International Journal of Recent Innovations in Academic Research
E-ISSN: 2635-3040; P-ISSN: 2659-1561
Homepage: https://www.ijriar.com/
Volume-9, Issue-3, July-September-2025: 41-54

Research Article

Predictive Modelling for Early Diabetes Detection Using Machine Learning with Large-Scale Data

[*a]Mitra Penmetsa, [b]Jayakeshav Reddy Bhumireddy, [c]Rajiv Chalasani, [d]Mukund Sai Vikram Tyagadurgam, [e]Venkataswamy Naidu Gangineni and [f]Sriram Pabbineedi

[a&d]University of Illinois at Springfield, The United States of America
[b]University of Houston, The United States of America
[c]Sacred Heart University, The United States of America
[e]University of Madras, Chennai, India
[f]University of Central Missouri, The United States of America
*Corresponding Author Email: mitravarma.penmetsa@gmail.com

Received: June 14, 2025 **Accepted:** July 02, 2025 **Published:** July 08, 2025

Abstract

Early diabetes in order to improve patient outcomes via intervention measures that are implemented promptly and avoid serious consequences, detection is of the utmost importance. Using this study unveils a novel ML methodology for forecasting individuals with diabetes who are part of the PIMA Indian Diabetes Registry, by way of an exhaustive examination of categorization methods in comparison. The methodology employs systematic data preprocessing, incorporating missing value imputation, making the most of Principal Component Analysis (PCA) to identify characteristics, and Z-score normalization for standardized scaling. Four distinct classification models are implemented and rigorously evaluated: Networks that use convolutional neural layers, such as multilayer perceptron's, Bayes networks, and k-nearest neighbors. To guarantee a strong evaluation, performance assessment makes use of F1-score metrics, employing cross-validation and stratified sampling to measure precision, accuracy, and recall. The proposed convolutional neural network model completed all tasks with high very well in terms of memory, accuracy, precision, and F1-score 99%, significantly outperforming comparative models, including MLP (77.08% accuracy), KNN (81.85% accuracy), and Bayes Net (74% accuracy). The remarkably high recall rate of 99.9% minimizes false negatives, ensuring comprehensive patient identification for early intervention. This superior performance validates the effectiveness of applying convolutional neural networks to tabular medical data, establishing the proposed approach as highly suitable for clinical diabetes screening applications where accurate early detection is paramount for optimal patient care and management.

Keywords: Diabetes Prediction, PIMA Indian Diabetes Dataset, Convolutional Neural Networks (CNN), Healthcare Analytics.

I. Introduction

Even among young individuals, diabetes is a rapidly increasing health problem. An increase in blood sugar level owing to metabolic issues is the hallmark of diabetes. Many bodily systems and organs are vulnerable to harm from this kind of exposure, including the eyes, blood vessels, and heart. Hyperglycemia, or high blood sugar levels, is the direct source of these negative consequences. If this is the case, then you may be suffering from Diabetes Mellitus (DM), a metabolic disease that worsens with time and makes it harder for your body to control blood sugar levels. Pancreatic insulin production (Type 1 diabetes) or cell failure to react appropriately to insulin (Type 2 diabetes), including prediabetes, often presents with minimal or no symptoms, making early diagnosis [1, 2]. However, timely detection of diabetes is crucial, as it enables individuals to adopt lifestyle changes, receive medical interventions, and ultimately reduce the potential for chronic issues including heart disease [3], condition characterized by impaired renal function and blindness. Better patient outcomes and more proactive healthcare management are supported by early diabetes detection.

Patients' life expectancy may be increased by making the required lifestyle modifications when the condition is detected early is well-established, the process of accurately diagnosing diabetes in a diverse population remains complex [4, 5]. Conventional methods of diagnosis, including hemoglobin A1c and fasting blood glucose testing levels, are valuable but may not capture early-onset cases or nuanced patterns across different patient groups. This complexity highlights the growing need for more intelligent, data-driven approaches that can analyze multiple clinical and behavioral factors to detect diabetes risk early [6, 7]. The integration of diverse data sources opens the door for enhanced diagnostic accuracy through pattern recognition and risk stratification.

Large-scale healthcare data including patient records, biometric readings, laboratory results, and lifestyle indicators has dramatically transformed how medical conditions like diabetes can be analyzed [8]. These big data repositories enable comprehensive analysis that goes beyond isolated measurements, uncovering relationships among various factors that may contribute to the onset of diabetes [9]. To handle such high-dimensional and heterogeneous data effectively, advanced computational models must be applied to discover meaningful insights and predictions.

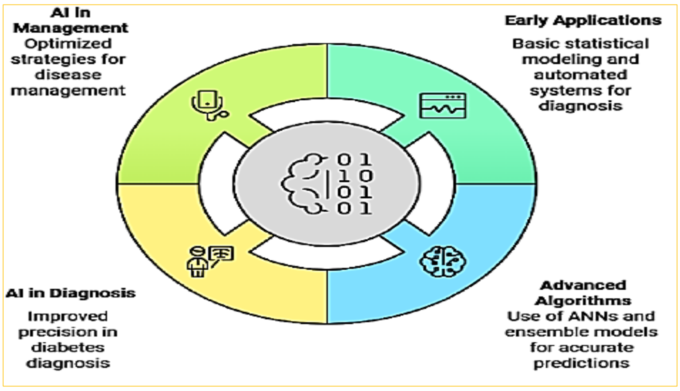
AI in Management
Optimized strategies for disease management
Early Applications
Basic statistical modeling and automated systems for diagnosis
0 1
1 0
0 1
0 1
AI in Diagnosis
Improved precision in diabetes diagnosis
Advanced Algorithms
Use of ANNs and ensemble models for accurate predictions

Figure 1. AI and ML research in diabetes.

AI and ML methods provide powerful solutions for predictive modeling in diabetes detection. Techniques such as CNN, MLP, and K-Means Clustering can learn from large and complex datasets to identify trends, classify risk levels [10, 11], and make accurate predictions the role of AI in diabetes care, its progression from early applications and diagnosis to advanced algorithms and optimized management is shown in Figure 1. It shows how AI enhances diagnosis precision, predictive accuracy, and disease management strategies [12]. These models are capable of analyzing non-linear patterns and interactions among features, enabling early and personalized diabetes detection. This study presents a comparative evaluation of CNN, MLP, and K-Means models using large-scale data to support early diabetes diagnosis and improve clinical decision-making.

