International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 05 July 2023 **Revised:** 25 August 2023 **Accepted:** 15 September 2023*

From Pixels to Diagnoses: Deep Learning in Diabetes Detection

Sai Prakash .S[1], Dr A C Subhajini[2]
[1]Research Scholar, Department of Computer Applications
NICHE, Noorul Islam Centre For Higher Education
Kumaracoil, India
e-mail: mailtosaitvm@gmail.com
[2]Associate Professor, Department of Computer Applications
NICHE, Noorul Islam Centre For Higher Education
Kumaracoil, India
e-mail: jinijeslin@gmail.com

Abstract— Diabetes mellitus is a chronic metabolic disease with a rising prevalence worldwide, affecting millions of individuals. Early detection and accurate classification of diabetes are crucial to reduce mortality rates and enhance the quality of life for affected individuals. Traditional diagnostic techniques for diabetes, such as blood testing and glucose tolerance tests, are costly, time-consuming, and often require substantial resources.This study proposes a deep learning model that utilizes the Pima Indian Diabetes dataset, consisting of health information from 768 individuals with attributes including blood pressure, glucose levels, BMI, etc. The aim is to overcome the limitations of traditional detection methods and develop a model capable of early detection and precise classification of diabetes.The classification of the data into diabetes and non-diabetic groups is done using a convolutional neural network (CNN) model. The experimental outcomes show the effectiveness of the proposed deep learning model, obtaining an accuracy of 94.2%, precision of 90.18%, recall of 98.9%, F1-score of 94.3%, Cohen's kappa of 88.5%, and ROC AUC of 94.4%. These findings indicate that a deep learning approach can be utilized to develop a model capable of accurately identifying diabetes in its early stages.Early identification of diabetes through the suggested deep learning framework holds promise for reducing the risk of complications associated with the disease. By leveraging the power of deep learning techniques, healthcare professionals can enhance their ability to detect and manage diabetes more efficiently, leading to improved patient outcomes and an overall reduction in the burden of this chronic condition.

Keywords- Diabetes mellitus (DM), CNN, PIDD, accuracy, Deep learning

I. INTRODUCTION

Public healthcare is a vital priority for safeguarding and avoiding illness outbreaks in the community. However, there has been a substantial surge in the emergence of chronic and hereditary conditions endangering public health in current years. One of the most hazardous disorders is Diabetes mellitus since it promotes to the development of other deadly conditions such as kidney disease, nerve damage, and heart disease. It is a rapidly spreading chronic disease in humans [12].

The human pancreas [14] secretes the hormone insulin, which enables our bodies to use the glucose from diet. Diabetes causes a decrease in insulin secretion, which reduces the effectiveness of insulin usage. Inadequate insulin synthesis in the pancreas prevents cells from absorbing glucose, leading in blood glucose accumulation [3].Diabetes can affect a person in four distinct ways, as types 1, 2, 3, and 4.An autoimmune disease known as type-1 diabetes causes damage to the essential cells needed to manufacture insulin for the body to absorb glucose and produce energy.Adolescents may experience it. In Type-2 diabetes [7], the body either challenges to consume insulin or is unable to manufacture insulin. It most commonly affects older persons. Pre-diabetes [22] is a Type-3 diabetes disorder in which the levels of sugar in the bloodare elevated but not to the degree of Type-2 diabetes. A body glucose level of 100 to 125 mg/dl is regarded as pre-diabetes [4]. A kind of diabetes that mostly affects pregnant women is gestational diabetes [6].

Diabetes stems from insulin production or response issues and early detection is vital for risk reduction [10]. Deep learning (DL), emulating the human mind, effectively addresses the selectivity-invariance challenge, yielding superior results, lower classification errors, and increased resilience to distortion compared to other methods in numerous studies [2].Several ML approaches [19, 20], bio-inspired computing techniques [18], and DL techniques [1, 13] have recently been applied in several medical prognoses. AI and DL approaches have revolutionised and impacted every industry [11]. In general, the medical field is one of the important domains where such technology is heavily used for the purpose of identifying and treatingseveral crucial disorders [21, 17]. The main goal of this studywas to recommend the creation of an

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

improved predictive model for earlier diabetes identification and prediction. Results from the experiments show that the proposed strategy surpasses in regards to accuracy as well as other evaluation criteria.

