 73.83% accuracy. The CNN model also demonstrated superior performance in all evaluation metrics, highlighting its potential as a reliable tool for diabetes diagnosis. These findings suggest that CNN-based models provide highly accurate and effective predictions for diabetes detection, offering a valuable contribution to early diagnosis and healthcare decision-making.

Keywords: Diabetes Diagnosis, Healthcare Systems, Machine Learning, Convolutional Neural Networks, Big Data Analytics, Deep Learning.

1. Introduction

The combination of medical devices and communication tools has caused huge changes in healthcare systems. Remote healthcare sensor nodes take vital signs from patients and store the information in medical files. These files are then sent to the cloud so that medical experts can access them and store them there [1]. These health care tools can be used for a lot of different things, like treating diabetes and high blood pressure and helping people get better. However, the incidence and Increased urbanization, obesity, ageing populations, and a decline in physical activity levels have all been linked to the incidence of diabetes mellitus globally. Approximately 600 million people will have diabetes by 2030, according to a WHO estimate, making it the ninth most deadly illness [2]. Diabetes mellitus is often linked to very bad outcomes, like going blind, having brain disease, or having complex coronary heart disease. Docs must follow the right clinical standards when diagnosing and treating diabetes because it is a serious and complicated disease. Only then can they give their patients the right diagnosis and care.

Diabetes is a group of metabolic diseases. High blood sugar is a sign of diabetes. This happens when insulin doesn't work properly or isn't made enough, or both. High blood sugar caused by diabetes can damage, malfunction, or fail many systems over time, including the kidneys, eyes, nerves, heart, and blood vessels. Diabetes arises due to a variety of pathological causes [3]. These include problems that make the body resistant to insulin and the death of pancreatic b-cells by the immune system, which causes a lack of insulin. Insulin not working properly in target organs is the main reason why diabetics' protein, lipid, and glucose metabolism is off [4][5]. A lack of insulin or decreased tissue responses to insulin at one or more places along the complex hormone action pathways are the two main reasons why insulin doesn't work well. As many as one patient can have problems with both insulin secretion and activity.

Big data in healthcare, which consists of vast and complex electronic health records, requires advanced analytics for effective processing and interpretation [6]. ML algorithms play a crucial role in healthcare analytics by extracting meaningful insights from medical datasets, enabling early disease prediction, accurate diagnosis, and personalized treatment plans [7]. In the context of diabetes, ML techniques, including predictive modeling and quantitative analysis, enhance

diagnosis by identifying critical patterns within large datasets. AI-powered models such as SVM, DNN, CNN, NB, RF, and MLP offer high accuracy in handling complex and nonlinear diabetes data [8]. These AI-driven approaches enable precision diagnosis by efficiently processing vast datasets, identifying hidden patterns, and improving early detection. By integrating real-time patient data with advanced analytics, big data enhances diabetes prediction, monitoring, and long-term disease management, providing healthcare systems with the tools to optimize treatment strategies and ultimately improve patient outcomes.

A. Motivation and Contribution of the Study

This research stems from the remarkable rise in diabetes frequency, which represents a serious global health crisis. In order to avoid problems and enhance patient outcomes, early diagnosis and precise prediction models are essential. Traditional diagnostic approaches, while effective, often face challenges related to accuracy, efficiency, and scalability. This work uses ML-based models to improve diabetes categorization and prediction in order to get beyond these restrictions, offering a more automated and dependable method for early detection. This study's primary contribution is stated below:

- Using the PIMA Indian Diabetes Dataset to correctly identify diabetes and paying close attention to how the data is prepared can help the model work better.
- Preprocessing methods are used, such as addressing missing numbers, eliminating outliers, and improving data for increased accuracy.
- Normalization of feature values through Min-Max scaling to enhance model convergence and performance.
- Application of AI-based classification models, such as Bayesian Networks, CNN, DNN, and LR for precise diabetes prediction.
- Scalable machine learning models are investigated to increase diagnostic effectiveness, and metrics such as F1-score, recall, accuracy, and precision are used to evaluate the models' performance.

B. Justification and Novelty

This study justifies its approach by utilizing the PIMA dataset, ensuring practical applicability to diabetes diagnosis. The dataset undergoes a rigorous preprocessing pipeline, addressing key challenges such as missing values and outliers through techniques like data imputation, outlier removal, and Min-Max feature scaling. This study's originality lays in its integration of feature standardization and refinement strategies to enhance model efficiency, coupled with the application of CNN, LR, DNN, and Bayesian Networks for improved diabetes prediction. Unlike traditional methods, this study leverages AI-driven models to optimize classification accuracy, ensuring scalability and robustness for real-world medical applications.