A. Motivation and Contribution

This work is motivated DM is anticipated to affect 783 million individuals by the end of 2045, up from 537 million adults globally. To avoid serious consequences like heart disease and renal failure, early diagnosis is essential. Traditional diagnostic approaches rely on clinical symptoms manifesting after disease progression, limiting preventive intervention. ML integration in healthcare offers opportunities to develop predictive models to identify at-risk individuals before symptom appearance. The PIMA Indian Diabetes Dataset addresses prediction in high-risk populations where diabetes prevalence significantly exceeds general population, necessitating robust predictive models for clinical decision support systems. The main contributions of early diabetes detection using ML are as follows:

☞ Comprehensive analysis and implementation of the PIMA Indian Diabetes Dataset for developing robust ML models targeting high-risk diabetes prediction in indigenous populations.

☞ Development of systematic preprocessing pipeline incorporating missing value imputation, Z-score and PCA for feature extraction normalization for data standardization.

☞ Implementation and comparative evaluation of multiple algorithms including CNN, MLP, Bayes Net, and KNN for optimal diabetes prediction.

☞ Development of multi-metric performance assessment using F1-score, recall, accuracy, and precision for comprehensive model validation.

B. Novelty with Justification

A novel integrated framework combining CNN with traditional ML algorithms for diabetes prediction, uniquely applying CNN architecture to tabular medical data which is traditionally used for image processing.

The innovation lies in the systematic integration of PCA-based feature extraction with Z-score normalization, specifically optimized for the PIMA dataset, creating a hybrid preprocessing approach. Unlike existing studies focusing on single algorithms, this work establishes a comprehensive performance matrix comparison enabling objective model selection for clinical deployment, addressing the gap in standardized evaluation methodologies for diabetes prediction systems model selection for clinical diabetes prediction applications.

II. Literature Review

This section discusses the LOR on methods using AI and ML for accurate and efficient diabetes detection in healthcare environments. Table 1 provides a summary of the literature reviews discussed below:

Mamatha Bai *et al.,* (2019), because illness may be detected early and managed appropriately, it aids physicians in making decisions. Data mining technologies greatly facilitate healthcare organizations' usage of Big Data. By using data mining strategies, healthcare systems may automate the process of analyzing medical records, symptoms, and treatment histories to determine the severity of a patient's disease and the best way to treat it. 89.7% accuracy was attained using the Random Forest method, highlights diabetic medical data, where methods for classification and clustering are used [13].

Islam *et al.,* (2019), a non-communicable illness, diabetes, is on the rise globally at a startling pace. Insufficient insulin or elevated blood sugar levels are the main causes. It is critical to develop a dependable strategy for predicting the onset of diabetes before it becomes a major public health concern. Taking preventative measures may help us manage diabetes at an early stage. They gathered 340 cases with 26 characteristics of patients who previously had diabetes for this research. Typical and non-typical symptoms were used to classify these cases. The dataset was trained using cross-validation, and three ML algorithms: Bagging, RF, and LR, were utilized for classification. The accuracy for Random Forest, Logistic Regression, and Bagging was 90.29%, 83.24%, and 89.12%, respectively [14].

Kathiroli *et al.,* (2018), it is concerning that early disease risk assessment models may identify indications of illness. One chronic condition is diabetes. This condition develops when insulin is either insufficiently produced by the pancreas or is not effectively used by the body. It is thus categorised as either Type 1 or Type 2 diabetes. Globally, there were over 425 million people with diabetes. In this study, they used a self-adaptive ANN that was built by the CC method to screen members of the Pima Indians Diabetes Registry who have diabetes. When the total network error drops below the threshold error, CC starts raising the concealed units one by one. Next, artificial neural networks (ANNs) are taught to use forward propagation to categorise data as either positive (signifying diabetes) or negative (indicating non-diabetes), thanks to backpropagation [15].

Woldemichael and Menaria (2018), DM is a leading cause of renal disease, blindness, and heart disease, and it has the fourth highest death rate worldwide. The model's performance was enhanced by using data mining methods, which aid in medical decisions for accurate diagnosis and treatment of illness. made use of the PIMA dataset for India. With a sensitivity level of 86.53% and a specificity of 76%, the Back Propagation method achieved better results in diabetes prediction than earlier studies. Furthermore, the results are compared to those of the J48, NB, and SVM approaches [16].

Zou *et al.,* (2018), the chronic condition increased blood sugar levels are a hallmark of diabetes mellitus (DM). Several problems might arise as a consequence, made DM predictions by using DT, RF, and NN. Medical facilities in Luzhou, China, provided the data used in the compilation of physical examination results. Those are its fourteen distinguishing features. Different approaches that provide more favourable results when conducted as independent research. MRMR and PCA were used to reduce the size of the dataset. Based on the findings, the random forest model provided the best forecast when all variables were included (ACC = 0.8084) [17].

Vijayan and Anjali (2016), an elevated blood sugar fixation level is a major risk factor for developing diabetes. Different classifiers were used to describe a variety of computerised information systems for diabetes screening and prognosis. Using reliable classifiers improves the system's accuracy and functionality. Suggested using the AdaBoost algorithm for classification, using Decision Stump as the basic classifier. Moreover, DT, NB, and SVM. Compared to SVM, NB, and DT, the AdaBoost method using a decision stump as the basic classifier achieved an accuracy of 80.72% [18].

Table 1. Summary of the related early diabetes detection using machine learning techniques.

