An Effective Approach for Detecting Diabetes using Deep Learning Techniques based on Convolutional LSTM Networks

P. Bharath Kumar Chowdary[1]*
Research Scholar, Department of Computer Science and Engineering, BIST, Bharath Institute of Higher Education and Research (BIHER), India

Dr. R. Udaya Kumar[2]
Research Supervisor, Professor, Department of Information Technology, BIST, Bharath Institute of Higher Education and Research (BIHER) Institution, India

***Abstract*—The most common disorder affecting millions of population worldwide due to insufficient release of insulin by pancreas is diabetes. Early detection or precaution of diabetes is necessary, otherwise leads to many complicated problems. Predicting diabetes at early stages with appropriate treatment, individuals can maintain a happy life. If the conventional diabetes detection method is tedious, the identification of diabetes from clinical and physical data requires an automated system. This paper proposes an approach to enhance diabetes prediction using deep learning techniques. Based on the Convolutional Long Short-term Memory (CLSTM), we developed a diabetes classification model and compared with the existing methods on the Pima Indians Diabetes Database (PIDD). We assessed the findings of various classification approaches in this study. The proposed approach is further improved by an efficient pre-processing mechanism called multivariate imputation by chained equations. The outcomes are promising compared to existing machine learning approaches and other research models.**

***Keywords*—Convolutional long short-term memory; diabetes prediction; machine learning; pre-processing**

I. Introduction

Diabetes is affecting the world's elderly population in a very drastic way [1]. By 2019, 463 million individuals around the globe had diabetes. It is expected by the International Diabetes Federation (IDF) that the number of patients rises to 700 million individuals in near future.

Diabetes occurs due to the inconsistency of glucose levels in the blood. Usually, diabetes is classified into type 1 and type 2 diabetes. Type 1 diabetes is due to little insulin production and type 2 occurs due to blood cells becoming insulin resistant. The fundamental cause of diabetes remains unclear, but scientists agree that diabetes plays a significant role in both genetic factors and environmental lifestyles. And though it is incurable, therapy and medicine can handle it by maintaining the levels in check.

Diabetes slowly causes different diseases in the long run. Mainly it affects the heart, nervous systems, retina, kidneys and other internal organs. The care taken at the early stages of diabetes helps in avoiding the damage of various organs. Although it is a chronic problem, researchers handled this by developing various prediction systems using machine learning algorithms [2]. The most popular algorithms were Support Vector Machine (SVM), Decision Trees, and Random Forest.

Another popular model to predict diabetes is an Artificial Neural Network (ANN) [2]. It is well-known for its high precision and performance. Present research includes Deep Learning (DL) for prediction due to the increasing size and complexity of data.

Recent studies [3] using DL have enhanced various prediction and classification parameters like accuracy and precision. PIMA diabetes dataset [4] is used by many researchers to test their models.

Diabetes occurs when the body is unable to metabolize the glucose. The body is unable to produce or react to the insulin produced in the case of diabetes. Once diabetes is attacked, it is tough to cure. Hence, the knowledge of how diabetes occurs helps individuals to prevent it. Early diagnosis helps in reducing the risk for the patient.

Practitioners require high amount of data. The healthcare industry collects a large amount of health-related data, but this data cannot perceive undetected patterns of good decision-making [5]. It is a tedious job for any individual to process a high amount of data. As a result of this, researchers developed various machine learning and classification techniques to handle the data.

This paper has used Traditional LSTM and convolutional LSTM models for prediction on the PIMA dataset. We have performed extensive experimentation using data mining algorithms such as decision trees (DT), Naïve Bayes classification, ANN, and DL to provide an insight into how different algorithms work for diabetes prediction. In a logical and well-organized way, the comparison of algorithms is interpreted, with more efficient and prominent results provided by DL. DL is a self-learning framework for knowledge used successfully to predict diabetes.

II. Related Work

A. Diabetes Prediction using Machine Learning (ML) Algorithms

ML algorithms are used by researchers to predict diabetes. The most famous approaches are SVM, J48, K-Nearest Neighbours (KNN), and Random Forest classifiers [8]. Ioannis

*Corresponding Author

et al. [7] applied ML and data mining (DM) techniques for diabetes prediction. This work [7] mainly focused on analysing the existing techniques in ML and DM. The authors have done extensive research on different databases containing diabetic data.