II. LITERATURE SURVEY

Das et al. [9] introduced a robust approach for accurately identifying risk variants associated with T2D. The method begins by creating Entropy-based digital representations of DNA sequences linked to T2D. These representations are then transformed into 224 x 224-pixel spectrum images. VGG19 and ResNet models are employed to extract distinct features from these spectrum images. Subsequently, k-NN and SVM algorithms are used to categorize the feature set, and the system's performance is evaluated using k-fold cross-validation. The results showcase the effectiveness of the proposed Entropy-based method, SVM, and ResNet, with a combined model achieving a maximum accuracy of 98.19%. Alex et al. [5] proposed research on an effectual prediction technique for DM categorization utilising Deep 1D-CNN values on an unbalanced dataset with missing values. First, missing values were removed using outlier detection. To mitigate the effect of the imbalanced class on prediction accuracy, the oversampling technique (SMOTE) was implemented. In the end, predictions were made with DCNN classifiers and evaluated with a standard set of metrics.

Utilizing the PIMA dataset, Naz et al. [24] provided a way for diabetes prediction using a diversified ML algorithm. Functional classifiers such as NB, ANN, DL, and DT all have an accuracy of 90-98%. The greatest findings for predicting diabetes onset on the PIMA datasets emerge from DL, with an accuracy of 98.07%. Kannadasan et al. [16] presented employing stacked auto encoders a DNNmodel for classification of diabetes data. After stacking auto-encoders are used to extract characteristics from the dataset, a softmax layer is used to categorise it. Using the training data as a reference, the networks were fine-tuned using a process called guided back propagation. Based on the outcomes, it is clear that their model is superior to the others, achieving an accuracy rate of 85.24 %.

An algorithm was used by Nai-Arun et al. [23] to categorise the risk of DM. Theyutilised four ML classification algorithms to achieve the goal: ANN, DT, Naive Bayes and LR.The proposed model was made more robust by the application of boosting and bagging methods. The results of the experiments show that the Random Forest technique surpassesevery other method analysed. Chaki et al. [8] investigated ML models for diabetes detection. Researchers' categorised 107 studies by kind of model or classifier used, size of datasets, number of features used in feature selection, and overall performance. Results were shown to be improved by including text, form, and texture qualities, as discovered by the authors.They also found that in terms of classification, SVMs and DNNs performed better than RFs.A diabetes prediction system using soft computing that incorporates three widely used supervised ML algorithms was developed by Kumari et al. [15]. Databases for from PIMA and cancers of breast were utilised for analysis. The technique outperforms with 78.9% accuracy when compared to cutting-edge individual and ensemble methods built on LR, RF, and naive Bayes. A. U. Haqet et al. [3] suggested detecting diabetic retinopathy using PNN, SVM, and a Bayesian classifier. A total of 250 images were used in the initial analysis of the system. After some preliminary processing, the necessary features could be extracted. They classified the images into three groups. SVM attained a maximum accuracy of 97.7%.

III. METHODOLOGY

According to the literature review, DL techniques outperform other approaches such as ANN and ML for diabetes mellitus identification. DL models generated more favorable and efficient outcomes. It is a method that, by integrating neural networks and ML ideas with additional layers, predicts diabetes with a higher accuracy rate. The proposed model is shown in figure 1