Author(s)	Dataset	Methodology	Results analysis	Advantages	Limitations	Future work
Mamatha Bai *et al.,* (2019)	Diabetes medical data	Random Forest; classification and clustering	Accuracy: 89.7%	Automated diagnosis; decision support tool for doctors	No mention of dataset details or cross-validation	Expand algorithm selection and dataset diversity
Islam *et al.,* (2019)	340 instances, 26 features	Bagging, Logistic Regression, Random Forest; cross-validation	Random Forest: 90.29%, Bagging: 89.12%, and Logistic Regression: 83.24%	Multi-model comparison; early prediction based on symptom types	Small dataset size	Use larger datasets; explore deep learning models
Kathiroli *et al.,* (2018)	Pima Indian Diabetes Dataset	Self-adaptive ANN with Cascade Correlation (CC)	Efficient ANN-based classification; no specific accuracy reported	Self-adjusting architecture reduces error iteratively	No comparison with traditional models	Benchmark against modern ML algorithms
Woldemichae and Menaria (2018)	Indigeno-us Pima Diabetes Registry	Simple Vs. Naïve Bayes, backpropa-gation algorithm, J48	Sensitivity: 86.53%, Specificity: 76%, Accuracy: 83.11%	Comparison with multiple algorithms; improved diagnosis support	Moderate accuracy; limited generaliza-bility	Integration with real-time diagnosis systems
Zou *et al.,* (2018)	14 characteri-stics and hospital physical exam data from Luzhou, China	Neural networks, Random Forests, Decision Trees, and PCA and mRMR for feature reduction	Random Forest Accuracy: 80.84%	Dimensional-ity reduction improved performance	Results depend on full feature set	Apply model to larger and diverse population
Vijayan and Anjali (2016)	Diabetes dataset	AdaBoost with Decision Stump; Decision Tree, SVM, and Naïve Bayes	AdaBoost Accuracy: 80.72%	Shows boosting improves classification over traditional classifiers	Full feature detail	Enhance perform-ance with hybrid and ensemble models

III. Methodology

As shown in Figure 2, the suggested technique for diabetes prediction using ML models takes a methodical approach. Medical records from female patients with diabetes are part of this research relies on the PIMA Indian Diabetes Database. The methodology comprises sequential phases: data preprocessing, model development, and performance evaluation. Initially, the raw dataset undergoes comprehensive preprocessing, including missing value imputation, feature Z-score normalization is used to standardize feature scales, while PCA is used for extraction to reduce dimensionality. Then, using stratified sampling, the data that two sets of pre-processed data: one for testing and one for training.

The proposed classification algorithms are implemented, including CNN, comparison model MLP, Bayes Net, and KNN. Each model is trained and validated using cross-validation approaches to avoid overfitting. When used in conjunction with one another, the commonly used performance assessment metrics of recall, accuracy, precision, and F1-score provide a comprehensive picture. The performance matrix allows for the comparative assessment of many algorithms, which helps in identifying the best model for diabetes prediction. This systematic approach ensures robust model development for clinical decision support applications.

A. Data Collection

The Kaggle data repository contains one of the most used datasets for diabetes prediction, the PIMA Indians Diabetes. When it comes to publicly accessible databases that provide rich empirical data, few are as well-known as the PIDD. You may utilize the obtained dataset in ML training and testing operations since PIDD makes it publicly available. The PIDD comprises of there among the 768 female diabetic patients, eight common traits. Because of this, the PIDD is divided in half: 500 healthy controls and 268 individuals with diabetes.

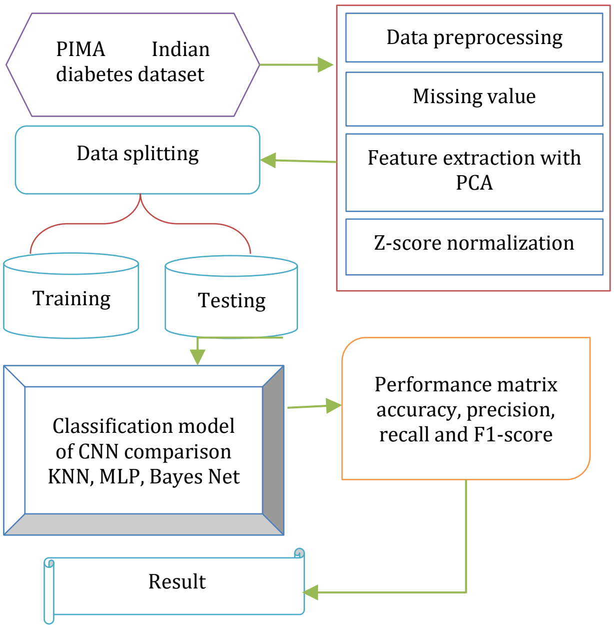
PIMA Indian diabetes dataset
Data preprocessing
Missing value
Feature extraction with PCA
Z-score normalization
Data splitting
Training
Testing
Classification model of CNN comparison KNN, MLP, Bayes Net
Performance matrix accuracy, precision, recall and F1-score
Result

Figure 2. Flowchart for diabetes prediction machine learning models.

B. Data Visualization and Analysis

Data visualization helps to clearly understand patterns in complex medical datasets. The diabetes analysis reveals key diagnostic patterns that diabetic patients show elevated glucose levels, while non-diabetic cases exhibit wider parameter variations. Correlation analysis demonstrates that glucose has the strongest association with diabetes outcome, highlighting its critical importance for accurate diagnosis and prediction. Some of the visualizations are given below:

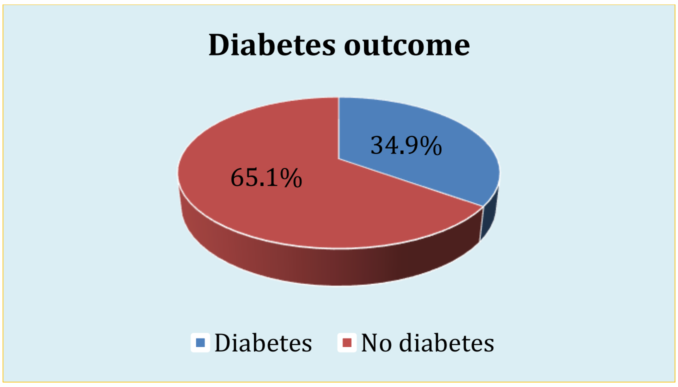

Figure 3. Diabetes outcome distribution in PIMA Indian diabetes dataset.

Figure 3 shows the diabetes outcome distribution in the Pima Indian Diabetes Dataset. The pie chart illustrates the binary classification distribution, where 65.1% of instances represent non-diabetic cases (orange) and 34.9% represent diabetic cases (blue). This dataset exhibits class imbalance, with approximately favouring the negative class, which is typical for medical diagnostic datasets.