C. Structure of the paper

This paper is structured in the following manner: Part II provides background information on diabetes diagnosis, Part III describes the study's methodology, Part IV shows the results of the experiments and the performance evaluation of the models, and Part V offers suggestions for future lines of inquiry.

2. Literature Review

This section reviews the body of research on diabetes diagnosis detection and categorization. Most of the evaluated research focuses on classification methods. Some of the key reviews are:

Agarwal and Saxena (2019) Millions of individuals worldwide suffer from diabetes, and women make up over half of those who have the disease. Although diabetes is relatively widespread, ML has been used in many areas of healthcare. The Pima Indians Diabetes Dataset includes data from Pima women. Comparing the various algorithms to find the most accurate one is the primary objective. SVM, KNN, NB, DT, and LR are the algorithms that are being compared. Using K-Fold and Cross Validation, managed to get an accuracy of 81.1% [9].

Rahman et al. (2019) proposed a study of one chronic condition that is characterized by Type 2 Diabetes Mellitus (T2DM), glucose homeostasis is disrupted. They used diabetic data and many cutting-edge ML methods, including RF, SVM, DT, and NB. The use of modern Bayesian optimisation (BO) has been proposed to optimise the hyper-parameters of machine learning classifiers for diabetic mellitus (DM). The NB classifier's accuracy 73.96%, the DT's 71.61%, the SVM's 76.04%, and the RF's 77.60% were the hyperparameters that were optimized using BO and without BO-optimized SVM, it reached 64.06% accuracy [10].

Yahyaoui et al. (2019) a DSS for diabetes prediction using ML. They contrasted traditional ML methods with DL methodologies. 768 samples, each with eight attributes, from the publicly accessible PIDD were used to test the suggested approach. 268 people had diabetes, whereas 500 samples were classified as non-diabetic. DL, SVM, and RF have respective overall accuracy of 76.81%, 65.38%, and 83.67% [11].

Gokulprasanth, Raja Kumari, and Kathiroli Diabetes is a long-lasting disease. Two different kinds of diabetes exist: Type 1 and Type 2. Unusually high blood sugar levels are a sign of a metabolic disease called diabetes mellitus (DM). There were about 425 million diabetics in the globe as of November 2017. The Pima Indians Diabetes Dataset's diabetic patients utilize an ANN that is self-adaptive. It automatically determines how many nodes are concealed and adjusts the number of hidden nodes and connection weights [12].

G., R. and K.P. (2018) Early diabetes testing is crucial for prompt treatment, which can prevent the condition from developing into severe problems. Heart rate variability (HRV) data, which are RR-interval signals obtained from electrocardiogram (ECG) readings, can be used to identify diabetes non-invasively. SVMs are fed these features in order to classify them. Their CNN and CNN-LSTM architectures have shown a performance gain of 0.03% and 0.06%, respectively, above their previous work without the use of SVM. With an extremely high accuracy of 95.7%, the suggested classification method can assist doctors in diagnosing diabetes using ECG data [13].

Kumar and Pranavi (2017) Cloud computing and big data are essential for solving healthcare issues. The amount of healthcare data is increasing dramatically every day these days, necessitating a quick, practical, and economical way to reduce the death rate. This research looks at and evaluates a lot of machine learning algorithms to see which one performs best in terms of accuracy, kappa, precision, recall, sensitivity, and specificity. A comprehensive analysis of the diabetes dataset is conducted using the RF, SVM, k-NN, CART, and LDA methods [14].

Rusanov, Prado and Weng (2016), laboratory data from diabetes patients may be applicable to other illnesses. Using glycated hemoglobin (HbA1C) discrete wavelet transforms, 2,365 diabetics were categorized. HbA1C trends were categorized using latent class growth analysis. ICD-9 codes, blood glucose, and creatinine were used to compare the clusters, and their stability was assessed. Out of the 'uncontrolled group', 98.9% had an average HbA1C of >7%, and 74.0% had an average of 154 mg/dL [15].

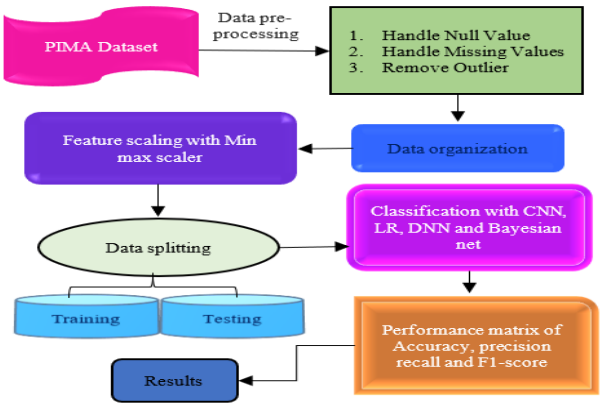
Table I presents a comparative study of the background studies based on their conclusions, limitations, and future research.

