Journal of Computing, Science &Technology
https://focjournal.unidel.edu.ng/
editorjcst@unidel.edu.ng
ISSN:.....Print
ISSN: ..... Online

Early Stage Diabetes Prediction Using Recurrent Neural Network Model

[1]Adedayo Sobowale, *[2]Adebimpe Esan, [3]Adetunji Omonijo, [4]Tomilayo Adebiyi,[5] Michael Adio, and [6]Oluwaseun Julius
[1,2,4,,6]Department of Computer Engineering, Federal University Oye-Ekiti, Nigeria
[3]Federal Teaching Hospital, Ido-Ekiti, Nigeria
[5]Ajayi Crowther University, Oyo, Nigeria
Corresponding author: adebimpe.esan@fuoye.edu.ng

Article Info
Date Received:20-02-2024
Date Accepted: 20-05-2024

Keywords:
Diabetes prediction, RNN, dataset, accuracy, early stage, prediction

ABSTRACT

Diabetes has dramatically increased the risk of various cardiovascular problems and previous approaches employed for diabetic prediction using machine learning algorithms are laced with curse of dimensionality and poor prediction accuracy issues. As a result, this research employs Recurrent Neural Network model, a subtype of deep learning to build a model for the prediction of Early-stage diabetes in patients. The dataset for training the model was obtained from Kaggle and a hospital in Nigeria. The dataset was preprocessed by removing duplicate and irrelevant entries, removing outliers, and handling missing data. Results from evaluation show that the diabetes prediction model performed well with an accuracy of 0.90, precision of 0.91, Recall of 0.95 and F1-Score of 0.93. The developed model outperformed conventional machine learning techniques and it has the potential to help healthcare professionals anticipate and prevent diabetes.

1.0 INTRODUCTION

Diabetes is a chronic condition that poses serious health risk to humans. It is defined by blood glucose levels that are greater than usual, which is caused by either faulty insulin secretion or its biological effects, or both. Diabetes can cause long-term damage and dysfunction in a variety of organs, including the eyes, kidneys, heart, blood vessels, and nerves [1]. Type 1 diabetes (T1D) and type 2 diabetes (T2D) are the two types of diabetes. Type1 diabetes occur mostly among young people under age thirty (30) years. Increased thirst, frequent urination and elevated blood glucose levels are common clinical symptoms of type 1 diabetes and insulin therapy as well as oral drugs are essential treatment for this type of diabetes. Obesity, hypertension, dyslipidemia, arteriosclerosis, and other disorders are typically connected with type 2 diabetes, and it is more common among middle-aged and elderly adults. Diabetes can also be diagnosed using fasting blood glucose, glucose tolerance, and random blood glucose readings in medicine [2]. Recently, numerous algorithms are used to predict diabetes, including the traditional machine learning method [3], such as support vector machine (SVM), decision tree (DT) and logistic regression. SVM are not suitable for large datasets likewise decision tree is unstable compared to other decision predictors. Quantum particle swarm optimization (QPSO) algorithm and weighted least squares support vector machine (WLS-SVM), linear Discriminant Analysis (LDA)[4] were also employed by previous research to predict early-stage diabetes. The limitations of machine learning approaches led to the introduction of deep learning models in the domain of medical prognosis. Deep learning models were introduced for: cancer prediction and detection, heart disease diagnosis [5] and blood pressure monitoring. Many studies have shown that deep learning techniques produce superior results, lower classification error rates, and are more noise resistant than conventional machine learning techniques. Hence, Recurrent Neural Network which is a deep learning approach was employed in this research for early-stage diabetes prediction to minimize the shortcomings of previous systems.