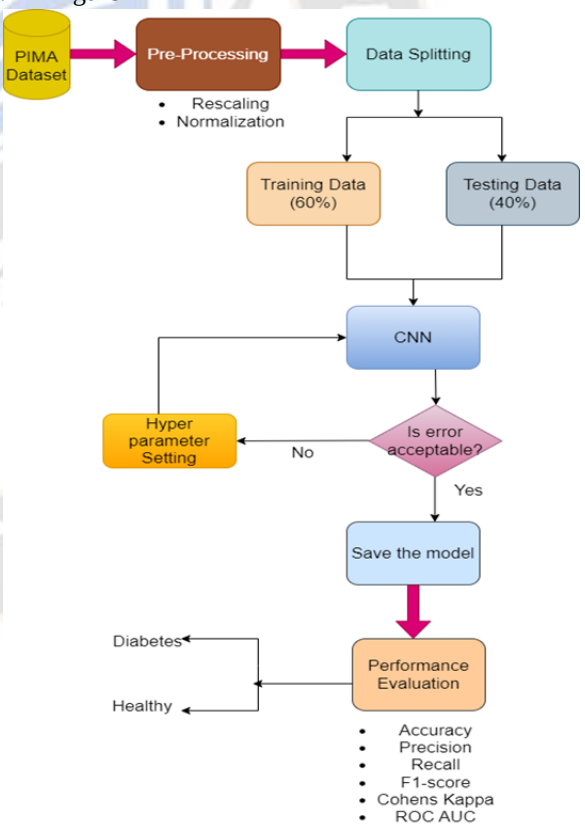


Fig.1: Proposed Methodology

A. Data Set

The study utilised the use of the PIMA Indian dataset (PID) from NIDDK, which is available from the Kaggle Repository. PIMA Indian assesses a user's likelihood of acquiring diabetes

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 05 July 2023 **Revised:** 25 August 2023 **Accepted:** 15 September 2023*

in the following four-year period in women. Due to the greatest risk of diabetes, the NIDDK has been doing long-term cohort research on PID since 1965.

Table.1: PIDD Attributes

Attribute	Description	Range
Pregnancies	Number of times pregnant	[0-17]
Glucose	Plasma glucose concentration based on a 2 hours oral glucose tolerance test	[0-199]
Blood Pressure	Diastolic blood pressure (mmHg)	[0-122]
Skin Thickness	Triceps skin fold thickness(mm)	[0-99]
Insulin	2-Hour serum insulin(mu U/ml)	[0-846]
BMI	Body Mass Index(BMI)	[0-67.1]
Diabetes Pedigree Function	Diabetes Pedigree Function	[0.078-2.42]
Age	Age(in years)	[21-81]
Outcome	Tested positive or negative	(0 or 1)

The data comprised measures and diagnostic indicators that made it possible to make an early diabetes or chronic illness diagnosis for the patient. The collection includes data on diverse women whose ages range from 21 to 81. PID consisted of 768 samples, of which 268 had diabetes and 500 did not. In each row, six attributes indicate physical examination details, whereas the remaining attributes pertain to chemical analysis information. Information on the patient's diabetes is contained in the last attribute in the row. The last column in each row contains either a 1 or a 0, with 1 denoting diabetes and 0 denoting non-diabetes. The dataset's properties are defined in Table 1.The *pandas head ()* method is used to analyse the data set. Figure 2 depicts the first five rows of a Data Frame that have been printed.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig.2: Visualization of dataset

B. Data Preprocessing

It is a crucial process that improves the data in order to facilitate the extraction of valuable data. When the dataset is first collected from many sources, it is in a rudimentary format, therefore there may be many divergences that the model cannot handle. As a result, pre-processing is required to remove any divergences and generate a clean data set. This includes fixing missing values, calculating new features, splitting data in the train-test set, data encoding, and data normalisation.Data that are null and inaccessible are routinely collected in the medical field. Missing data refers to this information. The number of approaches to handle missing data values is 'n'. The median, mean, and mode are a few examples of statistical techniques. In the present investigation, missing data are replaced with the "median," allowing to balance variables including blood sugar, blood pressure, skin thickness, insulin, and BMI. A check was performed to see whether there is any link between the dataset's properties. The result 'true' is replaced with a '1' and the result 'false' is replaced with a '0'.

After data augmentation, the dataset has a total of 2052 entries. As seen in Figure 3, *info ()* describes the column names and associated types of data. One quantitative discrete binary, six quantitative discrete integers, and two numeric continuous numeric floats with a 64-digit arrangement make up the data types of the attributes. The amount of memory used is at least 144.4 KB. The dataset has no null values, according to an examination of the columns and data types. As a result, the dataset utilised here has already been cleansed.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2052 entries, 0 to 2051
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2052 non-null   int64
 1   Glucose                   2052 non-null   int64
 2   BloodPressure             2052 non-null   int64
 3   SkinThickness             2052 non-null   int64
 4   Insulin                   2052 non-null   int64
 5   BMI                       2052 non-null   float64
 6   DiabetesPedigreeFunction  2052 non-null   float64
 7   Age                       2052 non-null   int64
 8   Outcome                   2052 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 144.4 KB
```

Fig.3: Diabetes data set description

Statistical analysis is used to determine the mean, standard deviation, and outliers in a data set. This will also help us create correlation matrices to determine which features substantially influence diabetes prediction. The *describe ()* method is used to learn about each of our columns' statistical information as shown in figure 4.