Table 1: Comparative Analysis of Machine Learning Approaches for diabetes diagnosis detection.

Author	Dataset	Methodology	Findings	limitation	Future work
Agarwal & Saxena (2019)	Pima Indians Diabetes Dataset	Compared Decision Trees, Logistic Regression, Naïve Bayes, SVM, and KNN using K-Fold Cross Validation	Achieved 81.1% accuracy using KNN	Limited to Pima dataset, lacks external validation	Test on diverse datasets and improve feature selection
Rahman et al. (2019)	Diabetes Mellitus dataset	Used Random Forest, SVM, Decision Tree, Naïve Bayes with Bayesian Optimization (BO)	RF (77.6%), SVM (76.04%), DT (71.61%), NB (73.96%)	No comparison with deep learning methods	Implement deep learning for further accuracy improvement
Yahyaoui et al. (2019)	Pima Indians Diabetes Dataset	Compared SVM, RF with CNN-based Deep Learning	CNN (76.81%), SVM (65.38%), RF (83.67%)	Small dataset size, limited feature engineering	Expand dataset and explore hybrid ML-DL models
Kathiroli, RajaKumari & Gokulprasanth (2018)	Pima Indians Diabetes Dataset	Self-adaptive ANN with cascade correlation algorithm	Improved classification of diabetic and non-diabetic cases	Lack of benchmark comparisons with other ML models	Test ANN performance with real-world clinical data
G., R. & K.P. (2018)	HRV signals from ECG data	LSTM, CNN, CNN-LSTM combined with SVM	CNN-LSTM improved accuracy to 95.7%	Limited dataset, focus only on ECG-based detection	Extend to multimodal diabetes detection with other biomarkers
Kumar & Pranavi (2017)	Big Data and Cloud-based Diabetes Data	Analyzed RF, SVM, k-NN, CART, LDA for diabetes prediction	Compared various metrics such as accuracy, precision, recall	Computational complexity of handling large-scale data	Optimize ML models for real-time cloud-based diagnosis
Rusanov, Prado & Weng (2016)	Diabetes patients' laboratory data	Used Discrete Wavelet Transform (DWT) and Latent Class Growth Analysis	Clustered 2,365 diabetic patients based on HbA1C trends	Focused on clustering only, no predictive modeling	Extend to real-time monitoring and risk stratification

3. Methodology

This study's objective is to evaluate diabetes detection algorithms powered by AI. The following steps of research design are shown in the Figure 1 flowchart. The PIMA dataset collection is the first step in the organized approach used by the suggested methodology for diabetes diagnosis. After handling null values, controlling missing values, and getting rid of outliers to improve the quality of the data, preprocessing can begin. The data set is sorted before the features are normalised using Min-Max scaling. For efficient model learning and evaluation, the data is pre-processed before it is split into training and testing sets. CNN, Bayesian networks, DNN, and LR are among the categorization methods used to forecast diabetes. These models are evaluated using F1-score metrics, accuracy, and precise recall in a performance matrix. By analyzing the final data, the optimal model for diabetes diagnosis is found.



The following stages are included in the diabetes diagnostic flowchart and are shown below:

A. Data Collection

The NIDDK analyzed 768 female patients aged 21 and older utilizing the UCI ML Repository, the PIMA Indian Diabetes dataset. Of the 500 individuals who do not have diabetes, 268 have positive diabetes cases. Pregnancy (F1), birth frequency and glucose measures (F2), plasma glucose levels following the glucose tolerance test and blood pressure (F3), skin thickness (F4), and triceps skin fold measurements are the eight characteristics included in the PIMA Indian Diabetes dataset. Other characteristics are age (F8), diabetes pedigree function (F7), BMI (F6), and insulin (F5). This dataset is widely recognized as a benchmark for diabetes prediction due to its reliability and extensive use in research. Some of the visualizations are as follows:

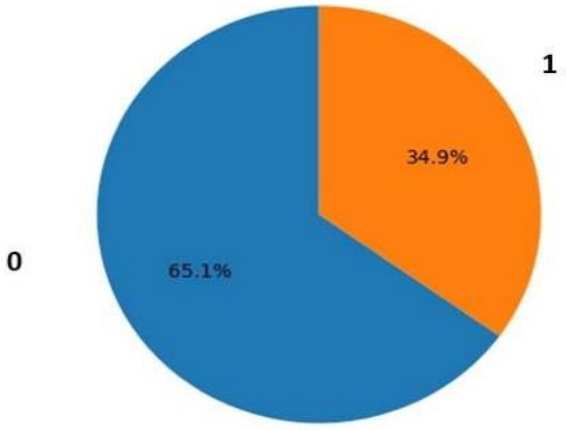


Figure 2: Pie Chart Distribution using PIMA Dataset

The distribution of the outcome variable in The PIMA dataset is utilized to forecast the likelihood of developing diabetes, is shown in a pie chart in Figure 2. "1" indicates people with diabetes, whereas "0" indicates those without the disease. The outcome variable is a binary categorization. The pie chart presents the percentage contribution of each category, indicating that 500 subjects (65.1%) are non-diabetic, whereas 268 subjects (34.9%) are diabetic. This distribution highlights an imbalanced dataset, with a majority class of non-diabetic cases.

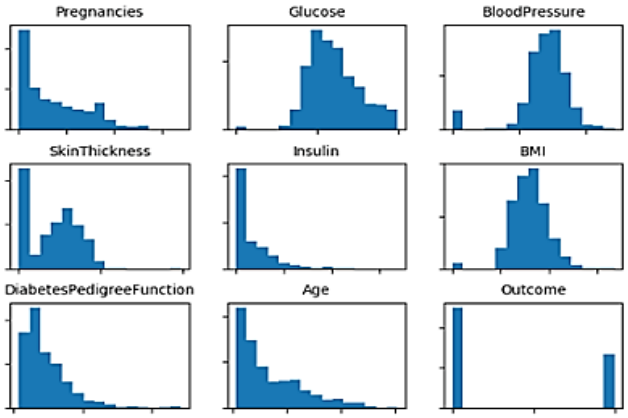


Figure 3: Histogram Data for Selected Features

Figure 3 illustrates histograms representing the distribution of various features in the PIDD. These include age, skin thickness, blood pressure, glucose, insulin, function, result, and diabetes pedigree. It is worth mentioning that Insulin and Diabetes Pedigree Function are among the factors with right-skewed distributions, suggesting a concentration of lower values. The histograms display the frequency distribution of each feature. In contrast, features like BMI and Blood Pressure appear more normally distributed. The Outcome variable, representing diabetes diagnosis (0 for non-diabetic, 1 for diabetic), shows a class imbalance, which can impact model performance.

A. Data Preprocessing

The initially model contains various inconsistencies, including Handle null value, missing values, and remove outlier, which can impact model performance. To address these issues, data preprocessing is performed, involving data cleaning, refinement, and organization. The cleaning process includes handling null values, imputing missing data, and removing outliers to enhance data quality. Additionally, numerical attributes are standardized using Min-Max scaling to ensure uniformity across features. The processing methods enable dataset development that produces usable information for model evaluation and training in effective classification models. The sequential list includes these preprocessing procedures:

Handle Null Values: This method handles null fields by replacing all general missing values with zeros and imputes F2 and F3 attributes with target outcome-based mean values. Data completion improves the accuracy of models because of this method.

Handle missing value: The mean of the related characteristic that corresponds to the desired outcome is used to replace missing values. It maintains data integrity as well as avoids creating bias in the predictive model.

Remove outlier: Statistical methods of Outliers are found and eliminated using tools like IQR and Z-score. This way helps to keep the data consistency and not let extreme values to hinder the model performance.

B. Data Organization

Data organization makes the dataset structured and ready for easy analysis. What it does is arrange the data in a consistent way, in particular, rows are taken to represent each individual sample, and columns represent some relevant features. To keep the data integrity, proper encoding of categorical variables, standardization of numerical values, and removal of duplicate and unnecessary features are some of the things that come handy.

C. Feature Scaling with Min-Max Scaling

In order to convert nonlinear data into linear form, data preprocessing is essential. Feature scaling is a critical stage in this process since it ensures that all features fall within a similar range and that no one feature dominates the learning process. Min-Max Scaling, one of the most popular normalizations, converts data into a defined range of [0 to 1], or [-1 to 1] Equation (1).

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

where X is the feature's initial value and X_{min} is its lowest value. X_{max} is The highest value of the characteristic, and X' Scaled feature value is within the range [0,1].

D. Data Splitting

The pre-processed data were used to make two sets of data: one for teaching and one for testing. The model's performance is checked using the testing set, which has 20% of the data, after it has been trained with the training set, which has 80% of the data.