2.0 RELATED WORK

Machine Learning technique is the most emerging technology for addressing inevitable problems in various domains [6] while Deep learning (DL) approach, a subset of Machine learning is extensively used in medical prognosis such as diabetes prediction [7]. Previous researchhave employed various machine learning and deep learning techniques for the prediction of diabetes and some other diseases. [8]employed Dynamic thresholding-based FP-Growth for treating unusual illnesses and certain types of diseases with unknown variations and different symptoms. The limitation of the work is poor prediction accuracy which led to misinterpretation in some

instances.[9] developed a classification-based illness prediction that includes various mechanisms such as Linear and Polynomial Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) for diabetes prediction. K-Nearest Neighbor and Logistic Regression were employed by[10]for diabetes prediction. The limitation of the approach is its divergence due to excessive training. [11] used Decision Tree, SVM, and Naive Bayes classifiers to detect diabetes. Although, Naïve bayes achieved the highest accuracy of 76.30% on Pima India dataset, the limitation of the research is poor prediction accuracy. [12] used six different classifiersJ48, Multilayer Perceptron, Hoeffding Tree, JRip, BayesNet, and Random Forest for diabetes prediction on Pima Indian dataset. Hoeffding Tree algorithm achieved the highest performance accuracy of 0.75. [13] employed Multilayer Feed-Forward Neural Network for diabetes prediction using the Pima Indian Diabetes dataset and an accuracy of 82% was achieved. [14] also employed Deep Neural Network (DNN) for diabetes prediction on the Pima Indian dataset the accuracy obtained is 88.41%.[15]proposed Artificial Neural Networks for illness prediction and the limitation of the research is its susceptibility to overfitting and time consumption during training. Considering the shortcomings of previous machine learning based diabetes prediction systems, this research employed recurrent neural networks model for early-stage diabetes prediction to improve prediction accuracy and reduce processing time.

3.0 METHODOLOGY

This research focuses on diabetes risk prediction using Recurrent Neural Network. The dataset utilized was obtained from Kaggle with eight characteristic features. The dataset was divided into two parts: 80% for training and 20% for validation. Seven of the attributes was utilized to train the RNN model, while the eighth attribute was used as an output parameter for the model to predict the output value, with 0 representing no diabetes and 1 representing diabetic. The performance of this model is assessed using four separate metrics: accuracy, precision, recall, F1-measures, and area under curve (AOC). The block diagram of the developed system is shown in Figure 1. The dataset consists of data of seven hundred and sixty-eight patients with 9 attributes (i.e. the feature name and description) as shown in Table 1. The table contains 9 attributes extracted from the dataset for training the RNN model.

3.1 Data Collection and Preprocessing

The dataset was obtained from Kaggle with seven hundred and sixty-eight data of patients. The dataset comprises of 768 patients with 500 non-diabetic(65.1%) and 268 diabetic patients (34.9%). Eighty percent of the dataset was used for training while twenty percent was used for testing the RNN model. The dataset was pre-processed by data cleaning, integrating the data, transformation of data. Exploratory Data Analysis (EDA) was carried out on the dataset to check for imbalance and validity of the data. Figure 2 shows the exploratory analysis of the attributes and the outcome. Figure 3 shows the Boxplot chart of the dataset's characteristics' highest, lowest, and middle values. Outliers that could cause the model performance to be over fitted are shown by the dotted dots above and below the graphs.