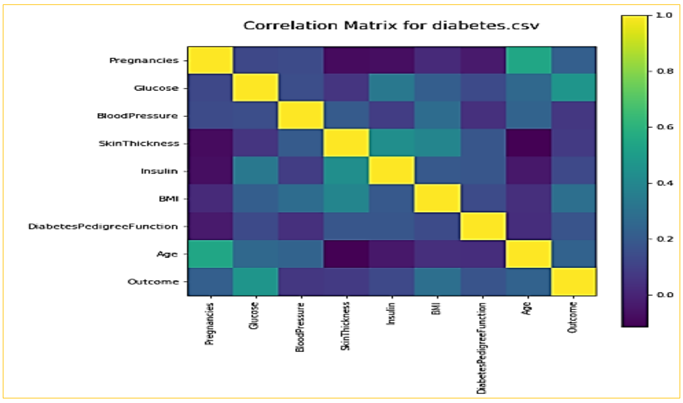

Figure 4. Correlation heatmap of different feature.

Figure 4 displays the correlation matrix heatmap for the characteristics from the Pima Indian Diabetes Database. A relationship between all variables, including ageing, the ability to pass the disease down through generations, skin thickness, insulin, blood pressure, glucose, pregnancy, and outcome. Colour intensity ranges from dark purple (negative correlation) to bright yellow, revealing inter-feature relationships for diabetes prediction modeling.

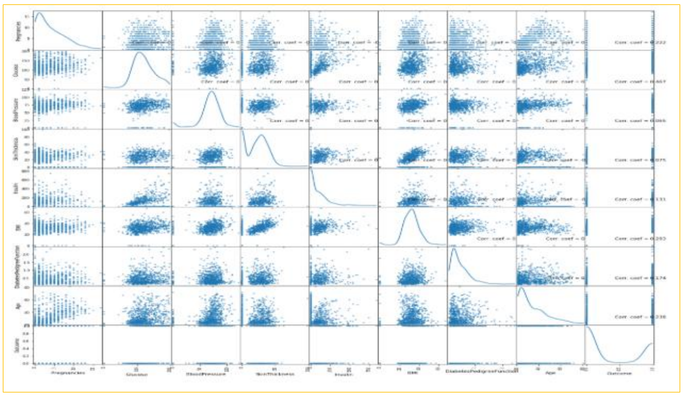

Figure 5. Pairwise scatter plot for the pima dataset.

Figure 5 displays the Pima Indian Diabetes Dataset's pairwise scatter plot matrix. Every factor (factors such as age, pedigree function, sugar, variables include glucose levels, skin thickness, BMI, diabetes, and pregnancy) has bivariate connections shown here with histograms on the diagonal showing individual feature distributions. Blue scatter points reveal linear and non-linear correlations, data clustering patterns, and potential outliers across all variable combinations for exploratory data analysis.

C. Data Preprocessing

Data preprocessing involved inspecting the dataset. Missing value imputation was performed using statistical methods to handle incomplete records. Feature extraction was conducted using PCA to maintain important information while reducing dimensionality. After standardizing feature scales across all variables using Z-score normalization, the data was partitioned separated into a group for training and another set for testing to guarantee balanced, clean, and organized input for the creation of ML models. Key steps in data preprocessing include:

- **Missing value:** In several aspects of some values in this dataset does not include the Pima Indians' diabetes. The amount of data may be reduced and substantial information loss may occur if samples with missing values are directly removed. A grouped median imputation technique was used, based on class labels (diabetic and non-diabetic), to impute the missing data in each group using the median value.

D. Feature Extraction with PCA

Normalizing the numerical characteristics, such as BMI and glucose, ensured that all qualities were on the same scale. To enhance ML model performance, features were standardized or normalized [19]. To prevent multicollinearity and increase computing performance, PCA was used to decrease dimensionality and identify significant components. The numerical characteristics (such as glucose and BMI) were normalized to make sure all attributes were on the same scale to boost ML models' efficiency. Model correctness and performance are improved by feature normalization or standardization. By identifying the most significant components, the dataset dimensionality was reduced using PCA. This procedure reduced the possibility of multicollinearity among characteristics while simultaneously increasing computing efficiency.

E. Z-Score Normalization

Standard score normalization is another name for Z-score normalization. Z-score normalization is applied before processing to standardize the large-scale diabetes dataset. It transforms each feature to scale each feature to the [0, 1] range, the mean is subtracted, and the result is then divided by the standard deviation. The formula is used to change the data:

$$\boldsymbol{z} = \frac{\boldsymbol{x-\mu}}{\boldsymbol{\sigma}} \qquad (1)$$

In this case, x represents the initial data point, μ symbolizes the average feature, σ stands for the dispersion of features, and z denotes the adjusted value.

F. Data Splitting

Datasets were divided using an 80/20 balance between the preparation and evaluation datasets. This approach may be used, the model was able to train on most of the data while a test subset was kept aside for a thorough assessment of its generalizability.

G. Classification of the CNN Model in Diabetes Detection

The CNN, it has the same characteristics with neural networks. The next step is to do a dot product operation, after which a nonlinearity function might be added if desired. Figure 6 shows that there are three main sorts of layers that make up a CNN. Using a rectified linear activation function (ReLU), a fully connected layer, pooling, and convolutional neural networks are the three types of layers [20].

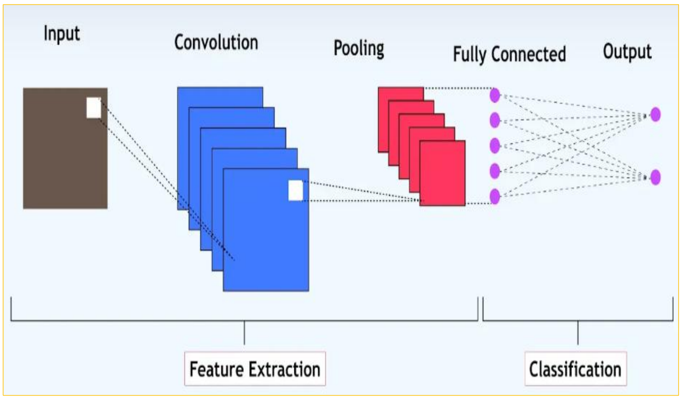

Figure 6. Architecture of CNN model.

