SN Computer Science (2023) 4:373
https://doi.org/10.1007/s42979-023-01831-z

ORIGINAL RESEARCH

Diabetes Prediction Using Bi-directional Long Short-Term Memory

Sushma Jaiswal[1] · **Priyanka Gupta**[1]

Received: 19 February 2022 / Accepted: 10 April 2023 / Published online: 3 May 2023

Abstract
The chronic nature of diabetes makes it the most complex and lethal health issue globally. This creates difficulties in managing this illness in the day-to-day life of a large population of affected people. Diabetes is also known as diabetes mellitus (DM). This research aims to reduce diabetic patients' pain and make their lives easier and more comfortable. This paper describes an expert system for diabetes mellitus classification. This study's evaluations presided over the Pima Indian Diabetes (PID) data set. We proposed a bi-directional long short-term memory (BLSTM)-based approach to diagnose diabetes mellitus at an earlier stage. The proposed methodology is novel, and the model is tuned for maximum attainment by varying the divergent range of parameters. The network attributes are trained using a tenfold cross-validation approach. Data pre-processing techniques play a vital role in this case. The class-balancer synthetic minority oversampling technique (SMOTE) was used to balance the data. The proposed BLSTM model provides better accuracy, sensitivity, specificity, and *F*1 score with values of 94%, 96%, 91%, and 93%, respectively. Our results demonstrate that the proposed model outperforms earlier studies in classifying diabetic patients.

Keywords ML · Classification · PID · DM · LSTM · BLSTM

Introduction

Diabetes mellitus is a silent disease; many individuals live with it for years without being aware of it. It is regarded as the most prevalent, incurable disease in the world [1]. According to Wikipedia statistics, the global prevalence of diabetes in 2019 is 9.3% (465 million people), which will climb to 10.2% (580 million people) by 2030, and 10.9% (700 million people) by 2045. The ballpark estimate for the number of diabetic patients who pass away each year is 2–5 million. The initial stage of diabetes is pre-diabetes. Diabetes is a chronic illness, also called a lifelong disease. Insulin, a hormone, is required by the body to convert glucose into energy. Predominantly there are three types of diabetes and one prediabetic condition. Type 1 diabetes develops when the process of producing insulin by pancreatic cells becomes insufficient. Type 2 diabetes occurs when the pancreas produces insulin, but the body cannot utilize it productively. The frequency of type 2 diabetes is rising, resulting in health problems for people worldwide.

In contrast, gestational diabetes is primarily seen in pregnant women. Sometimes it goes away after pregnancy, but other times it persists, much like type 2 diabetes, putting both mother and baby at risk. Borderline diabetes, also known as prediabetes, is a condition in which a person's blood sugar levels are elevated but not high enough to be classified as diabetes. Diabetes has numerous symptoms, including blurry vision, extreme thirst, frequent urination, slow wound healing, hunger, and tingling in the hands or feet.

The use of technology to improve health care has proven to be quite effective. This research tried to predict diabetes using a deep learning approach. Currently, the classification of data into various classes is one of the essential functions of deep learning (DL), machine learning (ML), and artificial intelligence (AI) in medical diagnostics [2]. Through machine learning, Knowledge Engineering accommodates a bridge of procedures and techniques; the computer system can change unprocessed data into some prosecutable, significant orientation. Machine learning methods that are immediately owned are classified as supervised, unsupervised,

✉ Priyanka Gupta
priyanka13666@gmail.com

Sushma Jaiswal
jaiswal1302@gmail.com

[1] Computer Science & Information Technology, Guru Ghasidas Vishwavidyalaya, Bilaspur (C.G.), India

semi-supervised, and reinforcement. Classification and regression are dealt with in supervised learning. The first obligation is data prophecy because the facsimile is data confirmation. The training practice of the supervised learning model is affected by external supervision. An unsupervised learning algorithm derives the outcomes from unlabeled training data and assists in forecasting outcomes for unlabeled data. Cluster-based analysis and Association rules are essential to unsupervised learning. Semi-supervised learning made use of both labeled (supervised) and unlabeled (unsupervised) data. Lastly, Reinforcement Learning is an assessment-situated Machine learning technique. The model learns to enforce within the domain by interacting with it. Following that, it performs the following action and changes states based on the feedback from the previous step. In this study, the BLSTM-based deep learning model for predictive analytics is implemented on the PID healthcare dataset. The BLSTM model gives more accurate and efficient results for classifying diabetes mellitus. For a variety of outcome evaluations, confusion metrics are used. This investigation aims to more accurately forecast whether a specific observation is at risk of diabetes or not.

The remaining sections of the paper are arranged as follows: the next section is the introduction to chronic disease. The subsequent section contains the related work in this field followed by which the proposed methodology is reported with the interpretation. The penultimate section discusses the practical implementation of the outcome. Finally, the conclusion and future works are given.

Related Work

As the medical health state of a person in the nation is a responsibility of healthcare organizations and professionals, numerous researchers have carried out a wide variety of studies in the past to acquire information on diabetes mellitus. Researchers utilize a variety of data mining, expert systems, and conventional methodologies in addition to a few deep learning models for predictive analysis of diabetes mellitus.