Zhu et al. [9] used a logistic regression-based model to predict diabetes. The authors have used principal component analysis and k-means algorithms to classify the developed model data correctly. The authors [10] developed a prediction model on diabetes data using classifiers based on the decision tree, naïve Bayes and random forest.

The authors [11] also used various MLK algorithms to classify diabetes data. This work [11] majorly focused on using decision trees and SVM to classify PIMA diabetes data. Dataset partitioning is carried out using a 10-fold method of cross-validation. The authors have not performed data pre-processing.

Negi and Jaiswal [12] also applied SVM to diabetes prediction on PIMA and Diabetes 130-US datasets.

The authors tested the existing ML algorithms on various datasets to predict diabetes. But the data consists of missing values and requires pre-processing. We are using data pre-processing techniques to enhance diabetes classification. The next part of this section covers various deep neural network models for diabetes prediction.

B. Deep Neural Networks

In the analysis of large datasets, researchers have begun to realize the capabilities of DL techniques [6]. Therefore, using DL techniques, diabetes prediction has also been carried out.

The authors [13] used a Deep Neural Network (DNN) for diabetes prediction. This approach was tested on the PIMA dataset. As DNN can filter the data and develop biases, the authors did not deliberately pre-process the dataset. For the research collection and the rest of the research, the dataset is divided into 192 samples. 88.41 percent was the accuracy rate stated by the authors.

Another approach [14] based on CNN and CNN-LSTM is developed to test the Electrocardiograms dataset.

The authors [15] used the logistic regression model as a basis for the multilayer neural network and CNN. The dataset used by authors [15] consists of nine patients. For each patient nine features are gathered. Moreover, each patient had data for 10,800 days, resulting in a total of 97,200 simulated days. There was no proper discussion of the attributes used in this analysis.

Miotto et al. [16] proposed the Deep Patient model, which is an unsupervised DNN. This model is used to classify electronic health records. The model is tested on a database consisting of 704,857 patients.

The authors [17] tested various deep learning methods on Australian hospital health records and developed a dataset.

The authors [18] used RNN model to predict both type 1 and type 2 diabetes. The authors used the PIMA dataset and predicted that the attribute “Glucose” has the highest significance followed by BMI, age, births, pedigree feature of diabetes, blood pressure, thickness of the skin and insulin.' The training dataset and 20 percent for the testing were split into 80 percent to validate the analysis.

This paper proposes a convolutional neural network with enhanced feature selection and data pre-processing mechanisms for diabetes prediction. The later section provides the proposed methodology.

The existing models fail to extract the features properly. The existing models failed to properly incorporate the data pre-processing techniques. The existing models do not fill the missing values. Moreover, neural networks and error propagation are not implemented by existing models. The proposed model overcomes all the above-mentioned problems and enhances the Diabetes prediction task. The remaining part of the paper is as follows. Section 2 gives the proposed methodology. The fourth section gives dataset description and selected results of the existing and proposed method.

III. Materials and Methods

90 percent of all forms of diabetes are diabetes types II. This disorder causes insulin resistance or insulin loss problems for the victim. The age at which diabetes type II typically takes place is 40 years old. Youth under the age of 30 are at risk for this disease with current eating habits and lifestyle. Early detection with routine checks and surveys allows people to diagnose the disease early and to take precautions.

Various research attempts were made to enhance the accuracy and applicability of various Clinical Decision Support Systems (CDSS) interpretability. However, it is still essential to optimize this issue. In the medical area, where interpretability is an essential question, fluid rules are relevant.

Many healthcare systems gain valuable information and produce a huge amount of clinical data. Machine learning techniques allow the practitioner to process this data and make quick decisions [9]. These decisions reduce the risk of diabetes, affecting the person severely, and preventing damage to other organs. Multiple machine training techniques for disease prediction and information from medical data have been developed.

The long short-term memory (LSTM) [21] is a form of RNN and consists of feedback connections. LSTM models can process a long input data sequence at ease.

A standard LSTM system consists of a cell, an entrance gate, an output gate, and a forgotten gate. The cell recalls values at arbitrary times, and the three gates monitor information flow in and out of the cell.