```
df.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	2052.000000	2052.000000	2052.000000	2052.000000	2052.000000	2052.000000	2052.000000	2052.000000	2052.000000
mean	4.093567	126.017057	69.551657	20.950780	84.649123	32.817349	0.491890	34.175926	0.512671
std	3.492140	33.126241	19.923201	16.413418	120.709134	7.831214	0.342711	11.675014	0.499961
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	102.000000	64.000000	0.000000	0.000000	28.000000	0.251000	25.000000	0.000000
50%	3.000000	123.000000	72.000000	24.000000	20.000000	32.800000	0.391000	31.000000	1.000000
75%	7.000000	147.000000	80.000000	33.000000	136.250000	37.200000	0.658000	42.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Fig.4: Method for viewing statistical information for the columns in the data set using *describe ()*

However, other features, such asskin thickness, blood pressure, glucose,BMI, and insulin have 0 values, which could be due to data entry errors. With the exception of the Pregnancies variable, none of these variables can be 0. Unless a person is very underweight, their BMI shouldn't be close to zero.The median or mean of a specific column must be used in place of 0 values. The difference between the maximum value for features such as Insulin, Skin Thickness, and Age and the third quartile suggests that there may be outliers in the data.Diabetes has an outcome variable of 1 while healthy people have a variable of 0. Figure 5 depicts a bar plot of the count in the dataset. According to the statistics, which is demonstrated to be skewed since there are disproportionately more people with diabetes than without it.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

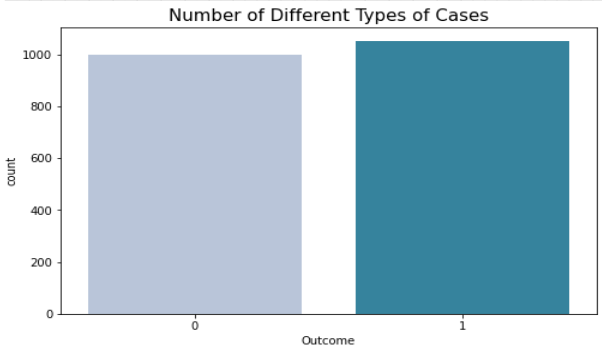


Fig.5: Bar plot visualization of number of non-diabetic and diabetic people in the dataset

Data visualisation is a critical component of data science. It aids in data comprehension as well as data explanation to others. Matplotlib, Seaborn, and other visualisation libraries are available in Python. To determine the distribution of the feature's data, we will utilise the pandas visualisation, that is built on top of matplotlib. Histograms can be used to readily visualise the distribution of each attribute. Figure 6 depicts the feature value distribution.

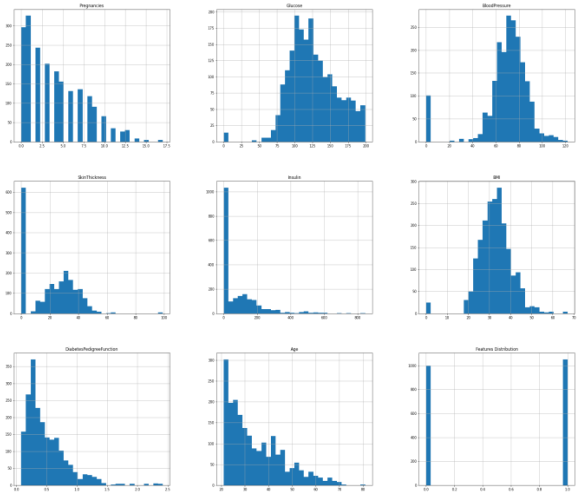
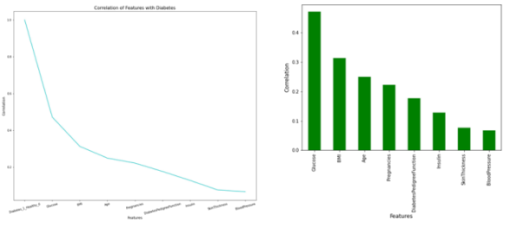