Saxena et al. [3] employed ML techniques such as multilayer perceptron, decision trees, K-nearest neighbours, and random forests in their study. In order to impute the missing data, they utilized a mean technique. The accuracy of the models for MLP, decision trees, K-nn, and random forest, respectively, is 77.60%, 76.07%, 78.58%, and 79.8%. The approach of stacking ensemble learning was proposed by Shivashankari [4]. They proposed that the current diabetic prediction models utilize a single algorithm. However, the unstructured and enormous datasets will not be acceptable for a single algorithm. With a classification accuracy of 93.1%, the suggested stacked ensemble model beat other current models for diabetes classification, including logistic regression (72%), Naive Bayes (74.4%), and LDA (81%). Wu et al. [5] have devised the K-means clustering method and the logistic regression (LR) model for diabetes prediction. Weka toolkit gives better results on Pima Indian Diabetes Dataset (PID). Syed et al. [6] employed the Chi-squared test to determine the most significant diabetes risk factor in the T2DM risk prediction model. SMOTE is used to balance the class. The two-dimensional Decision Forest algorithm obtained accuracy: 0.821, precision: 0.776, recall: 0.890, AUC: 0.867, and F1 score: 0.829 during the performance study, the model demonstrating a greater efficacy. Choubey et al. [7] recommended that Principal Component Analysis (PCA) is Superior to Particle Swarm Intelligence (PSO) for the Localized Diabetes Dataset and Pima Indian dataset. They used the WEKA library for different ML algorithms. Agarwal et al. [8–17] implemented ML techniques such as Decision Trees (DT), LR, Naïve Bayes, and Support Vector Machine (SVM), K-nn, Linear Discriminant Analysis (LDA), Random Forest (RF), and Artificial Neural Network (ANN) on PID Dataset. According to Suja A. et al. [18] findings, the PID dataset is used for diabetes diagnosis, and the Tukey method was employed to identify outliers. They used the Standard Scalar technique to normalize the dataset. After that, the SMOTE method is used to pre-process the data. Application of the deep convolutional network (DCNN) classifier resulted in 86.29% accuracy, 81% precision, 84% recall, and 91% AUC. Xu et al. [19] used XG-Boost classifiers on the PID dataset. The outgrowth of this utilization shows the consequence of data pre-processing techniques to classify the model, the accuracy of the XG-Boost classifier can be enhanced up to 88.28%. Xiaohua Li et al. [20] studied various metaheuristic algorithm combinations to increase the accuracy of KNN in diabetes diagnosis. The accuracy achieved by the hybrid model is 88.02%, 89.64%, and 91.65% for the three hybrids: GA-Kmeans, GA-PSO-Kmeans, and Harmony-Kmeans (HR-Kmeans). Zhou et al. [21] determined the essential characteristics of diabetic diseases by adding dropout regularization to address the overfitting problem on the PID dataset; the deep learning model achieved better accuracy. Hang Lai et al. [22] implemented a gradient boosting machine and logistic regression model. The AROC for the proposed model is 84.7% and 84.0%, with a sensitivity rate of 7.6% and 73.4%, respectively. Sainte et al. [23] focused on the DL technique convolutional neural network long short-term memory (CNN-LSTM) and PID dataset is used for prognostication of diabetes mellitus. The obtained accuracy achieved by the authors is 68–74%. Fathi et al. [24], the author's point of convergence is to develop a decision support system for people with Type 1 diabetes. The framework is adequately suitable for clinical data from 15 participants with physiologically plausible authorized system parameters. Roopa H. et al. [25] implemented

PCA-LRM. They use PCA for dimensionality reduction and apply the input features to the LR model. The model achieves 82.1% accuracy, F measure 86%, and precision of 81% on the PIMA diabetes dataset. Jayanthi et al. [26] Surveyed and gave comparative elaboration and predictive analysis of traditional and hybrid approaches. The studies show that ensemble or hybrid models provide better outcomes for the prediction of diabetes mellitus as compared to conventional machine learning models. According to the Saeedi et al. [27] survey, the global diabetes incidence occurrence is higher in cities (10.8%) than in rural areas (7.2%), and in high-income countries (10.4%) than in low-income countries (4.0%). One in every two (50.1%) diabetics is unaware that they have the disease. Ahmad et al. [28] the author uses Electronic Health Records that have been taken possession from five Saudi hospitals covering three central, eastern, and western regions. The dataset contains approximately 3000 records of diabetic patients assembled from 2016 to 2018 through various parts such as outpatient, inpatient, and emergency. The instate dataset incorporated 16 numerical, binomial, polynomial, and date types of attributes. FPG-labeled dataset gives a better result as compared to HbA1c-labeled dataset.

Research Gap and Motivation

Data imbalance and missing values in the data set are one of the most challenging issues in ML prediction [29, 30]. Many studies [31] used an upsampling approach, which essentially repeats the same data to balance the data set. However, this sampling approach has its limitations. Also noted by Roshi Saxena et al. [32] that the outcome of traditional machine learning classifiers is not much accurate. The models in medical data must give more precise results because accuracy is a crucial metric for diagnosing diabetes. Therefore, to overcome this research gap, there is a need to develop a deep learning-based strategy to detect diabetes at an early stage and improve the overall accuracy of diabetes prediction.

Contribution

Our research work's novelties and major contributions are as follows:

1. To the best of our knowledge, for the first time in literature, we proposed a BLSTM-based architecture for diabetes prediction on the PID dataset.
2. We are using Incremental Principal component analysis for feature reduction.
3. Optimizing the data set by addressing class imbalance issues, handling missing values, and rejecting outliers.
4. The proposed framework provides better accuracy, sensitivity, specificity, ROC-AUC, and $F1$ score Compared to previous state-of-the-art approaches.

Proposed Methodology

Dataset and data pre-processing techniques both have significant value for building the model. We employed a tenfold cross-validation technique for data partitioning and validating the model [33]. This gives assistance in overcoming the "Sample Variability" problem. The hyperparameter tuning plays a crucial role in the BLSTM model for predicting diabetes (see Fig. 1).

Dataset Description

In the suggested technique for predicting diabetes mellitus, we used the PID (Pima Indian Diabetes) dataset. According to prior studies [34–36], the Pima Indian Diabetes dataset is a standard and widely used dataset. It is one of the most often used databases. Data acquisition is obtained from the UCI machine learning repository. This dataset is primarily based on females between 21 and 80 years old and belonging to PIMA Indian heritage. The dataset comprises 768 records and 8 characteristics (Independent Variables) with one outcome (Dependent Variable) (see Tables 1, 2).