E. Classification using CNN model for PIMA Dataset

The CNN consists of a linked layer, a layer that drops off, pooling layers, convolutional layers, and an activation function. Using the given input, convolutional layers use filters to create feature maps[16]. Convolutional layers help reduce the amount of the input and expedite training without causing the model to become overfit. Figure 4 displays most pooling layers, including the max-pooling layer. To get the greatest value of the domain that the pooling kernels cover, the max-pooling approach has been used in conjunction with the parametric Equation 2 given below.

$$p^{l(i,j)} = max_{(j-1)W+1 \le t \le jW} \left\{ a^{l(i,j)} \right\} \quad (2)$$

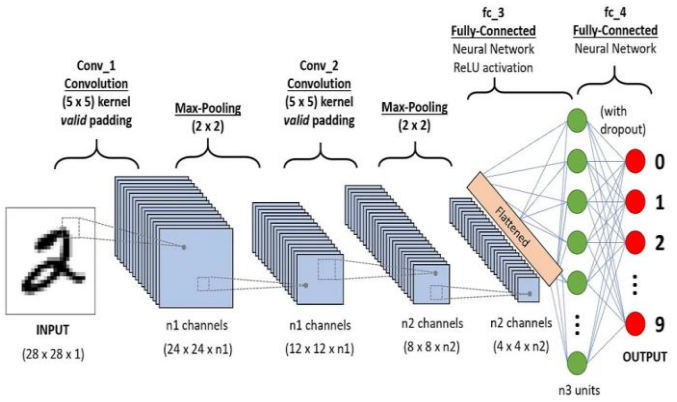


Figure 4: Architecture of CNN Model

where the values $p^{l(i,j)}$ and $a^{l(i,j)}$ represent the neuron's activation value and the pooling layer's breadth. An inseparable problem can be linearly transformed into a detachable one thanks to the activation function's enhancement of the model's flexibility[17]. The ReLU function is a valuable activation function because of its intrinsic capacity to generate values between 0 and 1 in Equation (3).

$$a^{l(i,j)} = f\left(y^{l\,(i,j)}\right) = max\{0,\} \qquad (3)$$

where a $l(i,j)$ indicates the layer output's activation value y $l(i,j)$, the ultimate forecast is produced by integrating the completely connected layers with the previously received data, as seen in the function that follows in Equation (4):

$$Z^{l+j} = \sum_{i=1}^{n} w_{ij}^{l}\, a^{l(i)} + b_j^l \,(4)$$

where w_{ij}^{l} $a^{l(i)}$and b_j^l represent the weight of the ith neurone, the bias values, and the ith neurone of the lth layer given the length of the input data n.

F. Performance Metrics

The efficacy of diabetes diagnosis was evaluated using a set of assessment measures, often called performance metrics shown in Figure 5. A confusion matrix compares the model's predicted and observed performance in a tabular format. Four main metrics—F1-score, recall, accuracy, and precision—were used to judge the end models. To begin, confusion matrices are used to look into how the model sorts things based on TP, FP, TN, and FN.

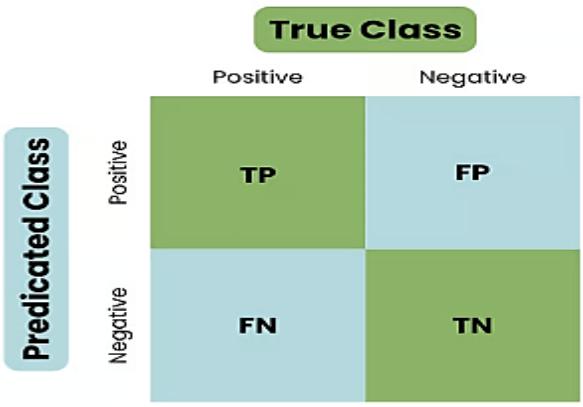


Figure 5: Confusion Matrix.

Accuracy: The accuracy of the prediction is shown by the percentage of all samples that were correctly forecast. This measure is found by dividing the total number of predictions by the number of correct predictions and incorrect predictions. This word is written as Equation (5):

$$Accuracy = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \qquad (5)$$

Precision: The ratio of true positives to fake positives shows how precise something is. It is written with an equation (6):

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (6)$$

Recall: One measure of sensitivity is the True Positive Rate, which is another name for recall; it measures the proportion of correctly recognized true positives. In Equation (7), it is shown:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (7)$$

F1-score: To evaluate a model's efficacy, taking accuracy and recall into consideration, one can utilize the F1-score, a single statistic. It is basically just the sum of the accuracy and sensitivity times one, divided by two. The F1 score can be defined as follows Equation (8):

$$\mathrm{F1} = \frac{2*(\mathrm{precision}*\mathrm{recall})}{\mathrm{precision+recall}} \qquad (8)$$

The comparison for the model performance for Diabetes Diagnosis in Healthcare Systems, these matrices are also being used.