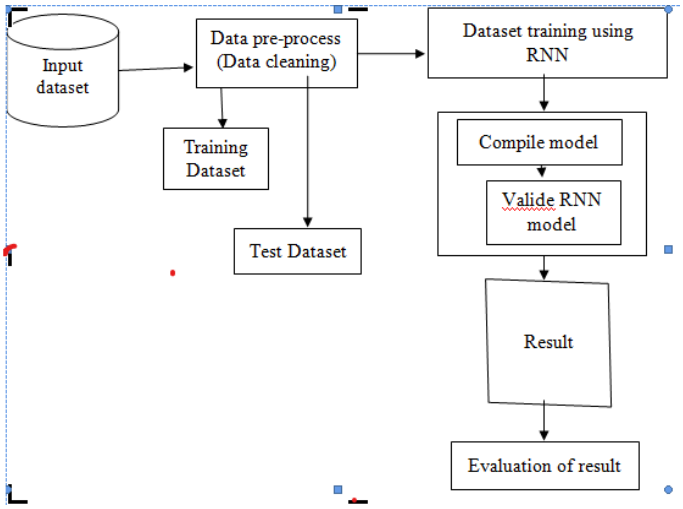

Figure 1: Block Diagram of the Recurrent Neural Network Model

Table 1: Description of 8 Attributes Obtained from the Dataset

S/N	Features Names	Description
1	Preg	Number of times pregnant
2	Plasma	Plasma glucose
3	Pres	Diastolic blood pressure
4	Skin	Triceps skin fold thickness
5	Insulin	Serum insulin
6	BMI	Body mass index
7	Pedi	Diabetes pedigree function
8	Age	Age of the patient
9	Outcome	Result of the prediction