Data Preparation (Data Preprocessing) Using Exploratory Data Analysis

Exploratory data analysis (EDA) is a method of visualizing, condensing and evaluating hidden information in a dataset in the form of rows and columns. It includes a perspective on the dataset. Following EDA, we ensure that we are working on a suitable dataset. Here we see that our dataset has a lot of zero values for several attributes. Except for pregnancies, absolute zero values are treated as missing values. In the series of existing procedures for manipulating zero values, the most straightforward approach is to blot out all the rows or columns that hold

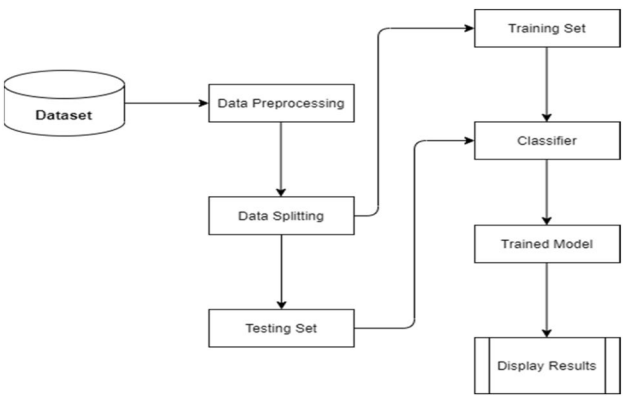

Fig. 1 Methodology of the proposed framework

Table 1 Brief description about the PID dataset [13]

Attribute	Description	Range	Missing value
Pregnancies	Numbers of times Pregnant (Numeric)	0–17	111
Glucose	Plasma glucose concentration 2 h in OGTT (oral glucose tolerance test) (Numeric)	0–199	5
Blood pressure	Diastolic blood pressure (mm Hg) (Numeric)	0–122	35
Skin thickness	Triceps skin fold thickness in mm (Numeric)	0–99	227
Insulin	2-h serum insulin (mu U/ml) (Numeric)	0–846	374
BMI	Body mass index (weight in kg/ (height in m)^2) (Numeric)	0–67.1	11
Diabetes pedigree function	Diabetes history in relatives (Numeric)	0.078–2.42	0
Age	Age in years (Numeric)	21–81	0
Outcome	Predictions of a person being diabetic or not 0—non-diabetic 1—Diabetic (True/False)	Yes/No	NA

Table 2 PID dataset statistics

Attribute	Count	Mean	STD	Min	Max
Pregnancies	768	3.84	3.36	0.00	17.00
Glucose	768	120.89	31.97	0.00	199.00
BP	768	69.10	19.35	0.00	122.00
Skin thickness	768	20.53	15.95	0.00	99.00
Insulin	768	79.79	115.2	0.00	846
BMI	768	31.99	7.88	0.00	67.10
DPF	768	0.47	0.33	0.07	2.42
Age	768	33.24	11.76	21.00	81.00

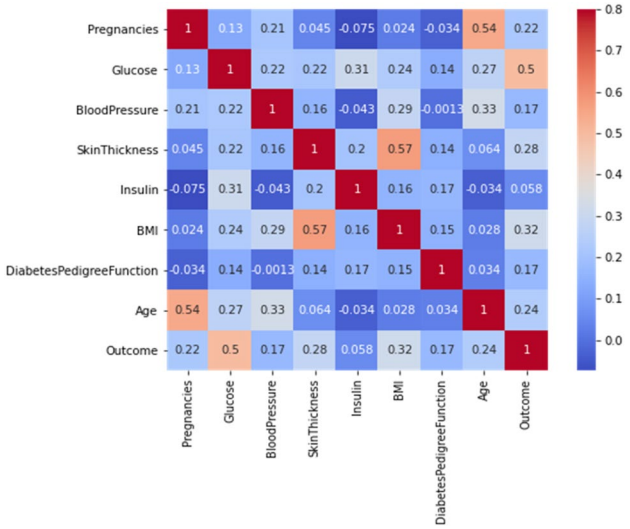

Fig. 2 Correlation between the features

missing phenomena. Nonetheless, much vital information is lost as a result of this. Data deletions cause bias in the data set because the number of records in the data set is reduced; this may cause an unsatisfactory outcome. It has been found that it is critical for predictive analysis not to remove all records (rows) or columns (columns) containing missing values. In this regard, the median strategy is accepted for re-creating missing data values, so crucial influential information stays in the dataset.

Dataset and data pre-processing techniques both have significant value for building the model. We employed a tenfold cross-validation technique for data partitioning and validating the model [33]. This assists in overcoming the "Sample Variability" problem. The hyperparameter tuning plays a crucial role in the BLSTM model for predicting diabetes.

After replacing zero values with the median, we must determine the relationship between one or more variables. Correlation measures how strongly one variable is dependent on another. It is a necessary tool for feature engineering while developing deep learning models. The correlation heat map, shown in Fig. 2, is used to fantasize about the concentration of values between two dimensions of a matrix. We can see that the degree of the association between outcome and glucose is substantial, implying that glucose is an essential factor. According to the heat map, other important factors include BMI, age, and insulin.

The numerical data distribution of the PID dataset is depicted by the violin graph in Fig. 3. A violin plot's intrinsic modules mostly depict the mean, median, and interquartile range. The violin graph demonstrates that diabetic patients have a higher blood pressure than non-diabetic people. It demonstrates that the BMI of diabetic individuals is higher than the BMI of non-diabetic people. Diabetes patients appear to have a higher pedigree function than non-diabetics. The heat map indicates that glucose is an essential feature for building the prediction model. It is possible to speculate that diabetic patients have lower insulin levels while non-diabetics have slightly greater insulin levels. Furthermore, diabetic individuals have thicker skin. Dealing with violin plots, it is notable that a pregnant woman is more prone to diabetes.