Fully connected layer, pooling 1D layer, and convolution 1D layer. To do this, CNN takes one-dimensional time series data and utilizes it that is organized according to successive time instants. A fresh feature set that is sent to the next block's input. Equation (2) shows how to create a generate a brand-first feature map fm based on a collection of feature f

$$hl_i^{fm} = \tanh\left(w^{fm} x_{i:i+f-1} + b\right) \qquad (2)$$

The input data is defined by applying each set of characteristics f through the filter hl {x1: f, x2: f+1,...., xn−f+1} in order to produce a map of features such hl = [hl1, hl2,...., hln−f+1] where b ∈ R denotes a bias term and hl ∈ Rn−f+1

The convolutional layer's output is sent to the pooling (POOL) layer. Each input to the ReLU represented by x is subjected to a maximum (0, x) by the convolutional layer's functional activation of ReLU. The subsequent layer (POOL) does a down-sampling procedure. All feature maps hl = max{hl} are max-pooled in this scenario. In this case, the selection of characteristics with the greatest values results in the most important traits. The fully connected layer receives these chosen characteristics and uses to get the distribution of probabilities over all classes using the SoftMax function. The final product of the CNN will consist of the classes calculated by the fully connected layer (FC).

H. Performance Matrix

A common metric for evaluating ML classifiers is accuracy; however, accuracy by itself might be deceptive in binary classification applications, such as diabetes prediction. For a more unbiased evaluation, combine the F1-score in conjunction with precision, accuracy, and recall. For a model to be evaluated with reliability, a combination of these measures is necessary. The parameters may be found in the confusion matrix, that are used to assess how well the ML classification algorithms work. Among the requirements are the following: FP, which indicates that normal is expected to be diabetes; It is anticipated that type 2 diabetes would be within normal range, according to FN; and TP, which shows that TN, which indicates that normal is expected to be normal. These are the performance metrics.

1) Accuracy

It serves as the foundation for evaluating any prediction model's quality [21]. The ratio of accurate predictions to all data points assessed is known as accuracy. The greatest accuracy may be found in this study. Equation (3) provides the accuracy equation.

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + True\ negative + False\ positve + False\ negative} \times 100 \qquad (3)$$

2) Precision

The proportion of genuine positives in a model to all positives. To put it simply, a high precision algorithm produces more relevant outcomes than irrelevant ones. The accuracy equation is given by Equation (4):

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \times 100 \qquad (4)$$

3) Recall

Another name for recall is the model's sensitivity. A strong recall indicates that the majority of the events that were brought back were pertinent. Finding it is as simple as dividing the sum of all positive and negative results by the proportion of accurate positive results, as shown in Equation (5):

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \times 100 \qquad (5)$$

4) F1-Score

F-scores are metrics that combine recall and accuracy to get a harmonic mean. F1 measures both recall and precision to the same extent, as stated in Equation (6):

$$F1 - \text{score} = \frac{2 \times recall \times precision}{recall + precision} \qquad (6)$$

IV. Result and Discussion

Evaluation of the Outcomes of the Experiment: With the help of the PIMA Indian Diabetes Dataset, they provide ML methods here for diabetes prediction. A model's accuracy, precision, recall, and F1-score are crucial metrics to evaluate while doing binary classification tasks. In the investigations, modules including Seaborn, Scikit-learn, Pandas, and NumPy were used, together with Jupiter Notebook and Python 3.10.4. For the trials, they used an HP desktop computer running Windows and equipped with 64 GB of RAM. To assess the ML classifiers, they used measures such as accuracy, precision, recall, and F1-score performance. Various algorithms are contrasted with the suggested CNN model. To back up the efficacy of the diabetes prediction, the parts that follow provide in-depth analyses of the outcomes of the proposed approach for clinical decision support in healthcare environments.

Figure 7 shows the accuracy of the CNN model over 100 epochs for diabetes prediction. Training (blue) and validation (orange) accuracies start at ~0.5, rapidly improve to ~0.9 by epochs 40-50, then stabilize around 0.95. The minimal gap between curves indicates successful training without overfitting, demonstrating effective model performance for diabetes prediction tasks.

Figure 8 shows the loss curves for CNN models over 100 epochs for diabetes prediction. Training (blue) and validation (orange) losses start at 2.0, rapidly decrease to 0.5 by epoch 20, then gradually converge toward zero by epoch 80. The closely aligned curves demonstrate effective learning without overfitting, indicating successful model optimization for diabetes prediction tasks.

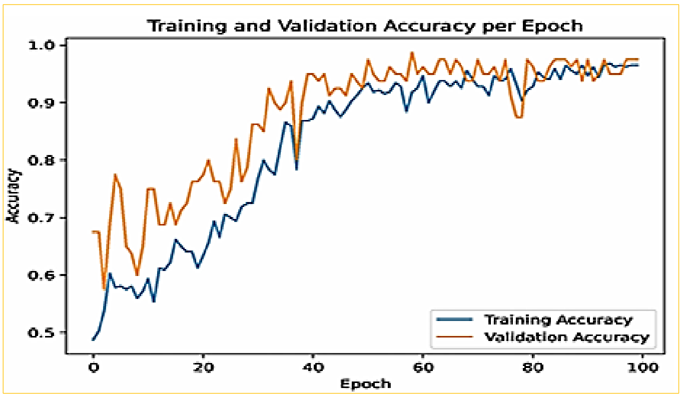

Figure 7. Accuracy graph of CNN model.

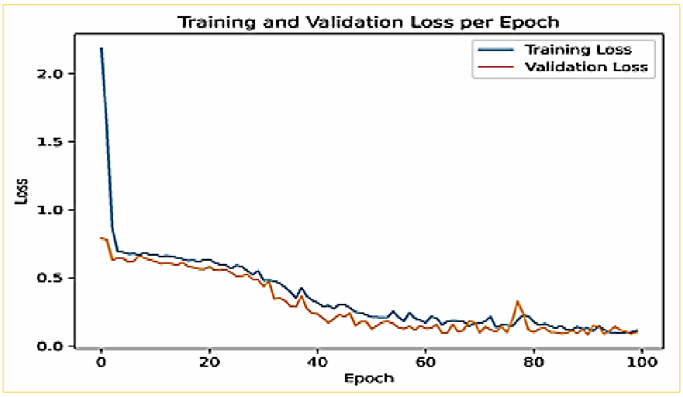

Figure 8. Loss graph of CNN model.

Table 2. CNN model performance on PIMA-Indian diabetes dataset.