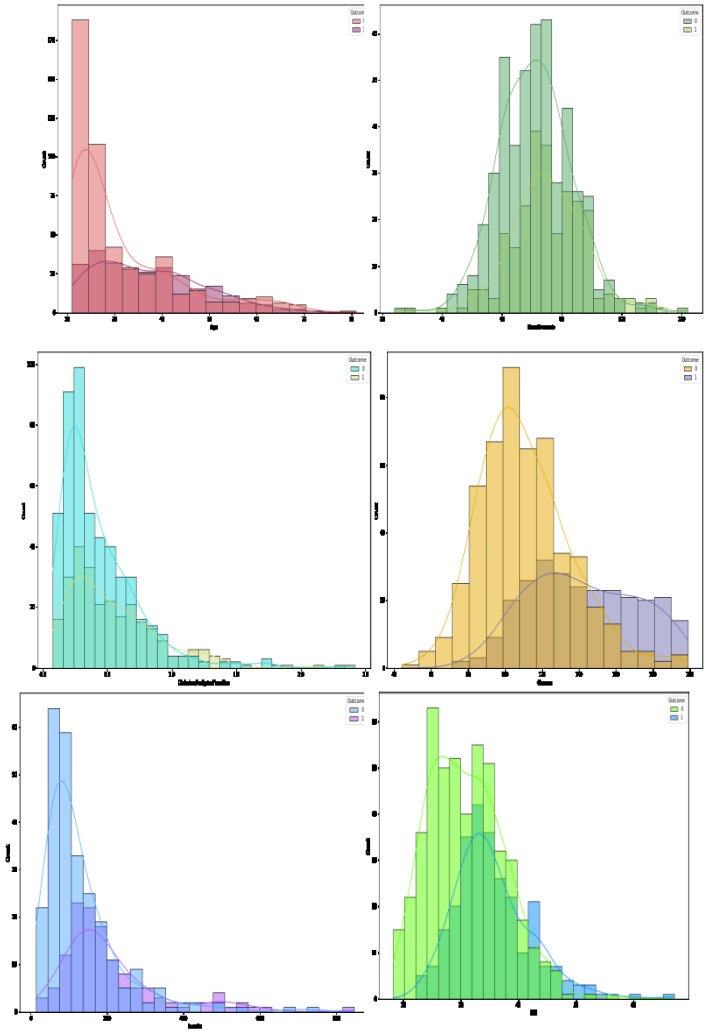

Figure 2: Exploratory Analysis of the Dataset

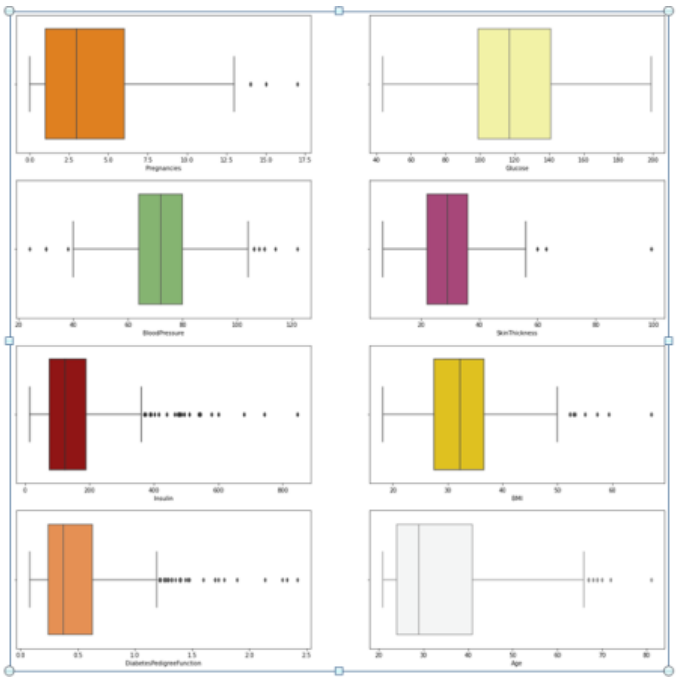

Figure 3: Boxplot of the Features in the Dataset

The percentage of missing values for each feature in the dataset is shown in the Figure 4. If the mean or median of the feature is not used effectively to replace those missing values, the model's accuracy may suffer. In this work, the missing value for a particular feature was replaced by the mean of the cases that were accessible, using the mean imputation approach, which replaces missing values in datasets. This approach keeps the sample size constant and is simple to apply.

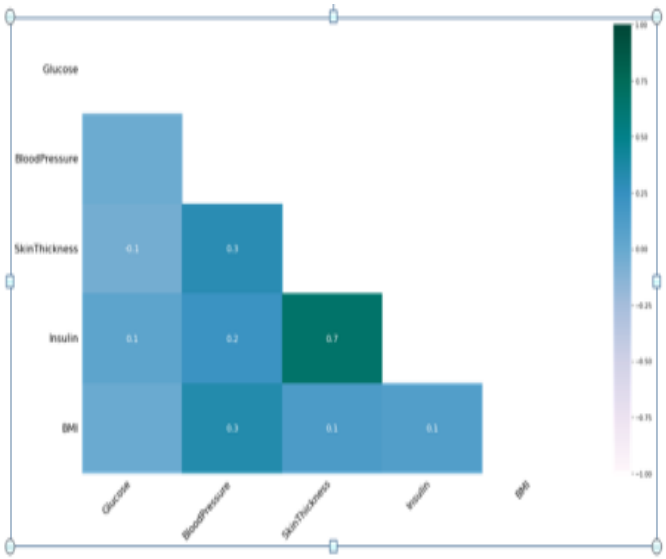

Figure 4: Missing Values Map of the dataset

3.3. Implementation

The developed model was deployed using Flask framework. Figure 5 shows the screenshot of the default page of the developed model that contained the attribute of the diabetes prediction. The attributes are times of been pregnant, the amount of sugar in the blood, blood pressure, skin thickness, amount of insulin in the blood, body mass index, diabetes pedigree function and age. Figure 6 shows the screenshot of diabetes prediction page.