LSTM networks are well suited for the classification, processing, and estimation of time series data because the period of uncertain events in a time series can be delayed. LSTMs have been developed to resolve the disappearance gradient problem that can be observed during conventional RNN training. Relative lack of attention to the length of gaps is an advantage of LSTM in multiple applications over RNNs, hidden Markov models, and other sequence learning methods.

Compared with a popular recurrent unit, an LSTM cell has the benefit of its cell memory unit. The cell vector can encapsulate the concept of missing some of its formerly saved memory and add some of the new details. The cell equations and the sorting of sequences under the hood must be inspected to demonstrate this.

A. Traditional LSTM

A LSTM network comprises of memory cell and four gates. The four gates in LSTM network are a) forget gate (f), b) input gate (i), control gate (c) and output gate (o) [19].

The underlying data pattern can be extracted and remembered, which addresses long-term data dependence on classic RNN algorithms [19]. Fig. 1 shows the TLSTM architecture [20]. Inputs of the architecture are h_{t-1}, x_t, and b. The term h_{t-1} represents previous cell sate, x_t represents current input vector and b represents bias. One of the outputs of the architecture is c_t, which represents the present memory content. Another output of the architecture is h_t, represents present cell state. These four gates listed above influences the data in the memory cell. Forget gate gives a value in the range 0-1. This value defines how much should be ignored from the previous memory cell. If the forget gate produces a value close to 0 it means that at the new time stamp, much of the previous timestamp's memory will be overlooked and the reverse occurs for the value close to 1. The gates in TLSTM are represented in the following equations as follows:

Equation 1 represents the forget gate of TLSTM [20] as,

$$f_t = \alpha_g(w_f x_t + u_f h_{t-1} + b_f) \quad (1)$$

Equation 2 represents the input gate of TLSTM as,

$$i_t = \sigma_g(w_i x_t + u_i h_{t-1} + b_i) \quad (2)$$

Equation 3 represents the control gate of TLSTM as,

$$c_t = f_t \times c_{t-1} + i_t \times \sigma_h (w_c x_t + u_c h_{t-1} + b_c) \quad (3)$$

Equations 4 and 5 represent the output of TLSTM,

$$o_t = \sigma_g(w_o x_t + u_o h_{t-1} + b_o) \quad (4)$$

$$h_t = o_t \times \sigma_h c_t \quad (5)$$

Here, the sigmoid function is represented by σ_g and hyperbolic tangent function is denoted by σ_h. The symbols w and u represent weights. These weights usually prevent the issue of gradients from disappearing.

We have used 50 T-LSTM units in each layer. In each layer, for every input an attention value is calculated. Attention value gives the significance of the input and is helpful in final prediction. The dense layer allows the final prediction of whether a patient has diabetes with the aid of an attention vector.

It can be noticed from in Fig. 1, that there is no correlation between the previous memory content with any of the gates in the network. This results in an abnormal situation if the output gate is locked. This reduces the efficiency of prediction and classification tasks. Hence, the primary goal of this work is to apply CLSTM to the classification of patients with diabetes and to illustrate how CLSTM overcomes the limitations faced by TLSTM.

