[1] Smita Panigrahy
[2] Sachikanta Dash
[3] Sasmita Padhy*

SMOTE-based Deep LSTM System with GridSearchCV Optimization for Intelligent Diabetes Diagnosis

Journal of Electrical Systems

Abstract: - Diabetes is a metabolic illness initiated by either inadequate insulin creation by the pancreas or the body's reduced responsiveness to insulin. It is characterised by consistently high levels of blood sugar and symptoms such as frequent urination, thirst, and increased appetite. Untreated diabetes can result in significant problems that impact crucial organs, presenting potentially fatal dangers. In order to address the need for accurate diagnosis, researchers have utilised artificial intelligence to develop the G-LSTM system. This system employs a novel method that combines SMOTE-based deep LSTM and GridSearchCV optimization to classify diabetes. This technique effectively tackles the issue of class imbalance in diabetes datasets, demonstrating an exceptional level of prediction accuracy. When tested on the PIMA dataset, G-LSTM demonstrated exceptional performance with an accuracy of 97.12%. Additionally, it produced high precision, recall, F1-score, AUC, and MCC values of 97.12%, 0.963, 0.954, 0.887, 0.989, and 0.882, respectively. The results highlight the higher performance of the G-LSTM method compared to other techniques, suggesting its use for clinical investigation of diabetes patients. This innovative intelligent diagnostic framework not only demonstrates the potential of artificial intelligence in healthcare, but also highlights its crucial role in enhancing the precision and effectiveness of diabetes diagnosis and treatment.

Keywords: SMOTE, G-LSTM, PIMA, GridsearchCV.

I. INTRODUCTION

Diabetes is a persistent medical condition that arises when there is insufficient processing of blood sugar. Over a period of time, it might have an impact on numerous organs, leading to persistent harm and dysfunction. Diabetes can lead to severe consequences, including neuropathy, myocardial infarction, renal failure, and cerebrovascular accident. Hyperglycemia is a medical disorder characterised by abnormally high amounts of glucose in the body. This study investigates the identification of diabetes, with a particular focus on its significant influence on the cardiovascular system and other essential human organs. The study examines diabetes patterns using the Pima Indian Diabetes Dataset (PIDD) from the UCI Repository. It reveals a significant rise in the prevalence of diabetes, with the number of affected individuals increasing from 108 million in 1980 to 422 million in 2014. This emphasises the importance of developing efficient detection techniques for this increasingly serious health issue and emphasises the requirement for sophisticated diagnostic instruments to tackle the rising difficulties presented by diabetes. It is expected that by 2045, the number of people with diabetes would reach an estimated 700 million [5]. Developing a sophisticated diagnostic paradigm for diabetes is essential, considering its substantial influence. There are three main forms in which the disease appears: type 1 diabetes, type 2 diabetes, and gestational diabetes [6]. Type 2 diabetes, which affects more than 95% of individuals with diabetes, arises from the body's diminished capacity to utilise insulin effectively, usually due to excess weight and insufficient physical exercise. Conversely, type 1 diabetes arises due to inadequate production of insulin, necessitating the administration of insulin injections, and its aetiology remains unidentified. Type 1 diabetes is characterised by specific symptoms such as polyuria (excessive urination), thirst, hungriness, weight loss, and vision loss. Machine learning (ML) has gained significant traction in the field of medicine, namely for its application in intelligent disease diagnosis. As a result, ML approaches have been extensively utilised for the intelligent detection of diabetes [7]. ML models can assist in early prediction, diagnosis, classification, and treatment planning by analysing and extracting information from data obtained from diabetic patients. Researchers are investigating the use of ML technology to enhance the diagnosis of diabetes, with the aim of enhancing its management and treatment. Nevertheless, despite recent progress, there are still several obstacles that need to be addressed. Many disease diagnosis domains face obstacles in data gathering

[1] Computer Science and Engineering, GIET University, Gunupur, Odisha, India, smita.panigrahy@giet.edu

[2] Computer Science and Engineering, GIET University, Gunupur, Odisha, India, sachikanta@giet.edu

[3] School of Computing Science and Engineering, VIT Bhopal University, Bhopal, Madhya Pradesh, India, pinky.sasmita@gmail.com

* Corresponding Author: Sasmita Padhy

and processing, such as dealing with little, imbalanced, or low-quality data. The PIMA dataset poses several issues of intricacy, such as class imbalance, a substantial amount of missing values, and bad data quality. Prior research employing rudimentary ML methods has produced below-average model performance and disappointing outcomes when applied to the PIMA dataset. Similarly, academics' efforts to utilise intricate deep learning models have not yielded meaningful results in tackling these difficulties.

Class imbalance refers to a scenario where the distribution of instances across different classes is unequal. In a dataset, classes with the most instances are majority classes, while those with fewer instances are minority classes. The class-imbalance ratio (CIR) quantifies dataset imbalance by calculating the ratio of minority instances to majority instances, offering a straightforward measure for assessing class distribution. The presence of a disproportionate distribution of classes adversely affects the efficacy of the ML system in the dataset. This issue is referred to as overfitting. To address the issue, it is necessary to transform the imbalanced dataset into a balanced dataset prior to commencing the training process. Data augmentation is a technique used to supply more data in order to address the issue of imbalanced data prior to the training phase In contrast, the Synthetic Minority Oversampling Technique (SMOTE) entails creating synthetic samples to tackle the issue of class imbalance. More precisely, it generates artificial samples that lie between each minority instance and its adjacent examples, so improving the representation of minority classes. Deep learning classifiers, particularly Long Short-Term Memory (LSTM), outperform ML classifiers and are widely used in the field of recurrent neural networks (RNN). LSTMs, which may be trained with numerous layers, enhance model accuracy, making them especially helpful in situations that demand intricate temporal relationships and effective memory retention during training.