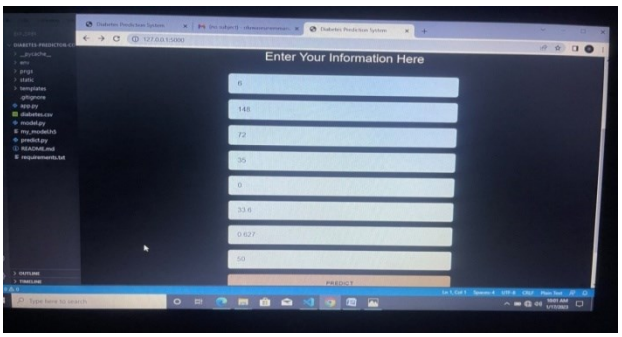

Figure 5: Default Page of the Deployed Model

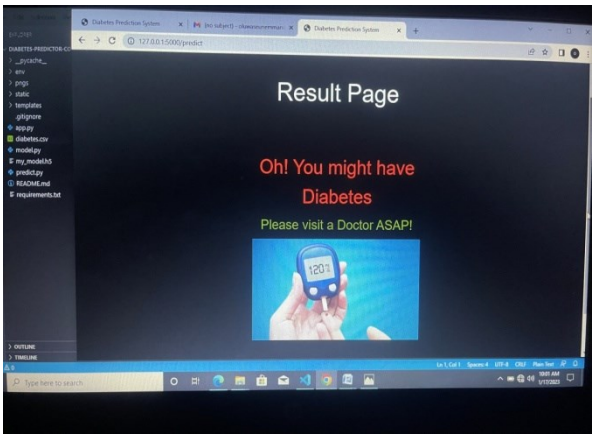

Figure 6: Diabetes Prediction page

4.0 RESULT and DISCUSSION

Confusion matrix was used to evaluate the performance of the developed model. 96 diabetic cases were classified and 10 non-diabetic cases were misclassified. The classifier also classified 42 Non-diabetic cases correctly while 5 Not-diabetic cases were misclassified as Diabetic as shown in Table 2.

Classifier		Prediction	
		Diabetic	Non Diabetic
RNN	True Result	96	42
	False Result	10	5

Table 2: Confusion Matrices for the developed RNN-based Model

Results obtained from model training at different epochs were: 0.65, 0.70 and 0.901 at 8th, 15th and 23rd epochs respectively as shown in Table 3. Results from evaluation recorded accuracy, recall, precision and F1-Scores of 0.901, 0.95, 0.905 and 0.758 respectively as shown in Table 4. The Graph of Model Accuracy during Training is shown in Figure 7.

Table 3: Accuracy at different epochs

S/N	Epoch	Accuracy
1	8	0.65
2	15	0.70
3	23	0.90

Table 4: Result from Confusion matrix

Metric	Score
Accuracy	0.90
Recall	0.95
F1-score	0.758
Precision	0.905

4.2 Discussion of findings

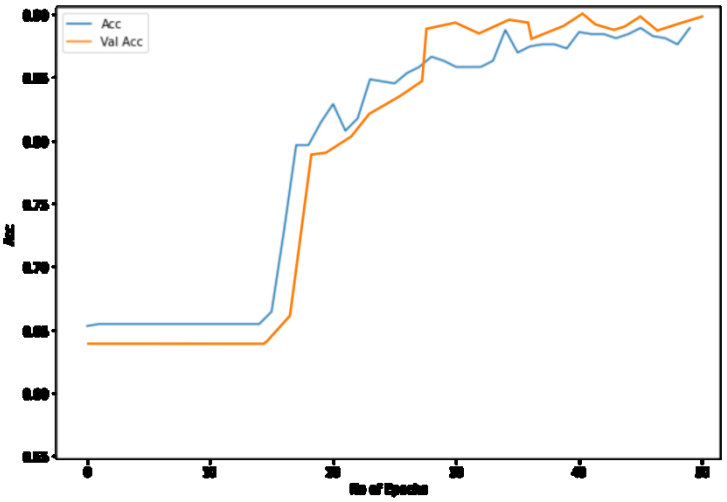

Figure 7: The Graph of Model Accuracy during Training

The result obtained from the developed system was compared to previous works using Multilayer Feed Forward Neural Network [14] and decision tree [12] and the results show an improved accuracy of up to 0.08 as shown in Table 5.

Table 5: Comparison of previous Machine Learning Algorithm for Prediction of Diabetes

Authors	Algorithm	Accuracy
Deepti (2019)	Decision Tree	76.30%
Mohebbi et al (2019)	Logistic Regression and CNN	77.5%
Olaniyi et al (2018)	Multilayer Feed Forward Neural Network	82%
Developed RNN model	Recurrent Neural Network (RNN)	90%

5.0 CONCLUSION

This research utilized Recurrent Neural Network (RNN) model for early-stage diabetes prediction.

The dataset used in training the model was obtained from Kaggle and a hospital in Nigeria. The system developed was evaluated using confusion matrix and results show that a good performance was achieved with an accuracy of 0.90, precision of 0.91, Recall of 0.95 and F1-Score of 0.93. When compared to previous works, this research show that RNN-based model performed better than some selected machine learning models for diabetes prediction. However, future research should employ indigenous dataset to make the work more suitable in Nigeria.