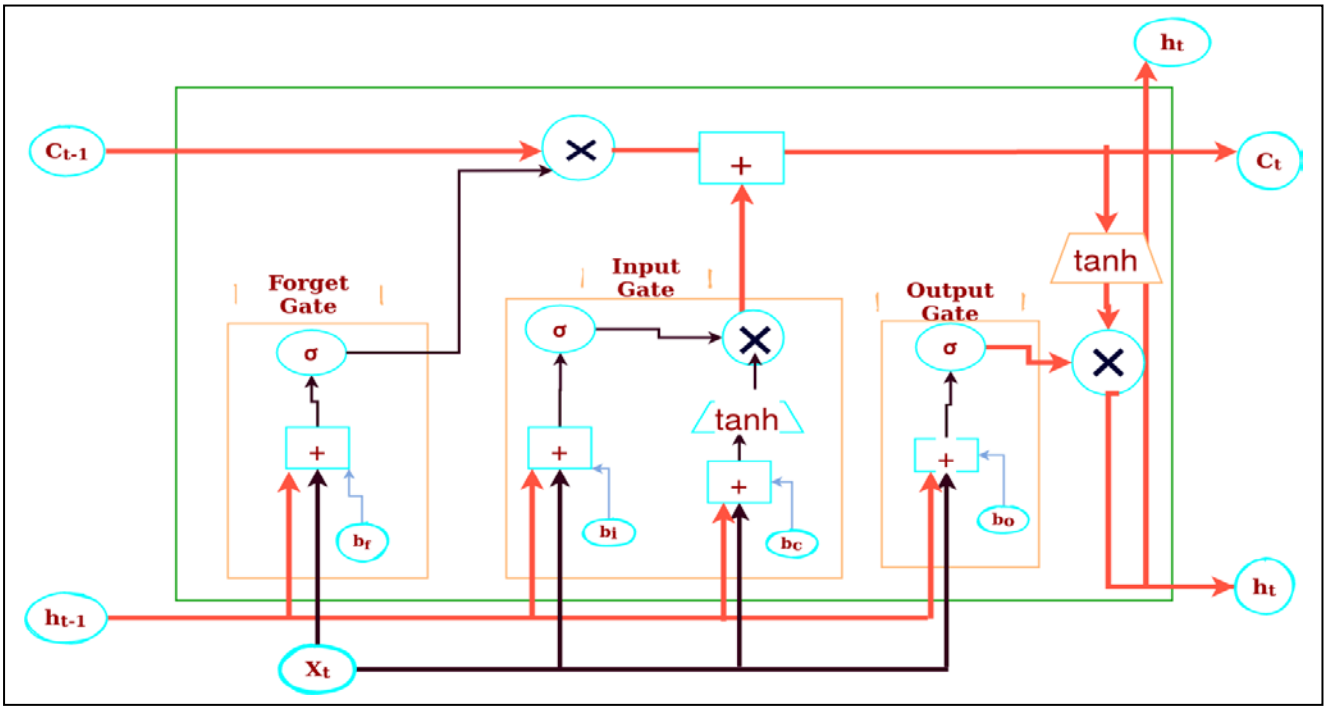

Fig. 1. Architecture of Traditional LSTM.

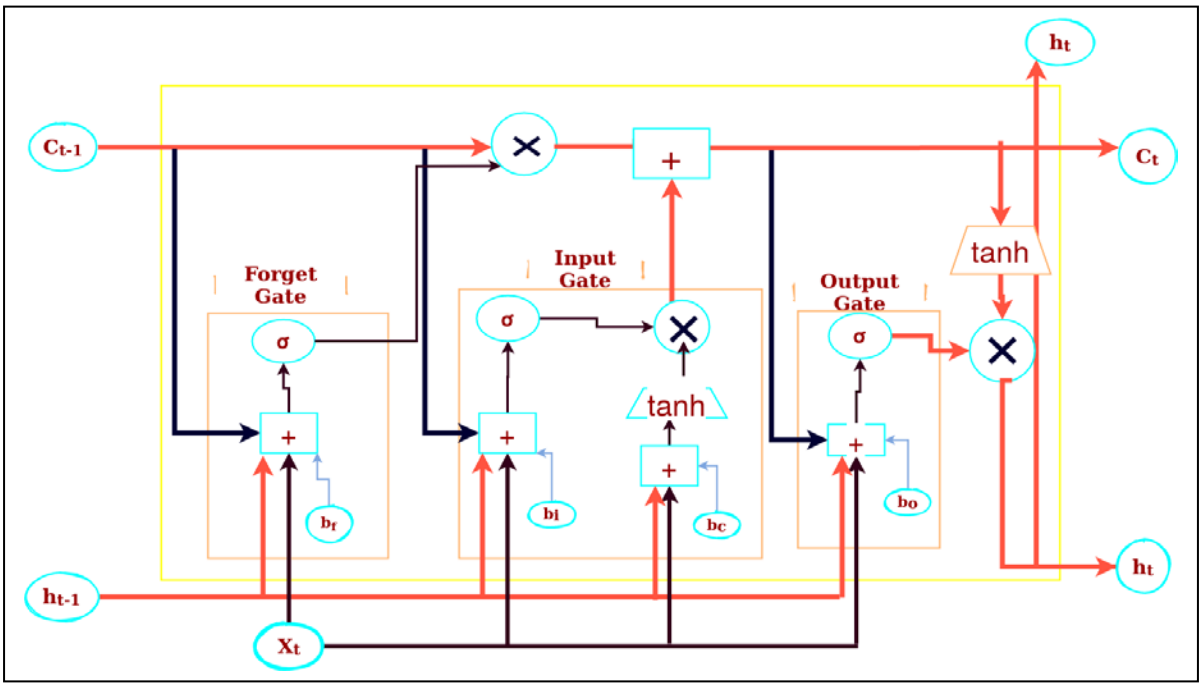

Fig. 2. Architecture of Convolutional LSTM.