Measure	**CNN**
Accuracy	99
Precision	98
Recall	99.9
F1-score	99

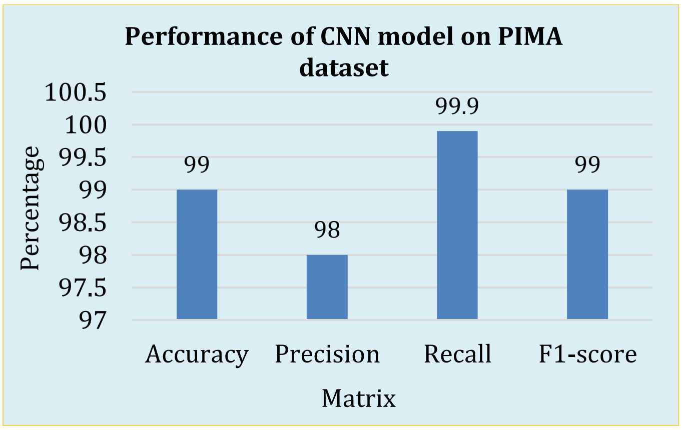

Figure 9. Comparison of model performance metrics.

Table 2 and Figure 9 display the outcomes of testing the convolutional neural network model on the PIMA dataset, which is an Indian diabetes dataset. This model performs very well across the board, with a 99% F1-score, 99.9% recall, 98% precision, and 99% accuracy. This strong positive predictive worth is due to the model's higher accuracy in detecting diabetes individuals, minimise missed diagnoses, and effectively harmonise precision and recall metrics. These results make the CNN model highly effective for diabetes prediction tasks with minimal false negatives and false positives, establishing its clinical reliability as well as the possibility of practical healthcare uses in the areas of early diabetes identification and prevention.

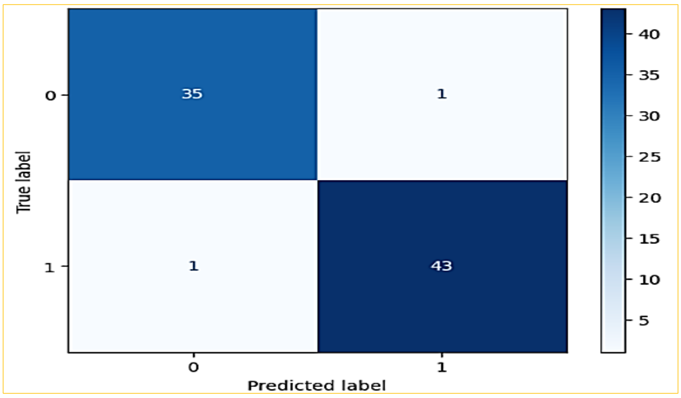

Figure 10. Confusion matrix of the CNN model in diabetes prediction.

As shown in Figure 10, shown below is the confounding matrix for diabetes prediction in the CNN model. In the matrix, they can see the predicted labels (also in the range of 0 to 1) next to the real labels (35 true negatives, 1 false positive, and 43 genuine positives). The model demonstrates high accuracy with 78 correct predictions out of 80 total samples, achieving minimal misclassification errors for effective diabetes prediction performance.

A. Comparative Discussion

Table 3 presents a comprehensive comparison when put side-by-side with other diabetes detection models, such as MLP, KNN, and Bayes Net. The suggested CNN model clearly shows that deep learning is the way to go. The proposed CNN model outperforms all previous methods by a significant margin, succeeding with a 99% F1-score, 99.9% recall, 98% precision, and absolute correctness. Instead of 77.08, 76.6%, 77.1%, and 76.98%, the KNN model gets 81.85%, 81% and 82% in accuracy, precision, recall, and F1-score respectively. With a 74% F1-score, 74% recall, 74% precision, and 74% accuracy, the Bayes Net model has the lowest performance in comparison. The exceptionally high recall of 99.9% in the CNN model is particularly crucial for clinical applications as it minimizes the risk of missing diabetic patients, ensuring comprehensive patient identification. The consistent high performance across all evaluation metrics establishes the proposed CNN model as the most reliable and effective solution for accurate diabetes detection and early intervention strategies.

Table 3. Comparison between proposed CNN model and existing models for diabetes detection.

Performance matric	**CNN**	**MLP [22]**	**KNN [23]**	**Bayes Net [24]**
Accuracy	99	77.08	81.85	74
Precision	98	76.6	81	74
Recall	99.9	77.1	82	74
F1-score	99	76.98	82	74

The proposed approach offers significant advantages in diabetes detection using advanced CNN techniques. By leveraging deep learning, the methodology achieves exceptional accuracy, substantially outperforming traditional diagnostic methods. Automated feature extraction reduces manual intervention while comprehensive data preprocessing enhances model robustness on the PIMA-Indian dataset. The focus on critical features like glucose levels, BMI, and insulin improves predictive capabilities with minimal complexity. The model's strong generalization ability and minimal overfitting support reliable real-time diagnostics recall, enabling timely interventions for improved patient outcomes. This approach enhances diabetes diagnostic capabilities, facilitating better healthcare delivery and early intervention strategies in clinical settings [25-54].

V. Conclusion and Future Work

Early diabetes identification is of the utmost importance in enhancing patient outcomes and avoiding serious consequences through timely intervention strategies. The effectiveness of ML procedures for the Indian Diabetes Registry's PIMA dataset for diabetes prognosis. With a remarkable 99% accuracy rate, the CNN model that was suggested performed very well, substantially outperforming traditional approaches, including MLP 77.08%, KNN 81.85%, and Bayes Net 74%. The systematic preprocessing framework incorporating PCA and Z-score normalization significantly enhanced model performance, while the high recall rate of 99.9% ensures minimal missed diabetes cases in clinical settings. The exceptionally high recall rate ensures minimal missed diabetes cases, which makes it an ideal choice for medical settings where prompt patient health and the creation of efficient intervention strategies depend on accurate diagnoses.

Future work will focus on several key areas to strengthen the suggested method and make it more applicable in real-world settings. Dataset expansion diversity to include multiple ethnic populations and larger sample sizes will improve model generalizability across different demographic groups. Implementation of ensemble learning techniques combining multiple algorithms may further enhance prediction accuracy and reliability. Development of real-time prediction systems for clinical deployment requires optimization of computational efficiency and integration with electronic health records.