In addition, the study presents the G-LSTM model, which has demonstrated encouraging outcomes in attaining a high level of accuracy in the diagnosis of diabetes. The G-LSTM model integrates generative adversarial networks (GANs) into its training process to generate synthetic samples continuously. This augmentation enhances the model's ability to identify intricate patterns crucial for precise diabetes classification. By incorporating GANs, the model's accuracy significantly improves, thereby enhancing its reliability for diagnosing diabetes through ML. The G-LSTM model, presented in this paper, demonstrates outstanding performance, notably notable in its excellent accuracy on the PIMA dataset. The integration of generative adversarial networks enhances the model's resilience and effectiveness, rendering it a powerful instrument for precise and quick diabetes categorization. This demonstrates the supremacy and efficacy of our proposed model in precisely forecasting and categorising instances of diabetes in various datasets.

The paper is organised in the following manner: The Related Work section offers a comprehensive summary of the existing research on diabetes classification and emphasises the challenges faced by scientists in this domain. The Materials & Methods section provides a detailed account of the dataset, the methodologies used for data preprocessing, and the diabetes classification model known as G-LSTM. The Results section showcases the outcomes of the framework, encompassing comparisons with alternative classifiers, diverse classification tasks, and results obtained from various datasets.

II. Literature Review

The progress of computer technology in recent years has resulted in the development of ML. A rising number of researchers employ ML to improve diabetes diagnosis and treatment, utilizing conventional classifiers for forecasting and categorizing the disease.

Researchers [8] employed the K-nearest neighbour approach and attained an accuracy of 79.8%, while Researchers [9] suggested logistic regression for the purpose of data classification. A thorough investigation [10] was conducted to examine the performance Random forest, multilayer perceptron, and logistic regression. Among these classifiers, the multilayer perceptron shown exceptional effectiveness, obtaining an accuracy rate of 86.06%. In another study conducted by Zo Authors [11], decision trees, random forests, and neural networks were utilised to predict diabetes. The results emphasised that the random forest algorithm exhibited a higher accuracy of 80.84% when all features were taken into account.

The utilisation of deep learning has witnessed a steady rise in recent years owing to its exceptional ability to effectively process intricate datasets. In their study [12], the authors employed a variational self-encoder and a sparse self-encoder to enhance data augmentation and feature augmentation, respectively. By jointly training a convolutional neural network with a sparse self-encoder, they achieved an impressive accuracy of 92.31%. Additionally, in another study [13], ensemble classifiers like AdaBoost and Gradient Boost were utilized, while in a separate study [14], authors introduced an enhanced artificial neural network (ANN) model without preprocessing

the data in advance. The authors [15] devised an innovative classification model utilising Conv-LSTM, with a record-breaking accuracy of 91.38%. In addition, they implemented a deep extreme learning machine (DELM) prediction model, which exhibited exceptional dependability, achieving an accuracy rate of 92.8%.

In [17], a hybrid model could be employed for diabetics prediction. The initial step was data cleansing to ensure consistency, followed by RF and XGB classifiers for selection of a subset of features. Subsequently, erroneous data were eliminated by the utilization of K-means clustering.

According to [18], the PIDD dataset was used to train seven distinct ML models, each with its own set of features. Two features were excluded in the feature selection process of this technique. SVM and LR showed strong predictive performance for diabetes; a complex neural network was trained with multiple hidden layers and epochs. The authors demonstrate that a neural network with two hidden layers has superior performance in comparison to previous methodologies.

A review [19] indicates that ML is robust enough to aid doctors in predicting the likelihood of future type 2 diabetes development. Machine learning (ML) was employed in a study [20] to conduct a comprehensive evaluation of predicting methods for diabetes. The Prediction Model Risk of Bias Assessment Tool (PROBAST) evaluated bias in ML models, whereas Meta-DiSc measured variability in a systematic review, demonstrating the greater effectiveness of ML compared to traditional methods..

The ensemble approaches utilized various supplemental ML techniques, such as SVM and Convolutional Neural Networks (CNN), to evaluate improvements in performance. However, the primary algorithm used in reference [21] was Logistic Regression (LR). The experiment utilized two distinct feature selection methods in conjunction with two datasets. The first dataset was chosen from the Pima Indians dataset, which comprises nine unique features. The subsequent dataset employed was the Vanderbilt dataset, which consisted of 16 features. The study's findings demonstrated that the LR algorithm ranks among the most efficacious methods for developing predictive models.

[22] employed ML techniques, specifically ten-fold cross validation, to analyze individuals with a history of non-diabetes and heart issues. The authors enhanced the accuracy of clinical prediction for early detection of diabetes type 2 mellitus by employing advanced ML forecasting algorithms like Glmnet, RF, XGBoost, and LightGBM. While it demonstrates efficacy with one dataset, it is unsuitable for another.

In their study, the researchers in [23] devised an innovative technique for detecting DD using the LS-SVM and GDA methodologies. A novel cascading learning system was implemented, utilizing the methodologies previously outlined. The constructed system comprised of two stages: firstly, GDA was utilized; secondly, LS-SVM was implemented to categorize the datasets connected to diabetes. Compared to previous findings obtained using alternative categorization techniques, the results demonstrated a favorable accuracy rate of 82.05% for classification.