B. Convolutional LSTM

Traditional LSTM do not access previous memory cell contents [19] even when the output gate of the model is closed. CLSTM negates this by adding an extra link to all the other gates from the previous memory. Fig. 2 [20] shows CLSTM diagram and its operation.

An additional parameter (former c_{t-1} memory content) is in CLSTM compared to TLSTM to provide previous memory cells' impact even when the output gate is closed.

The CLSTM four gates function with the help of the following equations:

Equation 6 represents the forget gate of CLSTM as,

$$f_t = \sigma_g(w_f x_t + u_f h_{t-1} + v_f c_{t-1} + b_f) \quad (6)$$

Equation 7 represents the input gate of CLSTM as,

$$i_t = \sigma_g (w_i x_t + u_i h_{t-1} + v_i c_{t-1} + b_i) \quad (7)$$

Equation 8 represents the control gate of CLSTM as,

$$c_t = f_t c_{t-1} + i_t \sigma_h (w_c x_t + u_c h_{t-1} + b_c) \quad (8)$$

Equations 9 and 10 represents the output of CLSTM as

$$o_t = \sigma_g (w_o x_t + u_o h_{t-1} + v_o c_{t-1} + b_o) \quad (9)$$

$$h_t = o_t \times \sigma_h (c_t) \quad (10)$$

This article developed a CLSTM-based model of diabetes prediction and is tested on the Pima Indian Diabetes dataset.

C. Proposed Model

Fig. 3 represents the proposed model for diabetes prediction. Initially, the PIMA dataset is pre-processed and later, essential features are selected. For evaluation and training purposes, the dataset is split into train and test sets.

The hyperparameters of TLSTM and CLSTM models are tuned in the next step. Once the training phase on the dataset is completed, we have calculated various parameters for performance evaluation. The accuracy for different test sizes are measured in the paper in order to estimate the performance of the proposed model.

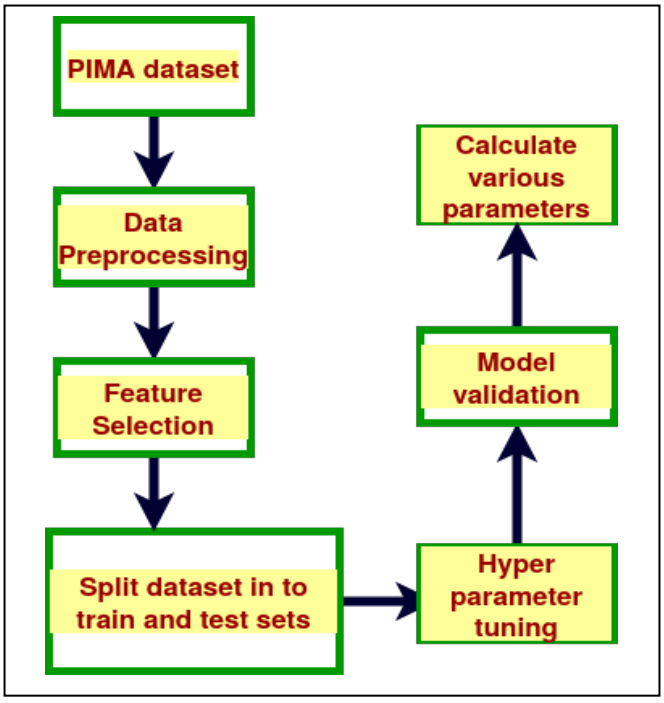

Fig. 3. Proposed Model for Diabetes Prediction.

D. Dataset Description

The initial process of our strategy is to apply dataset pre-processing techniques on the PIMA dataset. The dataset contains information about 768 patients, with nine attributes obtained for each patient. The data in the dataset consists of different female individuals between the ages of 21 and 81.

Six attributes represent physical examination specifics in each row, and the remaining attributes represent chemical examination information. The last attribute in-row is the data on whether the patient is diabetic.

The last column of each row is either 1 or 0, 1 indicating that the patient is diabetic and 0, indicating that the patient is not diabetic.