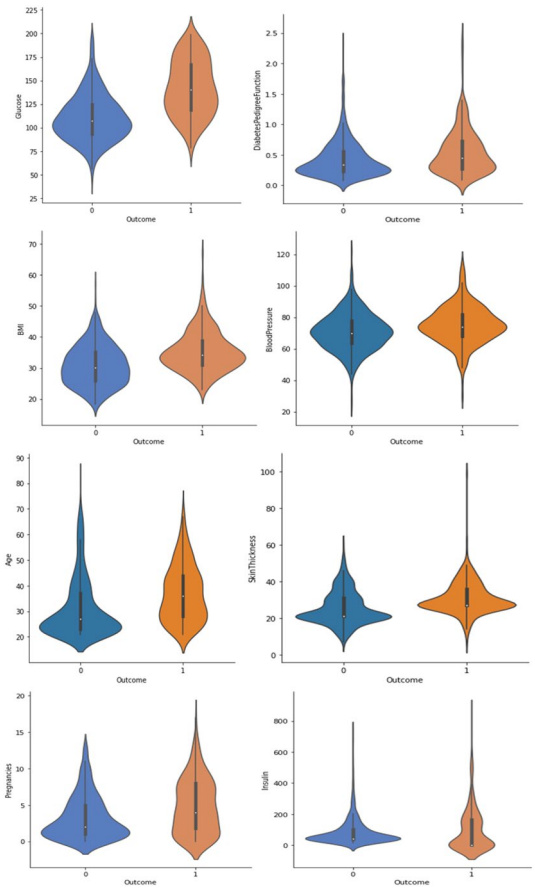

Fig. 3 Violin plot for the distribution of numeric data

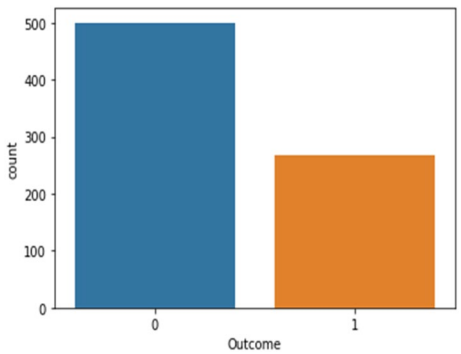

Fig. 4 Distribution of the dataset

The ratio of ordinary persons to diabetes patients is shown in Fig. 4. The dataset includes 500 observations of regular people (65.1%) and 258 observations of diabetic patients (34.9%). Our dataset contains an uneven class, as can be shown. The optimal classification model requires well-balanced datasets. To solve the data class imbalance problem, we employed the SMOTE approach. SMOTE is a data-driven algorithm that generates synthetic data points based on existing data points. SMOTE can be considered an improved version of oversampling or a unique data augmentation technique. A random sample is selected from the minority class by the SMOTE algorithm. The observations in this sample then allow us to determine the k closest neighbours. The vector that connects the current data point and the selected neighbour is then calculated using one of those neighbours. The vector is then multiplied by a random number between 0 and 1. The multiplied value is added to the original data point to create the synthetic data point. The data point would be slightly moved in the direction of its neighbour throughout this process.

Doing this ensures that the fake data point is not an identical replica of an already-existing data point and is also not too dissimilar from the known observations in your minority class.

In Fig. 5, we are looking for outliers in the dataset and using the IQR (interquartile range) approach to find them. After eliminating the outliers, the dataset was shrunk from (768, 9) to (639, 9). Figure 5 is a box plot representation for outlier detection. A box plot is a graph showing how the data values are spread out. There are five numbers of encapsulations: minimum, first quartile, median, third quartile, and maximum. It shows the outliers, their values, and the data skewness.

After that, Incremental Principal Component Analysis (IPCA) is used to scale down dimensionality and noise while preserving the majority of the data variance by removing the most significant principal components. By applying IPCA, we could remove the less predictable information while keeping the components that carry most of the predictive information. Depending on the input data's size, IPCA has considerably more memory and efficiency than PCA, which also allows for irregular variables.

The BLSTM model is acquired for diabetes diseases binary classification. In the BLSTM model the input data must be in the form of time series data. Hence, we reshaped the input data, including nine attributes. They are represented as Y_i in Eq. (1). The balanced records are reshaped in Eq. (2). It complies with the input requirement of the deep neural network framework.

$$\text{input} = \sum_{i=1}^{8} Y_i, \tag{1}$$

$$\text{Tensor input} = \text{reshape}\left(Y_i\right). \tag{2}$$

Long Short-Term Memory

The nature of the "LSTMs" is exactly like the recurrent neural network designed by Hochreiter and Schmidhuber in 1997 [37]. LSTM performs a variety of functions far more