The study conducted a performance comparison of three models and the accuracy measure was used for evaluation. The prediction accuracy value for LSTM was determined to be 60.6%, and the RNN successfully classified distinct forms of diabetes using all eight variables [24]. To ensure robust validation, the dataset was split into an 80:20 ratios, with 80% designated for training and the remaining 20% reserved for testing purposes. The accuracy of predicting type 1 diabetes was 78%, whereas the accuracy of predicting type 2 diabetes was 81%. Many academics support the use of deep learning in the field of medical data. They propose using a DBN to classify diabetes, using a small number of factors from a dataset of 50,000 diabetic cases. This strategy achieved a precision rate of 81.20%.

Another approach entails utilising the LM algorithm to train a Multilayer Neural Network (MLNN) for the purpose of predicting diabetes. This prediction is based on the Pima Indian Diabetes Dataset (PIDD). The MLNN's performance is then compared to that of a Probabilistic Neural Network (PNN) [25, 26] and prediction of glucose levels [27]. The strategy is a hybrid one, which involves the selection of optimal characteristics and then doing classification. The benefit of this method lies in its ability to prevent the use of unnecessary qualities. Additionally, it utilizes dimensionality reduction to address overfitting concerns, enhancing model robustness. The EAGA-MLP is a sophisticated adaptive-genetic algorithm multilayer perceptron that optimises attributes to create a dataset that is optimised for classification. The Long Short-Term Memory (LSTM) idea is offered as a method for predicting time-series data [29]. The LSTM, a distinct variant of recurrent neural networks, demonstrates immense value in the prediction of time-series data. LSTM, which consists of three layers (input, hidden, and output), is specifically designed to efficiently record and understand relationships in long sequences. This enables it to retain and recall important information necessary for making precise predictions in scenarios involving temporal data. The complex structure of LSTM enables it to effectively address the issues related to vanishing or inflating gradients, hence

improving its ability to capture long-term dependencies in sequential data. The hidden layer's memory cells are equipped with three gates that modify the cell's state. The vanishing gradient problem is absent in LSTM networks, unlike in RNNs. Accurate diabetes prediction relies on the network's ability to process outdated information [30]. The LSTM model is designed to handle input data in a three-dimensional manner, which includes information about samples, time steps, and features. This architecture facilitates the model's ability to effectively process and analyse sequential data, hence enabling it to accurately capture temporal dependencies and extract significant patterns from input sequences. The inclusion of a three-dimensional input representation is essential for ensuring the efficacy of LSTM models in processing time-series data. Each dimension of the input serves a unique purpose in capturing the temporal dynamics and complex relationships contained in the dataset. A batch may contain one or more samples. When working with natural-language processing, if we focus on analysing text on a sentence level, our sample size will consist of only one sentence. Prior work faces challenges like class-imbalance and dataset optimization.
The suggested method's features are outlined as listed below:

- Creating a deep-learning model to classify diabetic cases; • Enhancing the accuracy of the deep-learning model by implementing SMOTE to address the issue of imbalanced classes.
- Employed a GridSearchCV optimizer to enhance the prediction accuracy
- Compared with the existing models.

III. Resources and Approaches

3.1. Dataset

The study employed the publicly available PIMA Indian Diabetes Dataset from NIDDK, comprising data from 768 patients, 268 with diabetes. It includes eight physiological indicators used to forecast diabetes incidence: Pregnancies, Glucose, Skin Thickness, Blood Pressure, Insulin, Diabetes Pedigree Function, BMI, and Age. Table I presents a thorough analysis of each attribute, enabling a comprehensive comprehension of the dataset and its constituents. The diabetes dataset being analysed consists of 768 female patients, obtained from the UCI (University of California, Irvine) and the National Institute of Diabetes and Digestive and Kidney Diseases in November 2019. Out of the total number of participants, 500 do not have diabetes, but 268 have received a diagnosis for the ailment. The goal is to determine whether diabetes is present or not by examining specific diagnostic parameters in the dataset. The trial specifically targets individuals of Pima Indian ancestry, all of whom are at least 21 years old. Curiously, some patients display zero readings for crucial metrics. Significantly, there are 374 patients with a serum insulin level of zero, 27 with a body mass index of zero, 35 with a diastolic blood pressure of zero, 227 with a skinfold thickness of zero, and 5 with a glucose level of zero. These zero values are classified as null values and, in accordance with WHO criteria, function as crucial markers for forecasting diabetes. The goal variable in the dataset is dichotomous, representing the presence (1) or absence (0) of diabetes in a patient. By employing ML methods, this binary classification allows for the prediction of diabetes using particular criteria. The dataset's demographic attributes, specifically for patients diagnosed with diabetes, are displayed in Table I. This dataset has been crucial in studies focused on predicting diabetes, serving as a fundamental dataset for creating and assessing ML algorithms specifically designed for diabetes prediction. Scientists utilise the extensive data in this dataset to find key indicators that lead to precise predictions of diabetes. Due to its large sample size of female patients, this dataset is particularly valuable for comprehending and tackling the intricacies related to diabetes in this specific group. Scientists and professionals are still investigating and expanding upon the knowledge gained from this dataset, which is leading to progress in the prediction and treatment of diabetes.