The first column in the dataset represents the number of times a woman is pregnant, and the second column in the dataset represents the plasma glucose concentration. The third column in the dataset depicts the diastolic blood pressure and the fourth column gives the thickness of the triceps skin fold. The fifth column represents serum insulin for two hours, and the sixth column represents the person's body mass index (BMI). Pedigree feature is in the seventh column and the eighth column in the dataset reflects the individual's age and the last column reflects the incidence of diabetes (1/0). Fig. 4 shows the information of various attributes of the PIMA dataset. Fig. 5 shows the correlation of various attributes in the PIMA dataset.

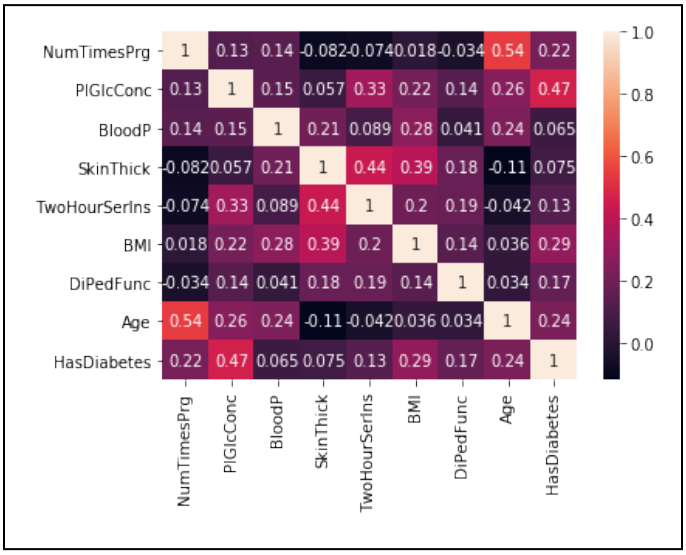

Fig. 4. Correlation of Various Attributes in the PIMA Dataset.