4. Experimental Results & Discussions

The experimental results are AI-based methods that were used to detect early diabetes in the PIMA dataset. The processing was done first in MATLAB 2021b and then tested in Python with Keras and TensorFlow. The research was conducted using an NVIDIA RTX 2060 GPU, 16GB DDR3 RAM, and an Intel i7 CPU. An evaluation of the model's performance was conducted using F1-score, recall, accuracy, and precision; CNN was employed for classification.

Table 2: Model Performance of CNN algorithm based on Diabetes diagnosis using PIMA Dataset.

Accuracy	97.49
precision	96.71
Recall	95.59
F1-score	95.42

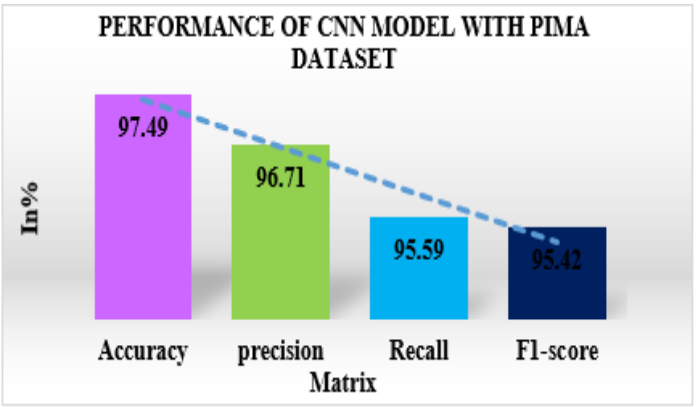


Figure 6: CNN Model Performance on PIMA Dataset.

The bar graph displaying the CNN model's performance is displayed in Figure 6, as well as in Table II. The results show that CNN is performing exceptionally well, with 97.49% accuracy, 96.71% precision, and 95.59% recall. It measures the precision and recall as 95.42%, which is a very high F1 score, meaning that the model is good at classifying instances with few false positives and negatives. The results obtained in these experiments indicate the robustness and reliability of the CNN in PIMA Diabetes Diagnosis data.

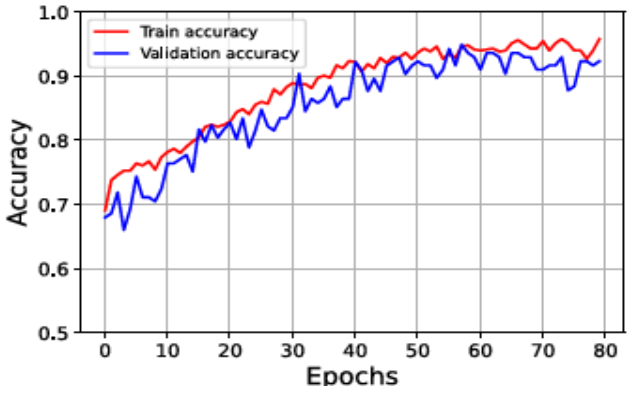


Figure 7: Accuracy Graph of Training and Validation for CNN.

Figure 7 shows the accuracy of the CNN model during training and validation on the PIMA Diabetes Diagnosis dataset. The validation accuracy (blue line) shows the fluctuations, which are in accordance with the model learning process, while the training accuracy (red line) continuously increases as usual. Both accuracies stabilize above 90% after 50 epochs, indicating strong model performance. It is clear from the small accuracy gap between validation and training that the model does a good job at generalizing, which reduces the possibility of overfitting and guarantees accurate diabetes diagnosis predictions.

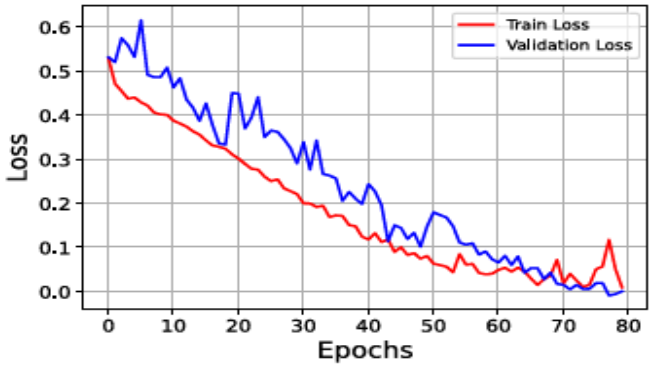


Figure 7: Loss Graph of Training and Validation for CNN.