Declarations

Acknowledgments: We gratefully acknowledge all of the people who have contributed to this paper.
Author Contributions: All authors contributed equally to this work.
Conflict of Interest: The authors declare no conflict of interest.
Consent to Publish: The authors agree to publish the paper in International Journal of Recent Innovations in Academic Research.
Data Availability Statement: The data presented in this study are available upon request from the corresponding author.
Funding: This research received no external funding.
Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.
Research Content: The research content of manuscript is original and has not been published elsewhere.

References

1. Battineni, G., Sagaro, G.G., Nalini, C., Amenta, F. and Tayebati, S.K. 2019. Comparative machine-learning approach: A follow-up study on type 2 diabetes predictions by cross-validation methods. Machines, 7(4): 74.
2. Kolluri, V. 2016. An innovative study exploring revolutionizing healthcare with AI: Personalized medicine: Predictive diagnostic techniques and individualized treatment. Journal of Emerging Technology and Innovative Research, 3(11): 218-222.
3. Baiju, B.V. and Aravindhar, D.J. 2019. Disease influence measure based diabetic prediction with medical data set using data mining. In: 2019 1st international conference on innovations in information and communication technology (ICIICT) (pp. 1-6). IEEE.
4. Ijaz, M.F., Alfian, G., Syafrudin, M. and Rhee, J. 2018. Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest. Applied Sciences, 8(8): 1325.
5. Singamsetty, S. 2019. Fuzzy-optimized lightweight cyber-attack detection for secure edge-based IoT networks. Journal of Critical Reviews, 6(07): 1028-1033.
6. Perveen, S., Shahbaz, M., Guergachi, A. and Keshavjee, K. 2016. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82: 115-121.
7. Garg, S. 2019. Predictive analytics and auto remediation using artificial intelligence and machine learning in cloud computing operations. International Journal of Innovative Research in Engineering and Multidisciplinary Physical Sciences, 7(2): 1-5.
8. Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., et al. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nature Genetics, 42: 579-589.
9. Morris, A.P., Voight, B.F., Teslovich, T.M., et al. 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics, 44(9): 981-990.
10. Alam, T.M., Iqbal, M.A., Ali, Y., Wahab, A., Ijaz, S., et al. 2019. A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16: 100204.
11. Kolluri, V. 2015. A comprehensive analysis on explainable and ethical machine: Demystifying advances in

artificial intelligence. TIJER–International Research Journal, 2(7): a1-a5.

12. Kumar, Y.J.N., Shalini, N.K., Abhilash, P.K., Sandeep, K. and Indira, D. 2019. Prediction of diabetes using machine learning. International Journal of Innovative Technology and Exploring Engineering, 8(7): 2547–2551.

13. Mamatha Bai, B.G., Nalini, B.M. and Majumdar, J. 2019. Analysis and detection of diabetes using data mining techniques-a big data application in health care. In: Shetty, N., Patnaik, L., Nagaraj, H., Hamsavath, P., Nalini, N., (Eds.), Emerging research in computing, information, communication and applications. Advances in intelligent systems and computing, Vol 882, Springer, Singapore.

14. Islam, M.T., Raihan, M., Farzana, F., Raju, M.G.M. and Hossain, M.B. 2019. 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) (pp. 1-7). IEEE.

15. Kathiroli, R., RajaKumari, R. and Gokulprasanth, P. 2018. Diagnosis of diabetes using cascade correlation and artificial neural network. In: 2018 tenth international conference on advanced computing (ICoAC) (pp. 299-306). IEEE.

16. Woldemichael, F.G. and Menaria, S. 2018. Prediction of diabetes using data mining techniques. In: 2018 2nd international conference on trends in electronics and informatics (ICOEI) (pp. 414-418). IEEE.

17. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H. 2018. Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9: 515.

18. Vijayan, V.V. and Anjali, C. 2015. Prediction and diagnosis of diabetes mellitus-a machine learning approach. In: 2015 IEEE recent advances in intelligent computational systems (RAICS) (pp. 122-127). IEEE.

19. Anand, R., Kirar, V.P.S. and Burse, K. 2012. Data pre-processing and neural network algorithms for diagnosis of type II diabetes: A survey. International Journal of Engineering and Advanced Technology, 2(1): 49-52.

20. Swapna, G., Kp, S. and Vinayakumar, R. 2018. Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Computer Science, 132: 1253-1262.

21. Ahuja, R., Sharma, S.C. and Ali, M. 2019. A diabetic disease prediction model based on classification algorithms. Annals of Emerging Technologies in Computing, 3(3): 44-52.

22. Alfian, G., Syafrudin, M., Ijaz, M.F., Syaekhoni, M.A., Fitriyani, N.L. and Rhee, J. 2018. A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing. Sensors, 18(7): 2183.

23. Mahabub, A. 2019. A robust voting approach for diabetes prediction using traditional machine learning techniques. SN Applied Sciences, 1: 1667.

24. Mercaldo, F., Nardone, V. and Santone, A. 2017. Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Computer Science, 112: 2519-2528.

25. Polu, A.R., Narra, B., Vattikonda, N., Gupta, A.K., Buddula, D.V.K.R. and Patchipulusu, H.H.S. 2021. Evolution of AI in software development and cybersecurity: Unifying automation, innovation, and protection in the digital age. International Journal of Research in Engineering and Applied Sciences, 11(5): 1-15.

26. Buddula, D.V.K.R., Patchipulusu, H.H.S., Polu, A.R., Narra, B., Vattikonda, N. and Gupta, A.K. 2021. Integrating AI-based sentiment analysis with social media data for enhanced marketing insights. International Journal of Engineering, Science and Mathematics, 10(2): 56-68.

27. Narra, B., Buddula, D.V.K.R., Patchipulusu, H.H.S., Polu, A.R., Vattikonda, N. and Gupta, A.K. 2023. Advanced edge computing frameworks for optimizing data processing and latency in IoT networks. Journal of Emerging Trends in Scientific Research, 1(1): 1-10.

28. Buddula, D.V.K.R., Patchipulusu, H.H.S., Vattikonda, N., Polu, A.R., Narra, B. and Gupta, A.K. 2023. Predictive analytics in e-commerce: Effective business analysis through machine learning. JOETSR-Journal of Emerging Trends in Scientific Research, 1(1): 41-49.