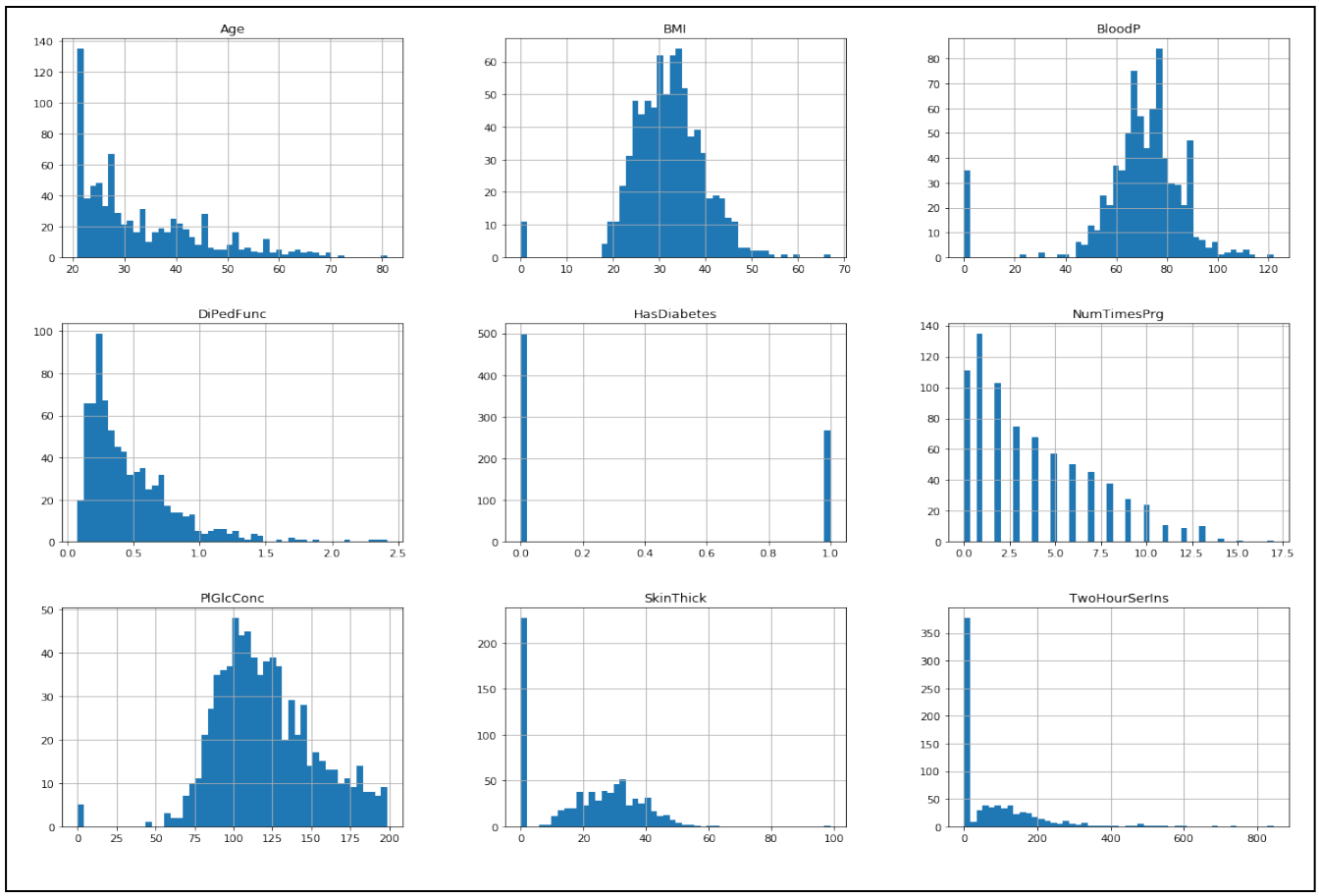

Fig. 5. Attributes in the PIMA Dataset.

IV. Results and Discussion

The comparison of various models like neural networks, machine learning and deep learning systems are presented in this section.

A. Experimental Setup

The TLSTM, CLSTM models are used in this paper to predict feature selection. Initially we have pre-processed the dataset with the mentioned techniques in the previous section. Random Forest algorithm is used for feature selection. We have found from our observation that five features (Glucose, Age, BMI, BP, Insulin) as important.

We have set the TLSTM and CLSTM models' hyperparameters with the following details mentioned in Table 1.

The values in Table 1 are hyperparameter optimization values where we obtained highest accuracy. We have used python inbuilt packages to develop our model. Pre-processing and feature selection of dataset are also carried out using python.

Table 2 presents the results of different models on the PIMA dataset. Naïve Bayes, SVM, DT, K-means have similar accuracy results. TLSTM and CLSTM models outperformed the accuracy results of other existing models. The machine learning algorithms reported in this section are traditional ones.

In Table 2 all the results are obtained from our experimentations. The results show that the TLSTM and CLSTM models outperformed all the existing machine learning models.

TABLE I. Hyperparameters of TLSTM and CLSTM

Parameter	TLSTM	CLSTM
Learning Rate	0.02	0.01
Batch size	32	32
Hidden layers	50	50
Epoch	50	50

TABLE II. Comparison of Various Models on PIMA Dataset

Model	Accuracy (test set 10%)	Accuracy (test set 20%)
Naïve Bayes	79.6%	78.6%
SVM	79.2%	78%
Decision Trees	78.4%	77.2%
MLP	80%	82%
K means	77%	72%
TLSTM	92.5%	93.7%
CLSTM	96.8%	95.6%

The results presented in this section specifies that the proposed model outperforms all the existing models. The TLSTM and CLSTM models have obtained higher accuracy results than all the existing machine learning models. The machine learning models do not capture the features properly and hence the results are less when compared with the proposed model. Moreover the proposed model takes care of the data pre-processing and feature selection properly and hence the results are high for our model.

V. Conclusion

This paper aims to introduce a CLSTM, TLSTM prediction model for diabetes. As diabetes is becoming a serious disorder now-a-days it is the need of the hour if the researchers come up with prediction models. The proposed approach enhances diabetes prediction using deep learning techniques. Moreover, the proposed approach also uses an efficient pre-processing mechanism called multivariate imputation by chained equations. This paper examines various classification approaches on the PIMA dataset. Existing ML and DL approaches are tested on PIMA dataset. As mentioned in Table 2, the result achieved by CLSTM model is higher than other methodologies. In the future, in the form of an application or a website, we plan to build a comprehensive framework using CLSTM algorithm, which will help practitioners to predict diabetes at early stages and reduce the risk of various diseases

References

[1] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. D. R. Fernandes, A. Ohlrogge and B. Malanda, “IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,” Diabetes Res. Clin. Pr., vol. 138, pp. 271–281, 2018.