The CNN model's loss during training and testing on the PIMA Diabetes Diagnosis dataset is shown in Figure 8. The blue line, which shows validation loss, changes shape more randomly as the model learns, while the red line, which shows training loss, stays the same. Small changes in confirmation loss could be a sign of overfitting, even if both losses go down. Cutting down on validation loss and bridging the gap between training and validation loss not only makes the model work better, but it also makes it better at adapting to new data.

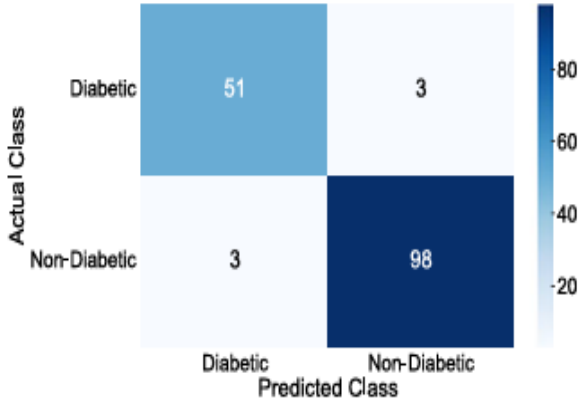


Figure 8: Confusion Matrix of CNN Model.

The CNN model's confusion matrix, as calculated using the PIMA Diabetes Diagnosis dataset, is shown in Figure 9. According to the matrix, 51 instances with diabetes and 98 cases without were accurately categorized by the model. Nevertheless, it incorrectly identified three cases of diabetes as non-diabetic and three cases of non-diabetes as diabetic. The high number of correct predictions and the low misclassification rate indicate strong model performance, demonstrating its ability to differentiate people with diabetes from those without the disease.

Table 3: Comparative analysis of proposed and base model based on diabetes diagnosis using PIMA Dataset.

	83.41	73.83	77.25	97.49
	89.76	64.45	77.40	96.71
	85.63	57.45	77.23	95.59
	87.65	60.55	76.34	95.42

The model performance comparison is summarized in Table III. Among the evaluated algorithms, CNN achieved the highest accuracy of 97.49%, outperforming DNN 83.41%, Logistic Regression 77.25%, and Bayesian Network 73.83%. Additionally, CNN demonstrated outstanding precision 96.71%, recall 95.59%, and F1-score 95.42%, highlighting its ability to minimize false positives while maximizing true positive detections. DNN, on the other hand, did very well too, with an F1-score of 87.65, an accuracy of 89.76%, and a memory of 85.63%. The Bayesian Network did the worst. Its accuracy was 64.45%, its recall was 57.45%, and its F1-score was 60.55%. In contrast, LR showed the greatest recall and F1-score, with 77.23% and 76.34%, respectively, indicating challenges in identifying all positive diabetes cases. All things considered, CNN was the best model for diagnosing diabetes, outperforming all other methods in every assessment criterion.

5. Conclusion & Future Work

The outcome is chronic metabolic disorder known as diabetes, in which Insulin resistance or deficiency can result in elevated blood sugar levels. It is one of the rapidly growing health problems across the globe, and its prevalence increases because of changes in lifestyle, obesity, and genetic predisposition. Early diagnosis coupled with accurate diagnosis is imperative to the management and prevention of some serious complications like cardiovascular diseases, kidney failure and neuropathy. The contribution of this study is on how AI-driven models, namely

CNN, improve diagnosis in diabetes. The CNN classification model then obtained 97.49% accuracy, 96.71% precision, 95.59% recall, and 95.42% F1-score, among other models. These findings show that, given a low number of true positives and negatives, DL is useful in distinguishing between instances with and without diabetes. In particular, the results demonstrate how ML has transformed healthcare by improving early detection, risk assessment, and customized treatment planning. For future research in this direction on utilizing AI-based ML for diabetes diagnosis, it would be beneficial to incorporate multimodal data (such as medical images and other genetic information) to make the diagnostic more accurate. Model interpretability and trust will be further improved by incorporating advanced methods, including deep reinforcement learning and explainable AI (XAI). The availability of real-time data from wearable devices might enable prompt and adaptive treatments for persons who are in danger. Federated learning also introduces the avenue for implementing privacy while collaborating even between healthcare systems. Future work should also take care of designing scalable algorithms that generalize well across many populations so that they can be widely applicable.