Table I Description of Attributes of Diabetics Dataset

Sl No	Attribute	Description	SD Vs Mean
1	Preg.	No. of times Pregnant	3.36 / 3.84
2	Plas	Plasma Glucose(2h)	30.46 / 121.67
3	Pres	Blood Pressure(mm Hg)	12.10 / 72.38
4	Skin	Skin-fold Thichness(mm)	8.89 / 29.08
5	Insu	Serum Insulin in two hours(μU/mL)	89.10 / 141.76
6	BMI	Body Mass Index(Kg/M)	6.88 / 32.43
7	Pedi	Diabetics Pedigree Function	0.33 / 0.47

8	Age	Age(Years)	11.76 / 33.24
9	O/p Class	Yes or No class for Diabetics	

3.2. Preprocessing Procedure

Initially, the missing values are addressed. It is typical for healthcare data sets to include missing values. The Tukey technique is a widely used outlier identification approach that effectively deals with missing information in various applications, making it more favourable than other methods. The dataset is normalised in this study using the min-max transformation method, which is a commonly used methodology for standardising attributes within a dataset. The purpose of the min-max transformation is to standardise attribute values by transforming them from their original range to a given new range. The procedure entails identifying the minimum and maximum values for each attribute and subsequently scaling the values proportionally within the specified range, typically [0, 1]. The process of standardisation is especially advantageous when working with characteristics that have varying measurement units or ranges. It guarantees that all attributes have an equal impact on the analysis, regardless of their original scales. The process of normalising the dataset helps to establish a consistent structure for further studies, improving the clarity and effectiveness of ML models by reducing the influence of varying attribute scales. The rigorous preprocessing stage enhances the overall dependability and efficacy of the study's predictive modelling or analysis, promoting a more precise comprehension of the underlying patterns in the data. Equation (1) provides the formula for min-max normalisation.

$$x_0 = (x_{max} - x_{min}) \times (x_i - x_{min}) \,/\, (x_{max} - x_{\min}) + x_{\min} \quad (1)$$

Here x_i represents individual instances of input, x_0 denotes the input data, and $x_{\max}$ indicates Maximum and $x_{\min}$ indicate the minimum values for input x. This scaling preserves data correlation.

3.3. Balancing Class Distribution

Classifiers trained on imbalanced data sets are likely to have poor robustness and generalisation performance. This issue is a fundamental challenge in the field of ML. SMOTE is a highly successful method for dealing with imbalanced data. Since the columns of PIDD dataset consist of statistical values, SMOTE is an appropriate technique to employ. The fundamental concept of SMOTE is to randomly insert fresh samples among smaller samples and neighbouring samples. The K-nearest neighbours are initially evaluated using the samples from the minority class. The interruption method utilized by SMOTE can be found in Equation (2).

$$S_i = X + rand(0,1) \times (y_i - X) \quad (2)$$

In the given context, X represents a data sample, rand (0, 1) represents a random number between 0 and 1, y_i represents the i^{th} nearest neighbours, and S_i represents the interpolated sample.

3.4. GridSearchCV optimization

The grid search methodology is one of the most commonly employed approaches for hyperparameter optimisation. Hyperparameter tuning employs a method to identify the optimal values that yield the highest performance in a specified model. This procedure involves training the model using all possible arrangement of properties specified by the user. It is a crucial step since the optimal parameters discovered have a significant impact on the overall performance of the model. Nevertheless, overfitting can arise during optimising. Applying the cross-validation (CV) method can mitigate the issue of overfitting. The CV technique involves training a model using a specific dataset and subsequently evaluating its performance using many datasets. In order to identify the optimal combination of learning, the GridSearchCV algorithm is employed. Subsequently, the algorithm identifies the parameter combinations that yield the maximum accuracy and selects them. Once the optimal parameter set has been chosen, the data estimating process commences. The 10-fold cross-validation method is used to completely evaluate the model's performance. It involves constructing ten distinct sets for both training and testing purposes. This procedure is iterated for every dataset, and the average of diabetes prognosis is computed. The study combines the grid search and cross-validation techniques to optimise hyperparameters through a set of experiments, ensuring the accuracy of diabetes prediction by robustly and reliably tuning the model parameters.

3.5. G-LSTM Prediction model

The LSTM network, which occurs repeatedly, handles the produced characteristics. The LSTM (Long Short-Term Memory) network consists of 64 cells featuring recurrent layers and three crucial gates: Input, Forget, and Output. Through these components, the network proficiently processes inputs to generate corresponding outputs. Furthermore, the cells possess memory units that retain all processed information. The memory cells are crucial for streamlining the computational process and generating the results. The memorial cell refreshes the processing inputs throughout each computation within a predetermined time interval. Figure 1 represnts the general architecture of classical LSTM model.

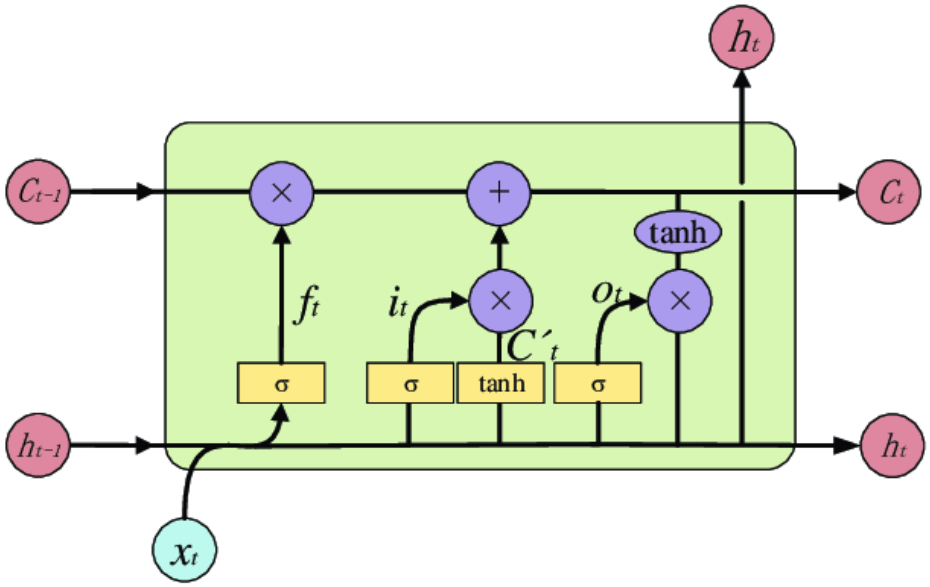

Fig. 1: Architecture of LSTM

During the calculation, the network consists of a single recurrent layer with 64 cells, as well as a dropout layer to remove extraneous input. The dropout layer mitigates overfitting issues, hence enhancing generalisation and prediction accuracy. Based on the explanation, the computation for each gate output is demonstrated in the equations. The formula 3-8 outlines the LSTM computations used to analyse gates and determine the accuracy of diagnosing diabetes in people:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f\right) \quad (4)$$