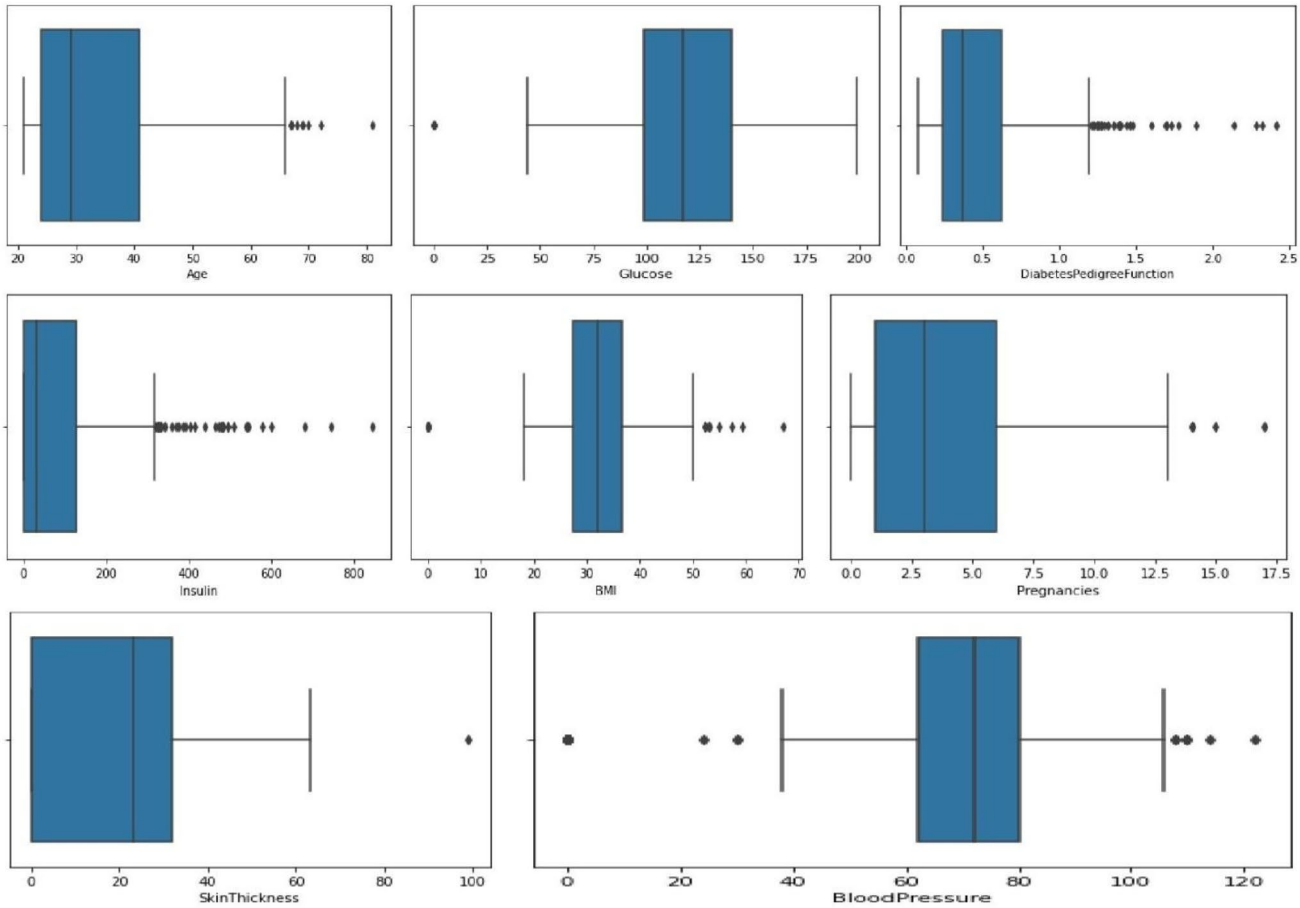

Fig. 5 Box plot for outlier detection

effectively than the standard version. RNN comes with vanishing gradient complexity. LSTM fulfils the promise of recurrent neural networks by overcoming technological challenges. LSTM has secured a memory unit for storing input/output data. As an alternative to a single-layer neural network, LSTM contains a memory cell and three interacting multiplicative units: forget gate, input gate, and output gate.

Figure 6 depicts the architecture of the LSTM cell [38] with input gate i_t, forget gate f_t, control gate c_t, and output gate o_t for a particular time step t.

The input gate determines what expletive knowledge will be stored in the cell state, which is stated as

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right). \tag{3}$$

The forget gate defines what prior knowledge from the cell state that is not significant from the previous time step should be remembered and is described as

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right). \tag{4}$$

The control gate oversees the rejuvenating cell state from Ct-1 to Ct, founded on Eqs. 5 and 6

$$\tilde{C}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right), \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * C_t. \tag{6}$$

The output gate is in authority for brought to pass the productivity in the current time step. This process can be lay it out as

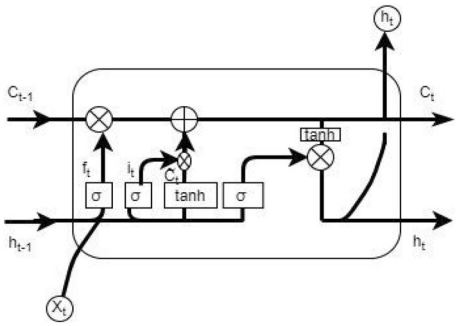

Fig. 6 The architecture of LSTM cell [38]

$$C_t = f_t * C_{t-1} + i_t * C_t, \tag{7}$$

$$h_t = O_t * \left((\tanh(C_t)\right). \tag{8}$$

In mathematical statement Eqs. 1–6, σ is the sigmoid activation function, which decides which values to let through 0 or 1. A value of 0 implies "let nothing through," whereas a value of one means "let everything through." The Ws are correlated with weight matrices. Tanh is assigns weightage to the values which are passed. Determine their level of relevance and restrict the values in the direction of through to the range of -1 to 1.

Bi-directional Long Short-Term Memory

Deep learning is a subset of machine learning and a type of neural learning based on artificial intelligence. Data in the digital age may be both unstructured and unlabeled. Deep learning handles it and builds a model [39]. BLSTM neural networks feature the same chain-like structure as LSTM neural networks, but the repeating module is recursive. We add an extra LSTM layer in Bidirectional long short-term memory, and the information flows in the reverse direction. BLSTM is a variant of the deep learning model employed for sequential processing.

The BLSTM is reckoned up by two LSTM nets defined as a bidirectional recurrent neural network. One processes the input in its original form, while the other processes the reversed input sequence. It improves the classification process, unlike the standard LSTM framework; there is a sequential input structure for BLSTM architecture to divide the neuron state into forward and backward. Videlicet, BLSTM gets hold of recurrent networks, forward and back. The output is calculated using forwards and backward RNNs on the BLSTM's hidden state. This permits the representation to be based on the past and the future for the given time range. A bidirectional RNN (BRNN) model is developed to alleviate numerous limitations of traditional RNNs (see Fig. 7).