References

1. F. Ma *et al.*, "Incorporating medical code descriptions for diagnosis prediction in healthcare," *BMC Med. Inform. Decis. Mak.*, 2019, doi: 10.1186/s12911-019-0961-2.
2. L. Chen *et al.*, "OMDP: An ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems," *Int. J. Distrib. Sens. Networks*, vol. 15, no. 5, p. 1550147719847112, 2019, doi: 10.1177/1550147719847112.
3. D. Of and D. Mellitus, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 37, no. SUPPL.1, pp. 81–90, 2014, doi: 10.2337/dc14-S081.
4. J. F. Plows, J. L. Stanley, P. N. Baker, C. M. Reynolds, and M. H. Vickers, "The pathophysiology of gestational diabetes mellitus," 2018. doi: 10.3390/ijms19113342.
5. A. T. Hattersley and K. A. Patel, "Precision diabetes: learning from monogenic diabetes," 2017. doi: 10.1007/s00125-017-4226-2.
6. S. Kamalakkannan, R. Thiagarajan, S. Mathivilasini, and R. Thayammal, "Big data analysis for diabetes recognition using classification algorithms," *Int. J. Recent Technol. Eng.*, 2019, doi: 10.35940/ijrte.A1333.078219.
7. A. Choudhury and D. Gupta, "A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques," in *Advances in Intelligent Systems and Computing*, 2019. doi: 10.1007/978-981-13-1280-9_6.
8. D. S. W. Ting *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA - J. Am. Med. Assoc.*, 2017, doi: 10.1001/jama.2017.18152.
9. A. Agarwal and A. Saxena, "Analysis of machine learning algorithms and obtaining highest accuracy for prediction of diabetes in women," in *Proceedings of the 2019 6th International Conference on Computing for Sustainable Global Development, INDIACom 2019*, 2019.
10. M. A. Rahman, S. M. Shoaib, M. Al Amin, R. N. Toma, M. A. Moni, and M. A. Awal, "A Bayesian Optimization Framework for the Prediction of Diabetes Mellitus," in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, 2019, pp. 357–362. doi: 10.1109/ICAEE48663.2019.8975480.
11. 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.
12. R. Kathiroli, R. RajaKumari, and P. Gokulprasanth, "Diagnosis of Diabetes Using Cascade Correlation and Artificial Neural Network," in *2018 Tenth International Conference on Advanced Computing (ICoAC)*, 2018, pp. 299–306. doi: 10.1109/ICoAC44903.2018.8939103.
13. S. G., V. R., and S. K.P., "Diabetes detection using deep learning algorithms," *ICT Express*, vol. 4, no. 4, pp. 243–246, 2018, doi: https://doi.org/10.1016/j.icte.2018.10.005.
14. P. S. Kumar and S. Pranavi, "Performance analysis of machine learning algorithms on diabetes dataset using big data analytics," in *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, 2017, pp. 508–513. doi: 10.1109/ICTUS.2017.8286062.
15. A. Rusanov, P. V Prado, and C. Weng, "Unsupervised Time-Series Clustering Over Lab Data for Automatic Identification of Uncontrolled Diabetes," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 2016, pp. 72–80. doi: 10.1109/ICHI.2016.14.
16. G. Swapna, K. P. Soman, and R. Vinayakumar, "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals," in *Procedia Computer Science*, 2018. doi: 10.1016/j.procs.2018.05.041.
17. X. Zhao *et al.*, "Fine-Grained Diabetic Wound Depth and Granulation Tissue Amount Assessment Using Bilinear Convolutional Neural Network," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2959027.
18. K. Kannadasan, D. R. Edla, and V. Kuppili, "Type 2 diabetes data classification using stacked autoencoders in deep neural networks," *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 4, pp. 530–535, 2019, doi: 10.1016/j.cegh.2018.12.004.
19. S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Current techniques for diabetes prediction: Review and case study," *Appl. Sci.*, vol. 9, no. 21, 2019, doi: 10.3390/app9214604.
20. G. Battineni, G. G. Sagaro, C. Nalini, F. Amenta, and S. K. Tayebati, "Comparative machine-learning approach: A follow-up study on type 2 diabetes predictions by cross-validation methods," *Machines*, vol. 7, no. 4, pp. 1–11, 2019, doi: 10.3390/machines7040074.
21. 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).
22. 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.

23. 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.
24. 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.
25. Routhu, K., & Jha, K. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. *Available at SSRN 5106490*.
26. 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*.