$$g_t = \sigma\left(W_{xg}x_t + W_{hg}h_{t-1} + W_{cg}C_{t-1} + b_g\right) \quad (5)$$

$$O_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}C_{t-1} + b_i) \quad (6)$$

$$c_t = (i_t g_t + f_t c_{t-1}) \quad (7)$$

$$h_t = (o_t \tanh(c_t) \quad (8)$$

x_t stands for the current input data, h_{t-1} is the output result of the data before it, $C_{t-1}$is the cell activation vector of the data before it, for the sigmod function, O_t is the output gate result of the data at the current time, and h_t is the output result of the data at the current time.

The sigmoid activation function is utilised in the computational process to generate the final output as $sigm$= 1/ (1+e^{-a}). The calculated output is compared to the training dataset in order to assess the error. The network parameters are adjusted iteratively to minimise misclassification error rates when the estimated values differ from the real values. Using LSTM networks with recurrent layers and gates improves the model's performance by enabling it to catch and utilise temporal dependencies in the input data, resulting in more accurate predictions or classifications.

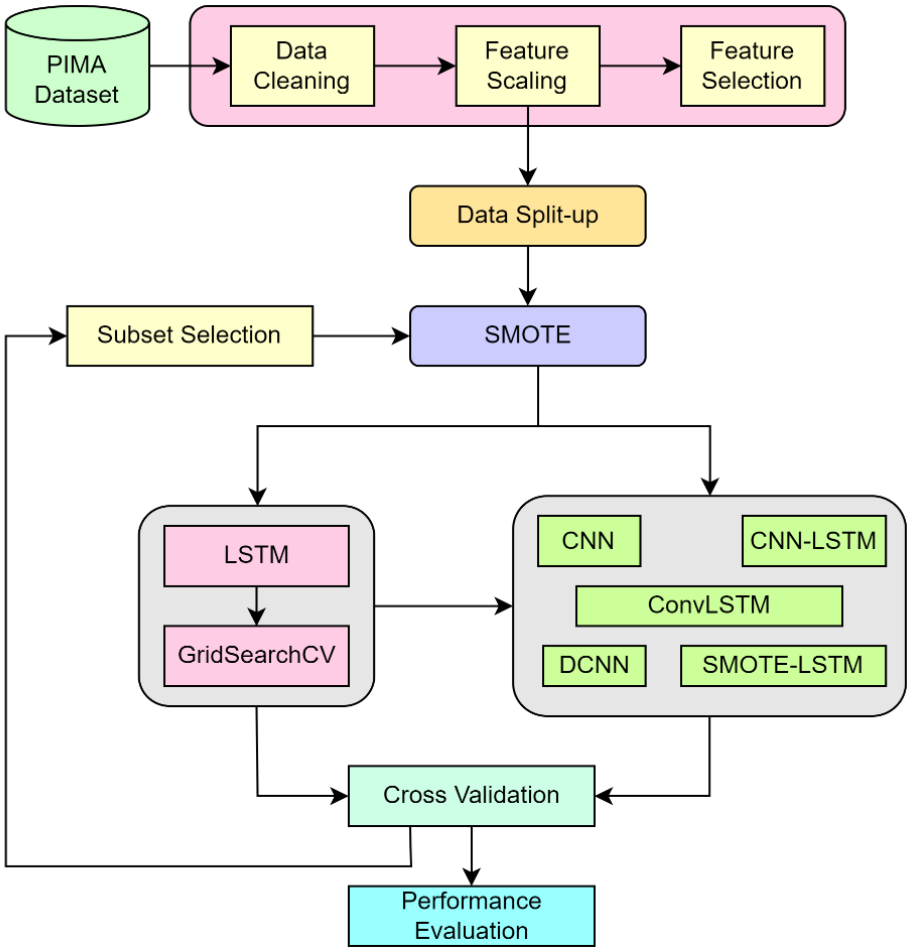

Fig. 2. Flow work of G-LSTM

The GridSearchCV optimisation method is used to select the most optimised parameters from the given search space. The optimisation approach aims to minimise the error rates in classification, hence improving the overall accuracy of predicting diabetic illness. Figure 2 shows the flow of proposed model to enhance prediction.

Algorithm:

Input: Training data consisting of X_train and Y_train
Test data consists of X_test and Y_test.
LSTM model structure
Grid of parameters for optimizing hyperparameters
Assessment metric

Result: G- LSTM model
Evaluation metrics for the top-performing model on the test dataset

Procedure:

- **Define a function for the Long Short-Term Memory (LSTM) model:**
  - Specify a function to generate the LSTM model.
  - Develop a function that accepts hyperparameters as arguments and outputs a constructed LSTM model.
  - Utilize a deep learning framework like Keras with a TensorFlow backend.
- **Create an instance of KerasClassifier:**
  - Instantiate a KerasClassifier object using create_lstm_model() as the build_fn.
  - Indicate the epochs, batch size, and verbosity level.
  - Parameter grid is a structured arrangement of parameters used for tuning and optimizing machine learning algorithms.
  - Create a parameter grid with hyperparameters for tuning, such as units and optimizer.
  - Enumerate the potential values for each hyperparameter.
- **Define Scoring Function:**

- Create a scoring function that is based on the chosen evaluation metric, such as accuracy.
- Transform the evaluation metric into a scorer object that is suitable for use with scikit-learn.