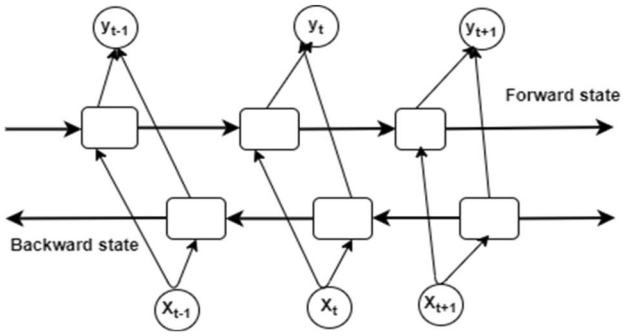

Fig. 7 Structure of the BLSTM [37]

Experimental Result

Hyperparameter Tuning

The BLSTM architecture accepts input vectors of 64 samples in a batch. The network is trained for more than 200 epochs using the Adam optimizer. It updates the weights and biases of the neural network. Further, the learning rates are considered as 0.002, and binary cross-entropy is used as a loss function. We have used the sigmoid activation function, which introduces non-linearity into the networks. Dropout layers and a tenfold cross-validation technique have been added to the proposed architecture to ensure no overfitting. Dropout refers to the practice of ignoring specific nodes in a layer at random during training in deep learning. A dropout is a regularization strategy that prevents overfitting by ensuring that no units depend on one another.

All of these parameters are chosen as the best parameters for neural network training. We used manual hyperparameter tuning, looping through different hyperparameter values and evaluating each combination.

Performance Evaluation Measures

The current study used the BLSTM model to predict diabetes. Knowledge gained from prior studies is critical to these accomplishments. Handling missing values and removing outliers are required for the intended results. This model's accuracy is improved using the IPCA approach. Up-sampling is also vital in establishing a balanced model. The confusion matrix tool is used to summarise the outcome for further investigation. The error matrix, meant to describe the classification algorithm's performance, analyses model performance and provides a holistic perspective of the model. The confusion matrix compares the ample category of the sampling and predicted variety of the selection with the help of their criterion, which are true positive (TP), true negative (TN), false positive (FP), and false negative (FN) in collaboration.where

- **True positive (TP)** denotes a circumstance in which a person has diabetes, and the model classifies their condition as diabetes, which falls under true positive.
- **True negative** (TN) denotes an observation in which a person is non-diabetic, and the model also categorises their case as non-diabetic.
- **False positives (FP)** refer to an individual who does not have diabetes, but their observation was classified as having diabetes by the model.

- **False negative (FN)** is a term used to describe situations in which a person has diabetes, yet the model classifies their observation as having no diabetes.

It is appropriate for therapeutical diagnosis; to utilise measures, accuracy, *f*1 score, sensitivity, and specificity are required. Accuracy is defined as the number of times the model correctly predicts output. Sensitivity manifests itself as some deterministic occurrences that are predicted to be positive. Sensitivity is also known as recall. The fraction of negative cases evaluated as negative determines Specificity. Recall's exact inverse is Specificity. Precision indicated that the number of correctly predicted difficulties was positive (see Table 3).

The equations and corresponding results are demonstrated here:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}),$$
$$\text{Accuracy} = (63 + 78) / (63 + 78 + 7 + 2) = 141/150 = 0.94,$$
$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}),$$
$$\text{Sensitivity} = (63 + 78) / (63 + 2) = 63/65 = 0.96,$$
$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}),$$
$$\text{Specificity} = 78/ (78 + 7) = 78/85 = 0.91.$$

Results and Analysis

In this study, the deep neural network BLSTM model is put into service to diagnose diabetes mellitus. Several criteria, such as accuracy, precision, recall, *F*1 score, and AUC, are employed to study the framework and assess the proposed approach's effectiveness. Several metrics, including accuracy, precision, recall, *F*1 score, and AUC, are used to evaluate the effectiveness of the suggested system.

Compared to earlier efforts, the proposed technique achieves 94% accuracy, 96% sensitivity, and 91% specificity, and its *F*1 score is 93%. Figure 8 depicts the graph for several performance metrics. This is productive for the diabetes classification issue. Using IPCA instead of PCA proved to be more expressive and decreased model convolution. In the PID dataset, SMOTE strategy performs better than under-sampling techniques. This thrash-out focuses on creating a very low-cost setup for diabetic illness detection because many people originate from underprivileged homes and do not have enough money to pay for testing in a lab.

The receiver operating characteristics curve (ROC) at the time of distinguishing false-positive and

Table 3 Comparison of previous research work on PID dataset

Author	Methodology	Result
Talha Mahboob Alamet al. [13]	ANN	Accuracy 75.7%
Suyash Srivastava et al. [40]	ANN	Accuracy 92%
Alessandro Massaroal et al. [41]	LSTM	Accuracy 75%
Muhammad Mazhar Bukhari et al. [42]	ABP-SCGNN	Accuracy 93%
Jayroop Ramesh et al. [43]	SVM	Accuracy 83.20%, Sensitivity 87.20% Specificity 79%
Umair Muneer Buttetal. [44]	LSTM	Accuracy 87%
Sourav Kumar Bhoiet al. [45]	LR	Accuracy—0.768 *F*1—0.760 Precision—0.763 Recall—0.768 AUC—0.825
Jobeda Jamal Khanam et al. [30]	Neural network	Accuracy 88.60%
Roshi Saxena et al. [3]	Random Forest	Accuracy 79.8%
Suja Alex et al. [18]	DCNN	Accuracy 86.29% Precision 81% Recall 84% AUC 91%
Hanaa Salem et al. [46]	TFKNN	Accuracy—90.63% Specificity—85% Precision—93% AUC—94.13%
Our proposed work, 2021	BLSTM	Accuracy 94% Sensitivity 96% Specificity 91% *F*1-score 93% Precision 90%