Fig.6: Feature value distribution

Only the distributions of BMI,glucose, and blood pressure are normal; the rest are skewed and contain outliers. Outliers in a dataset are uncommon numbers that might affect statistical studies and violate their assumptions. As a result, dealing with them is critical. We have to employ a variety of scaling and transformation techniques to address this situation since deleting outliers might lead to data loss.



(a) Correlation of features with diabetes (b) Correlation with case-type

Fig.7: Correlation Results

A correlation matrix is produced using the Pandas Data Frame's *corr()* function. A correlation of -1 or 1 indicates a complete negative or positive correlation. A score of 0 indicates that there is no link at all. As illustrated in figure 7a, every attribute is displayed on the x-axis of the correlation graph that corresponds with diabetes, and the coefficient of correlation is indicated on the y-axis.

The correlation of features with case type represents the link between different variables in a dataset and the classification of cases into different types or categories as shown in figure 9b. It helps determine how each feature is correlated or associated with the specific case types. The resulting graph is visualized as a bar or line chart, with the features represented on the x-axis and the corresponding correlation coefficients on the y-axis. The bars or lines is color-coded to signify the strength and direction of the correlation.

C. Feature Selection

The performance of the correlation is used to determine which features to use. A correlation heatmap is used to identify multi co linearity by listing all of the correlation coefficients. Based on the kind and degree of the correlation, it will compare and explain the linear relationship and connection between the two features. The values of the two traits are predicted to change in the same direction by a positive correlation and to change in the opposite direction by a negative correlation**.** A correlation matrix is a 2-D graphical representation of data that uses colours to display the matrix's value as shown in figure 8. The frequency and direction of a straight line connecting two quantitative variables are the two characteristics of correlation.

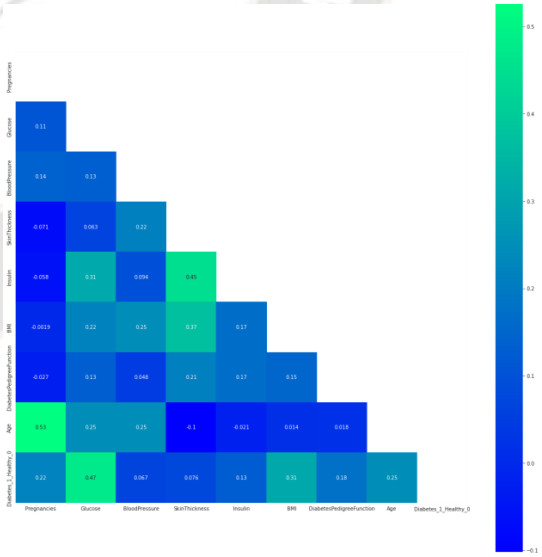


Fig.8: Heat map

D. Dataset Splitting

The dataset is prepared for testing and training purposes after preprocessing the data. The majority of the data gathered is susceptible to being altered as reckless. Aside from that, data

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 05 July 2023 **Revised:** 25 August 2023 **Accepted:** 15 September 2023*

quality is critical because it has a major impact on prediction outcomes and accuracy. Datasets must be properly balanced and split into training and testing data at a certain ratio as a consequence.The training dataset serves as the model's only source of learning, while the testing dataset serves as a performance evaluation of the model. The dataset is split in 6:4 between training and testing sets. The testing portion was chosen to guide the selection of hyper parameters. Theoretically, the testing set undergoes hyper parameter training before optimisation. A model is trained in the training sub-process, and its accuracy is then evaluated in the testing phase sub-process using the learned model.

E. Classification Model

Our suggested model used a DL neural network with three hidden layers for repeated dataset execution, one input layer for data entry, one output layer for prediction outcomes, and one input layer.