[2] Y. L. Sun and D. L. Zhang, “Machine Learning Techniques for Screening and Diagnosis of Diabetes: A Survey,” Teh. Vjesn., vol. 26, pp. 872–880, 2019.

[3] H. Naz and S. Ahuja, “Deep learning approach for diabetes prediction using PIMA Indian dataset,” Journal of Diabetes & Metabolic Disorders, vol.19(1), pp.391-403, 2020.

[4] D. Liccardo, A. Cannavo, G. Spagnuolo, N. Ferrara, A. Cittadini, C. Rengo and G. Rengo, “Periodontal disease: A risk factor for diabetes and cardiovascular disease,” International journal of molecular sciences, vol. 20(6), pp.1414, 2019.

[5] B. P. Nguyen, H. N. Pham, H. Tran, N. Nghiem, Q. H. Nguyen, T. T. Do, C.T. Tran and C. R. Simpson, “Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records,” Computer methods and programs in biomedicine, vol. 182, 2019.

[6] S. Spanig, A. Emberger-Klein, J. P. Sowa, A. Canbay, K. Menrad, and D. Heider, “The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes,” Artificial intelligence in medicine, vol. 100, 2019.

[7] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research,” Computational and structural biotechnology journal, vol. 15, pp.104-116, 2017.

[8] J. P. Kandhasamy and S. Balamurali, “Performance Analysis of Classifier Models to Predict Diabetes Mellitus,” Procedia Comput. Sci., vol. 47, pp. 45–51, 2015.

[9] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Informatics in medicine Unlocked, vol. 17, 2019.

[10] Z. Tafa, N. Pervetica, and B. Karahoda, "An intelligent system for diabetes prediction," In Proceedings of the 2015 4th Mediterranean Conference on Embedded Computing (MECO), pp. 378–382, 2015.

[11] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Comput. Sci., vol. 132, pp. 1578–1585, 2018.

[12] A. Negi and V. Jaiswal, "A first attempt to develop a diabetes prediction method based on different global datasets," In Proceedings of the 2016 Fourth International Conference on Parallel, Distributed and Grid Computing, pp. 237–241, 2016.

[13] A. Ashiquzzaman, A. Kawsar Tushar, M. D. Rashedul Islam, D. Shon, L. M. Kichang, P. Jeong-Ho, L. Dong-Sun and K. Jongmyon, "Reduction of overfitting in diabetes prediction using deep learning neural network," In IT Convergence and Security; Lecture Notes in Electrical Engineering; Springer, vol. 449, 2017.

[14] G. Swapna, K. P. Soman and R. Vinayakumar, "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals," Procedia Comput. Sci., vol. 132, pp.1253–1262, 2018.

[15] A. Mohebbi, T. B. Aradóttir, A. R. Johansen, H. Bengtsson, M. Fraccaro and M. Mørup, "A deep learning approach to adherence detection for type 2 diabetics," IEEE Engineering in Medicine and Biology Society, pp. 2896–2899, 2017.

[16] R. Miotto, L. Li, B. A. Kidd and J.T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," Appl. Sci., vol.6, pp. 4604-4612, 2019.

[17] T. Pham, T. Tran, D. Phung and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," J. Biomed. Inform., vol.69, pp.218–229, 2017.

[18] H. Balaji, N. Iyengar and R. D. Caytiles, "Optimal Predictive analytics of Pima Diabetics using Deep Learning," Int. J. Database Theory Appl., vol. 10, pp. 47–62, 2017.

[19] G. Zhu et al., "Redundancy and Attention in Convolutional LSTM for Gesture Recognition.," IEEE Trans. neural networks Learn. Syst., Jun. 2019.

[20] G. Zhu, L. Zhang, L. Yang, L. Mei, S. A. A. Shah, M. Bennamoun, and P. Shen, "Redundancy and attention in convolutional LSTM for gesture recognition," IEEE transactions on neural networks and learning systems, vol. 31(4), pp.1323-1335, 2019.

[21] Rahman and Siddiqui, "An Optimized Abstractive Text Summarization Model Using Peephole Convolutional LSTM," Symmetry (Basel)., vol. 11,2019.