- **Create an instance of GridSearchCV:**
  - Instantiate a GridSearchCV object.
  - Share the KerasClassifier, parameter grid, scoring function, and cross-validation settings, such as the number of folds.

- **Implement GridSearchCV to find the best model parameters.**
  - Train the GridSearchCV object using the training data (X_train, Y_train).
  - This will conduct hyperparameter optimization through k-fold cross-validation.

- **Obtain optimal parameters and performance score:**
  - Access the optimal parameters identified by GridSearchCV using the best_params_ property.
  - Retrieve the highest score attained (best_score_ attribute).

- **Assess Optimal Model:**
  - Access the optimal model (best_estimator_ attribute).
  - Assess the top-performing model using the test data (X_test, Y_test) to calculate assessment measures such as accuracy.

- **Displaying Output:**

Provide the top-performing LSTM model and its assessment metrics on the test dataset.

IV. Experimentation, Results and Analysis

The experiments were performed using a Jupyter Notebook (version 6.3.0) running Python (version 3.8.8) on the Windows 10 operating system. The study's simulation is performed on a personal computer including an Intel Core i7 8750 CPU processor and 16 GB of RAM. The simulation process comprises data pre-processing, data splitting, machine learning (ML) modelling, assessment, and charting. These jobs are made easier by utilising multiple libraries such as sklearn, PyOD, NumPy, pandas, SciPy, matplotlib, scikit-plot, and seaborn. Every library plays a unique role in the simulation, guaranteeing thorough and effective data analysis.

Our model was evaluated using a variety of factors, including precision, sensitivity, specificity, and others. The computations for the TN, TP, FN, and FP were used to calculate these assessment metrics. The terms TN and TP refer to the quantity of correctly classified negative and positive cases, accordingly. Additionally referred to as FN and FP, accordingly, is the quantity of wrongly identified positive and negative instances.

To ascertain the specific correctness of every single technique, the following metrics were taken and computed with the following formulas:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (10)$$

$$\text{MCC} = \frac{(TP*TN)-(FP+FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (11)$$

$$\text{F-Score} = \frac{2*(Precision*Sensitivity)}{Precision\ +Sensitivity} \quad (12)$$

$$\text{Accuracy} = \frac{TN+TP}{TP+TN+FN+FP} \quad (13)$$

4.1. Result analysis

The paper presents a bipartite model designed to enhance the precision of diabetes prediction. The model includes two main elements: the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance problems by creating artificial samples of the minority class, and a Gated Long Short-Term Memory (G-LSTM) classifier, which accurately captures temporal relationships in the data for accurate classification. The model improves its capacity to forecast diabetes accurately by integrating these strategies. Comparative analysis involved popular models and preceding model implementation, SMOTE was applied, and performance metrics such as precision, recall, accuracy, and Area Under the Curve (AUC) were assessed. Results, presented in Table II, highlighted the superiority of the G-LSTM model over others in diabetes prediction accuracy. This affirmed the

effectiveness of the proposed two-part model, emphasizing the synergy between SMOTE for class imbalance correction and the deep LSTM classifier for improved predictive capabilities in diabetes diagnosis.

Table II presents an extensive analysis of multiple deep-learning models used to predict diabetes, using a range of performance criteria. The evaluated models consist of CNN, CNN-LSTM , ConvLSTM, DCNN, SMOTE-LSTM, and the suggested G-LSTM. The precision values represent the ratio of correct positive predictions to all positive predictions. The G-LSTM model achieves a substantially high precision value of 0.963. The recall metric quantifies the capability to correctly identify all pertinent occurrences, demonstrating the robust performance of the G-LSTM model with a score of 0.954. The G-LSTM achieves a high accuracy and F-score of 0.9712 and 0.8876, respectively, indicating a high level of overall correctness. The Area Under the Curve (AUC), which demonstrates the model's capacity to distinguish between different classes, strongly supports the G-LSTM model with an impressive score of 0.989. The results highlight the superiority of the proposed model, specifically emphasising its precision, recall, accuracy, and AUC in contrast to other deep-learning approaches. The proposed method gives good accuracy with F1-score of 0.8876 and MCC of 0.8823. Figure 3 shows the graphical representation of the comparison of different models with proposed model.

Table II. Performance comparison of pre-trained models with proposed model

Model	Precision	Recall	Aaccuracy	F1-score	AUC	MCC
CNN	0.720	0.752	0.770	0.7829	0.805	0.6516
CNN-LSTM	0.795	0.810	0.825	0.7932	0.842	0.6432
ConvLSTM	0.815	0.828	0.840	0.7923	0.853	0.6954
DCNN	0.830	0.855	0.870	0.8513	0.905	0.6532
SMOTE-LSTM	0.943	0.934	0.968	0.8744	0.979	0.8476
G-LSTM	0.963	0.954	0.971	0.8876	0.989	0.8823