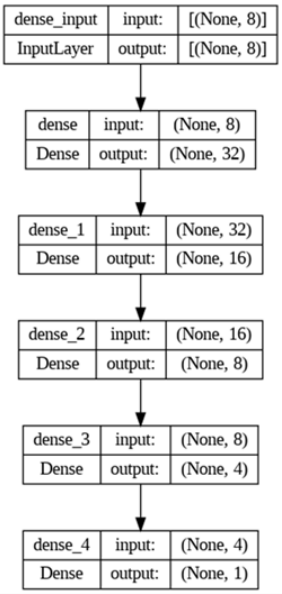


Fig.9: Model summary of a NN model

Each epoch comprises of all test cases being trained. A model with five thick layers was produced in this instance. The input and output layers of the network are the first and fifth layers, and they share the same input shape, neurons, and activation function as NNs with a single hidden layer. 16, 8, and 4 neurons, respectively, are present in each of the third and fourth hidden layers. The input layer has a value of 8 and the output layer has a value of 1. Figure 9 depicts the model summary of NN with three hidden layers.

We utilised an Intel Core i5 computer with 8GB of RAM for processing. We used Scikit-learn, an open-source Python machine learning framework that is integrated with Google Colab.Hyperparameters are the configuration settings that are specified before training a neural network model. These parameters cannot be learned during training, unlike the model's internal weights, and they directly influence the learning process and the model's performance. Table 2 shows the hyper parameters employed in this study.

Table.2: Hyper parameters

Hyper Parameters	Values
Epochs	2000
Learning Rate	0.01
Activation Function	RELU, Sigmoid
Loss	Binary Cross entropy
Optimizer	Adam
Batch size	512

F. Performance Model

1) Confusion Matrix

The performance of an algorithm can be seen using the error matrix, sometimes called the performance matrix. The confusion matrix is used to validate the CNN classifier's performance. The comparison of the algorithm's predictions to the ground-truth labels is shown in tabular form. The confusion matrix tabulates the expected and actual classes vertically and horizontally.

2) Evaluation Metrices:

The following widely utilised cutting-edge performance indicators are used to assess the performance of the suggested approaches.

Table:3 Evaluation Metrices

Accuracy = (TP + TN) / (TP + TN + FP + FN)	(1)
Precision = TP / (TP + FP)	(2)
Recall = TP / (TP + FN)	(3)
F1 Score = 2 * (Precision * Recall) / (Precision + Recall)	(4)
Specificity = TN / (TN + FP)	(5)
where TP (true positives) is the number of correctly identified Diabetes cases, TN (true negatives) is the number of correctly identified non-diabetes cases, FP (false positives) is the number of non-diabetes cases incorrectly identified as diabetes, and FN (false negatives) is the number of diabetes cases incorrectly identified as non-diabetes.	

3) Roc Curve

The true positive rate (TPR) and false positive rate (FPR) are compared at various threshold levels to produce the ROC curve. It is a graph that depicts the effectiveness of classifier at several thresholds.

IV. RESULT AND DISCUSSION

Predicting diabetes mellitus is essential, and so is improving the accuracy of diabetes predictions. Hence, the main aim is to present a suite of deep learning and machine-based algorithms for diabetes prognosis. Building a computer model that can reliably detect diabetes in its earliest stages is the focus of this research. Results from the experiments validate the model and show that both the diabetes-type and Pima Indians diabetes data sets are suitable for training the model.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

A. Accuracy and Loss Visualization

Visualising the model's performance is a straightforward way to make sense of the data generated by a DL model and to decide whether modifications to the model's attributes or hyper parameters are necessary.Figure 10 displays both the training and validation accuracy/loss graphs. Training accuracy appears to have been high for the model, as seen by the accuracy plot. It suggests the fact that the model has not yet over-learned the training data.The model accuracy had the best convergence, confirming that a satisfactory balancing between test and train models was accomplished.

Figure 11 illustrates the classification report, which demonstrates the reliability of the suggested models in recognising diabetes individuals at an early stage. The suggested model outperforms with 90.18% Precision, 98.99% Recall, and 94.27% Accuracy.