REFERENCES

[1] Krasteva et al. "Type 2 Diabetes, Pre-Diabetes, and the Metabolic Syndrome." *JAMA*, vol. 306, no. 2, 13 July 2011, 10.1001/jama.2011.970. Accessed 2 Dec. 2019.

[2] Feig, Denice. "Preventing Diabetes in Women with Gestational Diabetes." *Diabetes/Metabolism Research and Reviews*, vol. 28, no. 4, May 2012, pp. 305–306, 10.1002/dmrr.2280.

[3] Kavakiotis,I.,Tsave,O.,Salifoglou,A.,Maglaveras,N.,Vlahavas,I.,&Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural bio technology journal.

[4] Duygu, and Cooper, G.F.(2011). System to predict diabetes called LDA-MWSVM, used linear discriminant analysis. 3rd International Conference on Recent Trends in Computing (pp. 500-508). Elsevier B.V.

[5] Esan A, Akingbade J, Omonijo A, Sobowale A, Adebiyi T (2024). Development and Performance evaluation of Heart Disease Prediction Model using Convolutional Neural Network. ABUAD Journal of Engineering Research and Development, 7(1), 1-6.

[6] Shafi, J. (2022) "Machine learning through supervised semi-supervised approached" *SN Computer Science*, vol. 4, no. 1, 29 Nov. 2022, 10.1007/s42979-022-01485-3. Accessed 1 Dec. 2022.

[7] Konishi, T.; Matsukuma, S.; Fuji, H.; Nakamura, D.; Satou, N.; Okano, K. Principal Component Analysis applied directly to Sequence Matrix. *Sci. Rep.* **2019**, *9*, 19297. [Google Scholar]

[8] Mallik S. et al., (2022). "Dynamic thresholding based FP- Growth used for treating unusual illnesses and certain types of diseases with unknown variation with different symptoms. *Obstetrics & Gynecology*, vol. 127, no. 4, Apr. 2022, p. 800, 10.1097/aog.0000000000001363.

[9] Mallik, S.; Mukhopadhyay, A.; Maulik, U.; Bandyopadhyay, S. Integrated analysis of gene expression and genome-wide DNA methylation for tumor prediction: An association rule *mining-based approach. In Proceedings of the 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Singapore, 16–19 April 2013; pp. 120–127. [Google Scholar]*

[10] Huang S. Keshavjee, K.; Guergachi, A.; Gao, X.(2022). "Classification based illness prediction in a supervisory approach. Includes various mechanism such as linear and polynomial regression, vol. 2, no. 1, Jan. 2022, pp. 2–3, 10.1016/s2213-8587(13)70116-5.

[11] Parry, R.M.; Jones, W.; Stokes, T.H.; Phan, J.H.; Fang, H.; Fischer, M.; Tong, W.; et al. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharmacogenom. J.* **2010**, *10*, 292–309. [Google Scholar]

[12] Deepti Sisodia and Dilip Singh Sisodia (2018)Prediction of Diabetes using Classification Algorithms.Procedia Computer Science.132,pages 1578-1585 url={https://api.semanticscholar.org/CorpusID:267857241}

[13] Ahamed, B Shamreen, Arya, Meenakshi, Skb, Sangeetha, Nancy, V Auxilia (2022). Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers. Applied Computational Intelligence and Soft Computing. pages 1-11. DOI - 10.1155/2022/7899364

[14] Olaniyi and Adnan (2018). "Multilayer Feed-Forward Neural Network, vol. 7, no. 3, Jan. 2000, p. 42, 10.1016/s1322-7696(08)60378-9.

[15] Ashiquzzaman, R.K.; Arivuselvan, K. Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier. *Int. J. Cogn. Comput. Eng.* **2020**, *1*, 55–61. [Google Scholar]

[16] AnifatO, Rotroff, D.; Ma, J.; Shojaie, A (2022). "Artificial neural networks based on illness prediction mechanisms "journal of Neural Network, vol. 19, no. 1, 14 Apr. 2022, pp. 391–403, 10.1007/s40200-020-00520-5. Accessed 3 Nov. 2022.