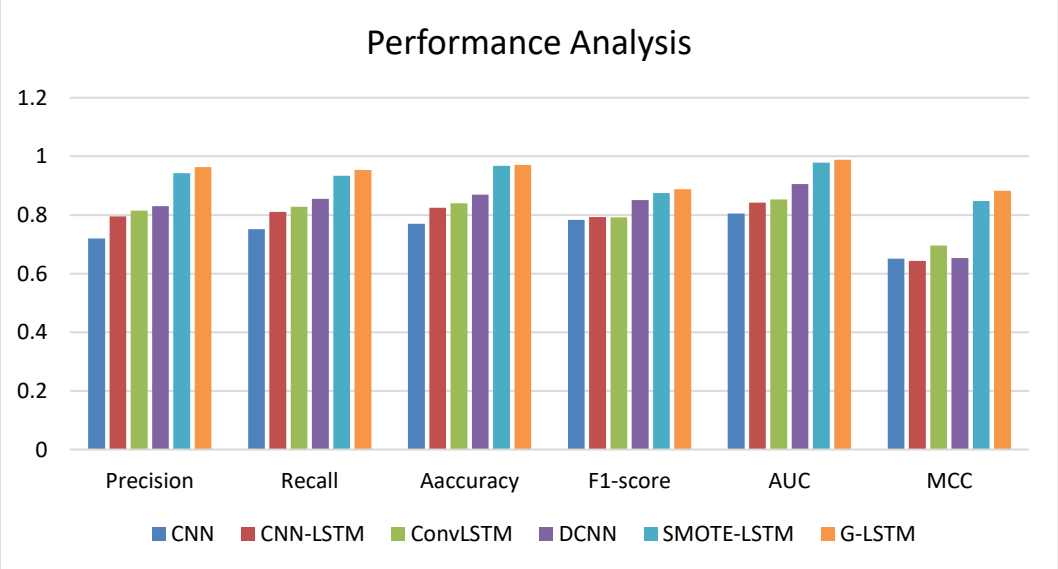

Figure 3. Performance analysis of different models

Table III and figure 4 gives the performance of various existing state-of-the-art techniques for predicting diabetes, with emphasis on the G-LSTM, which is shown to be the most effective. The study achieved diabetes prediction with a remarkable accuracy of 97.12% by utilising the G-LSTM model and nine variables from the Pima Indian Diabetes Dataset (PIDD). The results emphasise the effectiveness of the suggested approach in obtaining extremely precise and dependable predictions for diabetes, as demonstrated by its exceptional performance across all evaluation measures.

Table III. Performance of various existing state-of-the-art techniques

Author , Year	Models	Accuracy
Nadesh et al. , 2020 [10]	Deep Neural Network (DNN)	96.26
Wang et al., 2020 [13]	PSO+ LSTM	95.05
Ezzat et al., 2021[15]	Gravitational Search Optimised DeepLearning (GSODL)	95.03
AlZubi et al., 2020 [22]	Tubu Optimised Sequence Module Net (TOSMN)	96.52
Yin et al., 2021 [23]	DiabDeep	95.32
Yang etal., 2023 [27]	Blood Glucose through Temporal-multi-head Attention Mechanism (BGTAM)	96.35
Langarica et al., 2023 [5]	Input and State Recurrent Kalman Network (ISRKN)	95.56
Koutny et al., 2022 [29]	Meta-Differential Evolution	95.81
Proposed Model	(GridSearchCV + LSTM)	97.12