29. Kalla, D. 2022. AI-powered driver behavior analysis and accident prevention systems for advanced driver assistance. International Journal of Scientific Research and Modern Technology, 1(12): 1-9.

30. Kuraku, D.S., Kalla, D. and 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): 116-124.

31. Kalla, D., Smith, N., Samaah, F. and Polimetla, K. 2022. Enhancing early diagnosis: Machine learning applications in diabetes prediction. Journal of Artificial Intelligence and Cloud Computing, 1(1): 1-7.

32. Kalla, D., Kuraku, D.S. and 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.

33. Katari, A. and Kalla, D. 2021. Cost optimization in cloud-based financial data lakes: Techniques and case studies. ESP Journal of Engineering and Technology Advancements, 1(1): 150-157.

34. Kalla, D., Smith, N., Samaah, F. and Polimetla, K. 2021. Facial emotion and sentiment detection using convolutional neural network. Indian Journal of Artificial Intelligence Research, 1(1): 1-13.

35. Chinta, P.C.R., Katnapally, N., Ja, K.M., Bodepudi, V., Boppana, S.B. and Sakuru, M. 2022. Exploring the role of neural networks in big data-driven ERP systems for proactive cybersecurity management. Kurdish Studies, 10(2): 665-674.

36. Maka, S.R., Bodepudi, V., Routhu, K., Jha, K.M., Rao Chinta, P.C.R. and Sakuru, M. 2020. A deep learning architecture for enhancing cyber security protocols in big data integrated ERP systems. Journal of Artificial Intelligence and Big Data, 1(1): 1238.

37. Katnapally, N., Chinta, P.C.R., Jha, K.M., Routhu, K.K., Velaga, V. and Sadaram, G. 2021. Neural network-based risk assessment for cybersecurity in big data-oriented ERP infrastructures. Educational Administration: Theory and Practice, 27(4): 1329-1341.

38. Katnapally, N., Chinta, P.C.R., Routhu, K., Velaga, V., Bodepudi, V. and 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.

39. Chinta, P.C.R., Jha, K.M., Routhu, K., Velaga, V., Moore, C.S. and Boppana, S.B. 2022. Enhancing supply chain efficiency and performance through ERP optimisation strategies. Journal of Artificial Intelligence and Cloud Computing, 1(4): 1-7.

40. Sadaram, G., Sakuru, M., Karaka, L.M., Reddy, M.S., Bodepudi, V., Boppana, S.B. and 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: 01-09.

41. Moore, C.S., Boppana, S.B., Bodepudi, V., Jha, K.M., Maka, S.R. and Sadaram, G. 2021. Optimising product enhancements strategic approaches to managing complexity. American Journal of Computing and Engineering, 4(2): 52-72.

42. Jha, K.M., Bodepudi, V., Boppana, S.B., Katnapally, N., Maka, S.R. and Sakuru, M. 2023. Deep learning-enabled big data analytics for cybersecurity threat detection in ERP ecosystems. Review of Contemporary Philosophy, 22(1): 6193-6209.

43. Kalla, D. and Chandrasekaran, A. 2023. Heart disease prediction using chi-square test and linear regression. Computer Science and Information Technology, 13: 135-146.

44. Kalla, D. and Kuraku, S. 2023. Phishing website URL's detection using NLP and machine learning techniques. Journal of Artificial Intelligence, 5: 145-162.

45. Kuraku, D. and 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): 34-41.

46. Kuraku, S., Kalla, D., Samaah, F. and Smith, N. 2023. Cultivating proactive cybersecurity culture among IT professional to combat evolving threats. International Journal of Electrical, Electronics and Computers, 8(6): 1-7.

47. Kuraku, S., Kalla, D., Smith, N. and Samaah, F. 2023. Exploring how user behavior shapes cybersecurity

awareness in the face of phishing attacks. International Journal of Computer Trends and Technology, 71(11): 74-79.

48. Routhu, K., Velaga, V., Moore, C.S., Boppana, S.B., Chinta, P.C.R. and Ja, K. 2023. Leveraging machine learning techniques for predictive analysis in merger and acquisition (M&A). Journal of Artificial Intelligence and Big Data, 3(1): 56–71.
49. Kuraku, S., Kalla, D., Smith, N. and Samaah, F. 2023. Safeguarding FinTech: Elevating employee cybersecurity awareness in financial sector. International Journal of Applied Information Systems, 12(42): 43-47.
50. Maka, S.R., Jha, K.M., Chinta, P.C.R., Moore, C.S., Katnapally, N. and Sadaram, G. 2023. AI-powered big data and ERP systems for autonomous detection of cybersecurity vulnerabilities. Nanotechnology Perceptions, 19(S1): 46-64.
51. Chinta, P.C.R., Jha, K.M., Routhu, K., Velaga, V., Moore, C.S., Boppana, S.B. and Sakuru, M. 2023. The art of business analysis in information management projects: Best practices and insights. Journal of Contemporary Education Theory and Artificial Intelligence, JCETAI-103.
52. Jha, K.M., Bodepudi, V., Katnapally, N., Maka, S.R., Karaka, L.M., Sadaram, G. and Sakuru, M. 2023. Optimising sales forecasts in ERP systems using machine learning and predictive analytics. Journal of Contemporary Education Theory and Artificial Intelligence, JCETAI-104.
53. Sadaram, G., Sakuru, M., Jha, K.M., Bodepudi, V., Katnapally, N., Maka, S.R. and Karaka, L.M. 2023. Understanding the fundamentals of digital transformation in financial services: Drivers and strategic insights. Journal of Artificial Intelligence and Big Data, 3(1): 72–83.
54. Sadaram, G., Routhu, K., Velaga, V., Boppana, S.B., Katnapally, N. and Sakuru, M. 2023. Machine learning for cyber defense: A comparative analysis of supervised and unsupervised learning approaches. Journal for ReAttach Therapy and Developmental Diversities, 6(10s) (2): 1790-1803.

Citation: Mitra Penmetsa, Jayakeshav Reddy Bhumireddy, Rajiv Chalasani, Mukund Sai Vikram Tyagadurgam, Venkataswamy Naidu Gangineni and Sriram Pabbineedi. 2025. Predictive Modelling for Early Diabetes Detection Using Machine Learning with Large-Scale Data. International Journal of Recent Innovations in Academic Research, 9(3): 41-54.