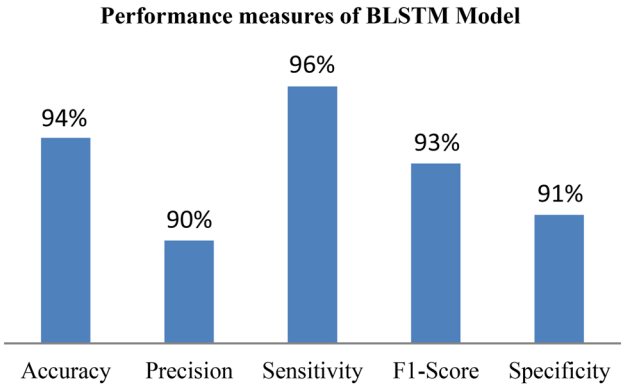

Fig. 8 Statistical comparison of the model

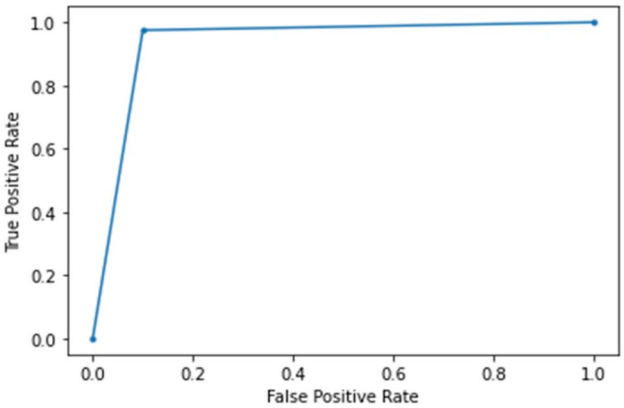

Fig. 9 ROC curve analysis

genuine-positive rates is shown in Fig. 9. It displays a relationship between sensitivity and specificity. AUC scores ranging from 0.9 to 1 are considered magnificent. The AUC value of the presented BLSTM model is 0.93, as seen in the graph in Fig. 9. The higher the AUC value, the better the model. The proposed research intends to create a model that accurately predicts diabetes mellitus and produces desirable outcomes.

Conclusion and Future Work

Diabetes is a chronic illness that can affect various parts of the body. The proposed BLSTM framework is designed for the early detection of DM. The model outperforms earlier research work and provides the maximum sensitivity rate and accuracy, respectively, 96% and 94%. The PID dataset, often used in diabetes prediction, has several outliers, missing values, and class imbalance regarding issues. To reduce these problems, the SMOTE-BLSTM model is used. Researchers could attempt to deploy different models on a greater dimension and a larger sample dataset. The pre-process data must be used for newly found approaches and ensemble learning to enhance the model's performance.

Data availability Data available on request from the authors.

Declarations

Conflict of interest There is no conflict of interest.

References

1. Dremin V, Marcinkevics Z, Zherebtsov E, Popov A, Grabovskis A, Kronberga H, Geldnere K, Doronin A, Meglinski I, Bykov A. Skin complications of diabetes mellitus revealed by polarized hyperspectral imaging and machine learning. IEEE Trans Med Imaging. 2021;40(4):1207–16. https://doi.org/10.1109/TMI.2021.3049591.
2. Krishnamoorthi R, Joshi S, Almarzouki HZ, Shukla PK, Rizwan A, Kalpana C, Tiwari B. A novel diabetes healthcare diabetes prediction framework using machine learning techniques. J Healthc Eng. 2022;1684017:10. https://doi.org/10.1155/2022/1684017.
3. Saxena R, Sharma SK, Gupta M, Sampada GC. A novel approach for feature selection and classification of diabetes mellitus: machine learning methods. Comput Intell Neurosci. 2022;3820360:11. https://doi.org/10.1155/2022/3820360.
4. Sivashankari R, Sudha M, Hasan MK, Saeed RA, Alsuhibany SA, Abdel-Khalek S. An empirical model to predict the diabetic positive using stacked ensemble approach. Front Public Health. 2022;9:792124. https://doi.org/10.3389/fpubh.2021.792124.
5. Wu H, et al. Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlock. 2018;10:100–7 (**ISSN 2352-9148**).
6. Syed AH, Khan T. Machine learning-based application for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi Arabia: a retrospective cross-sectional study. IEEE Access. 2020;8:199539–61. https://doi.org/10.1109/ACCESS.2020.3035026.
7. Choubey DK, Kumar P, Tripathi S, Kumar S. Performance evaluation of classification methods with PCA and PSO for diabetes. Netw Model Anal Health Inform Bioinform. 2020. https://doi.org/10.1007/s13721-019-0210-8.
8. Katarya R, Jain S. Comparison of Different Machine Learning Models for diabetes detection. In: IEEE International Conference on advances and developments in electrical and electronics engineering (ICADEE), 2020; p. 1–5.
9. Kulkarni BP. Analysis of classifiers for prediction of type II diabetes mellitus. In: Fourth International Conference on computing communication control and automation (ICCUBEA), Pune, India. 2018; p. 1–6.
10. Woldemichael FG, Menaria S. Prediction of diabetes using data mining techniques. In: 2nd International Conference on trends in electronics and informatics (ICOEI), Tirunelveli, 2018; p. 414–418.
11. Agarwal A, Saxena A. analysis of machine learning algorithms and obtaining highest accuracy for prediction of diabetes in women. In: 6th International Conference on computing for sustainable global development (INDIACom), New Delhi, India, 2019; p. 686–690.
12. Kowsher M et al. Prognosis and treatment prediction of type-2 diabetes using deep neural network and machine learning classifiers. In: 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2019; p. 1–6.