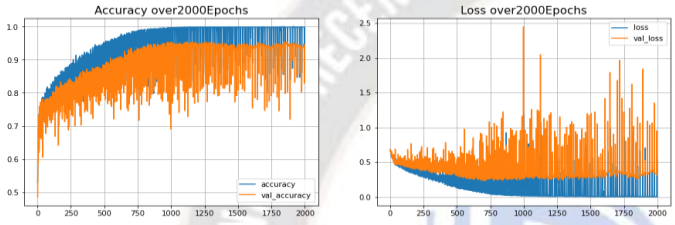


Fig.10: Data visualization of the trained model

The model's performance is assessed and validated by predicting the classifications of these new unbiased data that were not used to train the model using a test set of 821 records. There are 422 occurrences for the target class label of 0 and 391 examples for 1.

```
#classification report
print(classification_report(y_test, y_pred))

26/26 [==============================] - 0s 1ms/step
Accuracy: 0.942753
Precision: 0.901826
Recall: 0.989975
F1 score: 0.943847
Cohens kappa: 0.885720
ROC AUC: 0.944040
[[379  43]
 [  4 395]]
Specificity:  0.8981042654028436
              precision    recall  f1-score   support

           0       0.99      0.90      0.94       422
           1       0.90      0.99      0.94       399
```

Fig.11: Classification Report

Figure 12 depicts a table with columns representing expected values and rows showing actual values.

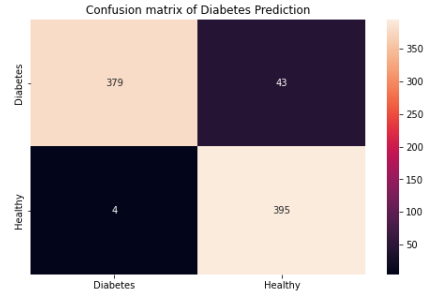


Fig.12: Confusion matrix of the proposed method

ROC curves were also used to evaluate the model. They are a valuable tool for creating classifier and visualising their performance. Since they create complete scenarios of the trade-off between sensitivity and false-positive rates over a range of thresholds, they are also commonly used in healthcare decision-making. The ROC curve's X-axis is used to draw false positive rate graphs, while the Y-axis is used to plot true positive rate graphs. Figure 13 depicts the ROC for the diabetes prediction algorithm.The effectiveness of a random classifier is shown as a diagonal line running from bottom left to top right. The ROC curve is above the diagonal line for the suggested CNN model, showing that it performs better than random selections.

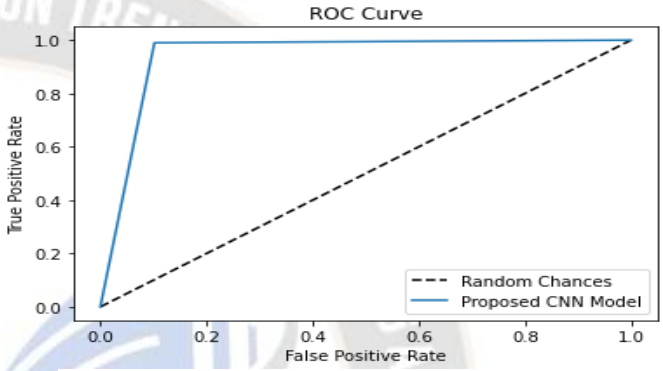


Fig.13: The diabetes prediction system's ROC curve

V. CONCLUSION

Using the Pima Indian Diabetes dataset, we suggested a deep learning-based model in this paper for the identification of diabetes mellitus. The rising prevalence of diabetes worldwide underscores the importance of early detection to mitigate its severe consequences and improve patient outcomes. Traditional detection methods are expensive and time-consuming, necessitating the exploration of more efficient and accurate alternatives.Our proposed CNN model demonstrated promising results in accurately classifying individuals into diabetic and non-diabetic categories. The framework attained an impressive accuracy of 94.2%, a precision of 90.18%, a recall of 98.9%, and an F1-score of 94.3%. Additionally, Cohen's kappa reached 88.5%, and the ROC AUC score was 94.4%. These performance metrics indicate the model's capability to effectively identify diabetes in its early stages, allowing for timely intervention and reducing the risk of complications associated with the disease.

REFERENCES

[1] A. Ashiquzzaman, A. K. Tushar, M. Islam, J.-M. Kim et al., ``Reduction of overfitting in diabetes prediction using deep learning neural network," arXiv preprint arXiv:1707.08386, 2017

[2] A. Thammano, A. Meengen, "A New Evolutionary Neural Network Classifier," Springer-Verlag Berlin, pp. 249-255, (9), 2005

[3] A. U. Haqet al., "Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data," Sensors (Switzerland), vol. 20, no. 9, 2020.

[4] AD Association. Classifcation and diagnosis of diabetes: standards of medical care in diabetes-2020. Diabetes Care. 2019. https://doi.org/10. 2337/dc20-S002.

[5] Alex, S. A., Nayahi, J. J. V., Shine, H., &Gopirekha, V. (2022). Deep convolutional neural network for diabetes mellitus prediction. *Neural Computing and Applications*, *34*(2), 1319-1327.

[6] Anna V, van der Ploeg HP, Cheung NW, Huxley RR, Bauman AE. Sociodemographic correlates of the increasing trend

[7] Bellamy L, Casas JP, Hingorani AD, Williams D. Type 2 diabetes mellitus after gestational diabetes: a systematic review

[8] Chaki J, Ganesh ST, Cidham SK, Theertan SA. Machine learning and artifcialintelligence based diabetes mellitus detection and self-management: a systematic review. J King Saud Univ Comput Inf Sci. 2020. https://doi.org/10.1016/j.jksuci.2020.06.013

[9] Das, B. (2022). A deep learning model for identification of diabetes type 2 based on nucleotide signals. *Neural Computing and Applications*, *34*(15), 12587-12599.

[10] Deng L, Yu D. Deep learning: methods and applications. Tech Rep. 2018:198–239https://www.microsoft.com/en-us/research/publication/deeplearning-methods-and-applications/, Accessed date: 10 September 201

[11] Huang CL, Chen MC, Wang CJ. Credit scoring with a data mining approach based on support vector machines. Expert Syst Appl. 2007; 33(4):847–56.

[12] International Diabetes Federation. Diabetes. Brussels: International Diabetes Federation; 2019.

[13] J. Vijayashree and J. Jayashree, "An Expert System for the Diagnosis of Diabetic Patients using Deep Neural Networks and Recursive Feature Elimination," International Journal of Civil Engineering and Technology, vol. 8, pp. 633-641, Dec. 2017.

[14] J.S. Kaddis, Human Pancreatic Islets and Diabetes Research, JAMA 301 (15) (2009) 1580, https://doi.org/10.1001/jama.2009.482.

[15] K.G.M.M. Alberti, P.Z. Zimmet, Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO Consultation, Diabet. Med. 15 (7) (1998) 539–553.

[16] Kannadasan, K., Edla, D. R., &Kuppili, V. (2019). Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clinical Epidemiology and Global Health*, *7*(4), 530-535.

[17] M. Jamjoom, "Data mining in healthcare to predict cesarean delivery operations using a real dataset," in Proceedings of the First International Conference on Computing and Emerging Sciences ICCE'2020, pp. 20–26, Erbil, Iraq, December 2020

[18] M. K. Hasan, M. M. Islam, and M. M. Hashem, "Mathematical model development to detect breast cancer using multigene genetic programming," in Proc. 5 th International Conference on Informatics Electronics and Vision (ICIEV), pp. 574-579, Dhaka, 2016

[19] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of Breast Cancer using Support Vector Machine and K-Nearest Neighbors," in Proc. IEEE Region 10 Humanitarian Technology Conference (R10- HTC), pp 226-229, Dhaka, 2017.

[20] M. R. Haque, M. M. Islam, H. Iqbal, M. S. Reza, and M. K. Hasan, "Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder," in Proc. International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, pp. 1-5, 2018.

[21] M. Rizwan, A. Shabbir, A. R. Javed et al., "Risk monitoring strategy for confidentiality of healthcare information," Computers & Electrical Engineering, vol. 100,

[22] Meigs JB, D'Agostino RB Sr, Wilson PW, Cupples LA, Nathan DM, Singer DE. Risk variable clustering in the insulin

[23] Nai-Arun, N., Moungmai, R., 2015. Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69, 132–142. doi: 10.1016/j.procs.2015.10.014.

[24] Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, *19*, 391-403.