13. Alam TM et al. A model for early prediction of diabetes. Inform Med Unlock. 2019;16:1–6 (**ISSN 2352–9148**).
14. Sarki R, Ahmed K, Wang H, Zhang Y. Automatic detection of diabetic eye disease through deep learning using fundus images: a survey. IEEE Access. 2020;8:151133–49.
15. Sarwar MA, Kamal N, Hamid W, Shah MNA. Prediction of diabetes using machine learning algorithms in healthcare. In: 24th International Conference on Automation and Computing (ICAC), 2018; p. 1–6.
16. Islam MT et al. An empirical study on diabetes mellitus prediction for typical and non-typical cases using machine learning approaches. In: 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019; p. 1–7.
17. Islam MA, Alvi HN, Mamun KAA. DiaHealth: a smart app for complete diabetes lifestyle management. In: International Conference on Medical Engineering, Health Informatics and Technology, 2016; p. 1–6.
18. Alex SA, Nayahi JJV, Shine H, et al. Deep convolutional neural network for diabetes mellitus prediction. Neural Comput Appl. 2022;34:1319–27. https://doi.org/10.1007/s00521-021-06431-7.
19. Xu Z, Wang Z. A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and XGBoost ensemble classifier. In: Eleventh International Conference on Advanced Computational Intelligence (ICACI), Guilin, China, 2019; p. 278–283,.
20. Li X, Zhang J, Safara F. Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm. Neural Process Lett. 2021. https://doi.org/10.1007/s11063-021-10491-0.
21. Zhou H, Myrzashova R, Zheng R. Diabetes prediction model based on an enhanced deep neural network. JbWireless Com Netw. 2020;148:1–13.
22. Lai H, Huang H, Keshavjee K, et al. Predictive models for diabetes mellitus using machine learning techniques. BMC EndocrDisord. 2019;19:101. https://doi.org/10.1186/s12902-019-0436-6.
23. Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T. Current techniques for diabetes prediction: review and case study. Appl Sci. 2019;9(21):4604.
24. Fathi AE, Kearney RE, Palisaitis E, Boulet B, Haidar A. A model-based insulin dose optimization algorithm for people with type 1 diabetes on multiple daily injections therapy. IEEE Trans Biomed Eng. 2021;68(4):1208–19.
25. Roopa H, Asha T. A linear model based on principal component analysis for disease prediction. IEEE Access. 2019;7:105314–8.
26. Jayanthi N, Babu BV, Rao NS. Survey on clinical prediction models for diabetes prediction. J Big Data 4. 2017;26:1–15.
27. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Colagiuri S, Guariguata L, Motala AA, Ogurtsova K, Shaw JE, Bright D, Williams R. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045. In: Results from the International Diabetes Federation Diabetes Atlas. 2019;157:107843. https://doi.org/10.1016/j.diabres.2019.107843. (**Epub 2019 Sep 10. PMID: 31518657**).
28. Ahmad HF, Mukhtar H, Alaqail H, Seliaman M, Alhumam A. Investigating health-related features and their impact on the prediction of diabetes using machine learning. Appl Sci. 2021;11:1173.
29. Ramraj S, Sunil NUR, Banerjee S. Experimenting XGBoost algorithm for prediction and classification of different datasets. Int J Control@eory Appl. 2016;9:651–62.
30. Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 2021;7(4):432–9. https://doi.org/10.1016/j.icte.2021.02.004. (**ISSN 2405-9595**).
31. García-Ordás MT, Benavides C, Benítez-Andrades JA, Alaiz-Moretón H, García-Rodríguez I. Diabetes detection using deep learning techniques with oversampling and feature augmentation. Comput Methods Programs Biomed. 2021;202:105968. https://doi.org/10.1016/j.cmpb.2021.105968. (**ISSN 0169-2607**).
32. Saxena R, Sharma SK, Gupta M, Sampada GC. A Comprehensive review of various diabetic prediction models: a literature survey. J Healthc Eng. 2020;8100697:15. https://doi.org/10.1155/2022/8100697.
33. Battineni G, Sagaro GG, Nalini C, Amenta F, Tayebati SK. Comparative machine-learning approach: a follow-up study on type 2 diabetes predictions by cross-validation methods. Machines. 2019;7(4):74. https://doi.org/10.3390/machines7040074.
34. Saha PK, Patwary NS, Ahmed I. A widespread study of diabetes prediction using several machine learning techniques. In: 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2019; p. 1–5.
35. Gupta SC, Goel N. Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method. In: Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020; p. 980–986.
36. Huang L, Lu C. Intelligent diagnosis of diabetes based on information gain and deep neural network. In: 5th IEEE International Conference on CloudComputing and Intelligence Systems (CCIS), Nanjing, China, 2018; p. 493–496.
37. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
38. Olah C. Understanding LSTM Networks. 2015. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.
39. Özal Y. A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. Comput Biol Med. 2018;1(96):189–202.
40. Srivastava S, Sharma L, Sharma V, Kumar A, Darbari H. Prediction of diabetes using artificial neural network approach. Eng Vib Commun Inform Process. 2018;478:679–87.
41. Massaro A, Maritati V, Giannone D, Convertini D, Galiano A. LSTM DSS automatism and dataset optimization for diabetes prediction. Appl Sci. 2019;9(17):3532.
42. Bukhari MM, Alkhamees BF, Hussain S, Gumaei A, Assiri A, Ullah SS. An improved artificial neural network model for effective diabetes prediction. Complexity. 2021;5525271:10.
43. Ramesh J, Aburukba R, Sagahyroon A. A remote healthcare monitoring framework for diabetes prediction using machine learning. HealthcTechnol Lett. 2021;8(3):45–57.
44. Butt UM, Letchmunan S, Ali M, Hassan F, Baqir A, Sherazi HR. Machine learning based diabetes classification and prediction for healthcare applications. J Healthc Eng. 2021;9930985:17. https://doi.org/10.1155/2021/9930985.
45. Bhoi SK, Panda SK, Jena KK, Anshuman P, Sahoo KS, Sama NU, Pradhan SS, Sahoo RR. Prediction of diabetes in females of PIMA Indian heritage: a complete supervised learning approach. Turk J Comput Math Educ. 2021;12:3074–84. https://doi.org/10.17762/turcomat.v12i10.4958.
46. Salem H, Shams MY, Elzeki OM, Abd Elfattah M, Al-Amri FJ, Elnazer S. Fine-tuning fuzzy KNN classifier based on uncertainty membership for the medical diagnosis of diabetes. Appl Sci. 2022;12(3):950. https://doi.org/10.3390/app12030950.