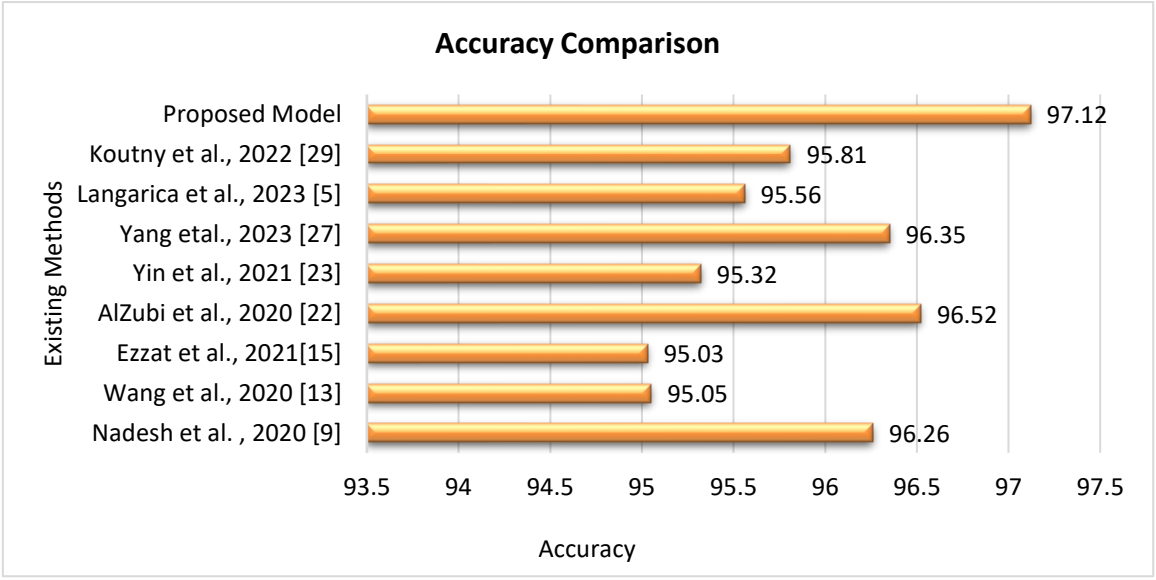

Figure 4. Accuracy comparison of proposed model with state-of-the-art model

V. DISCUSSION

To address the worldwide health risk presented by diabetes, scientists are progressively ML and DL technologies to automate the process of diagnosing the disease. This work aims to apply the G-LSTM algorithm to a dataset for the purpose of detecting diabetes. The train-test split is performed using a 10-fold Stratified Cross-Validation (SCV) technique. The G-LSTM algorithm's hyperparameters are optimised by a grid search employing a 10-fold CV technique. The Python programming language and accompanying libraries are used to do several tasks, such as data preparation, preprocessing, creating the G-LSTM algorithm, and conducting statistical analyses. The main goal is to assess and contrast the efficacy of four separate deep learning algorithms in forecasting diabetes. The study concludes with the presentation of findings, which demonstrate the evaluation criteria for current DL algorithms in Table 3 and Figure 4. The G-LSTM algorithm exhibits exceptional results, attaining a 97.12% accuracy rate. The findings support the ongoing attempts to improve automated diagnosis systems for diabetes by utilising advanced deep learning techniques. When tested on the PIMA dataset, G-LSTM demonstrated exceptional performance with an accuracy of 97.12%. Additionally, it produced high precision, recall, F1-score, AUC, and MCC values of 97.12%, 0.963, 0.954, 0.887, 0.989, and 0.882, respectively.

VI. Conclusion

The proposed methodology utilised the PIMA Dataset to accurately forecast the occurrence of type-2 diabetes specifically in females. This study utilised deep-learning principles to develop a predictive model for type-2 diabetes in the Pima Indian Diabetes Dataset, achieving accurate findings. The article utilised an SMOTE implemented deep LSTM approach with GridSearchCV optimisation to efficiently address class-imbalance handling and prediction, effectively capturing the entire process's features. The G-LSTM model was evaluated against other established deep learning models. Upon conducting comparisons, it was found that the G-LSTM model exhibited outstanding performance, with an accuracy rate of 97.12%. In addition, it achieved good levels of precision, F1-score, recall, AUC, and MCC, with values of 97.12%, 0.963, 0.954, 0.887, 0.989, and 0.882, respectively. By employing G-LSTM, we can promptly identify concealed hazards and implement proactive measures to enhance the patient's well-being and mitigate the risk of diabetes. The G-LSTM model demonstrated high accuracy and will be further expanded to forecast various diseases using different medical datasets.

Funding Statement:

There is no external funding received for doing this research work.

Conflicts of Interest:

The authors declare that they have no financial or other conflicts of interest.

Acknowledgment:

The authors express their sincere gratitude for the combined effort and unwavering dedication of all those who contributed to this scientific project. Every author made significant contributions to the conceptualization, design, execution, and interpretation of the findings of this study. Their combined knowledge, perspectives, and dedication greatly enhanced the quality and comprehensiveness of the research reported in this work.

References

[1] American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care.* 2011;37 (Suppl_1):62-S69.

[2] Priya G, Kalra S, Dasgupta A, Grewal E, Diabetes insipidus: a pragmatic approach to management. *Cureus*. 2021;13(1): e12498- e12498.

[3] Prabhakar PK, Pathophysiology of secondary complications of diabetes mellitus. *Pathophysiology*. 2016;9(1):32-36.

[4] Sun H, Saeedi P, Karuranga S, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract.* 2022;183:109119.

[5] S. Langarica, M. Rodriguez-Fernandez, F. J. Doyle III and F. Núñez, "A Probabilistic Approach to Blood Glucose Prediction in Type 1 Diabetes Under Meal Uncertainties," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 10, pp. 5054-5065, Oct. 2023, doi: 10.1109/JBHI.2023.3309302.

[6] Sönmez A, Ozdoğan O, Arıcı M, et al. Diyabette kardiyovasküler ve renal komplikasyonların önlenmesi, tanısı ve tedavisi için Endokrinoloji Kardiyoloji Nefroloji (ENKARNE) Uzlaşı Raporu. *Turk J Endocrinol Metab.* 2021;25(4):392-411.

[7] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* 2022;28(1):31-38.

[8] Ghaffar Nia N, Kaplanoglu E, Nasab A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence.* 2023;3(1):5. doi:10.1007/s44163- 023-00049-5

[9] Nadesh, R. K., & Arivuselvan, K. (2020). Type 2: Diabetes mellitus prediction using deep neural networks classifier. *International Journal of Cognitive Computing in Engineering*, *1*, 55-61. https://doi.org/10.1016/j.ijcce.2020.10.002

[10] Ali YA, Awwad EM, Al-Razgan M, Maarouf A, Hyperparameter search for machine learning algorithms for optimizing the computational complexity. *Processes.* 2023;11(2):349.

[11] Birjais R, Mourya AK, Chauhan R, Kaur H, Prediction and diagnosis of future diabetes risk: A machine learning approach. *SN Appl Sci.* 2019;9(1):1-8.

[12] Tigga, NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. *Procedia Comput Sci.* 2020;167: 706-716.

[13] Wang, W., Tong, M., & Yu, M. (2020). Blood glucose prediction with VMD and LSTM optimized by improved particle swarm optimization. *IEEE Access*, *8*, 217908-217916. https://doi.org/10.1109/access.2020.3041355

[14] Singh, N, Singh P. Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybern Biomed Eng.* 2020;40(1):1-22.

[15] Ezzat, D., Hassanien, A. E., & Ella, H. A. (2021). An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization. *Applied Soft Computing*, *98*, 106742. https://doi.org/10.1016/j.asoc.2020.106742

[16] Lyngdoh AC, Choudhury NA, Moulik S. Diabetes disease prediction using machine learning algorithms. 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), Langkawi Island, Malaysia. 2021:517-521.

[17] Kumari S, Kumar D, Mittal M, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int J Cog Comp in Eng.* 2021;2:40-46

[18] Chang V, Ganatra MA, Hall K, Golightly L, Xu QA. An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics.* 2022;2(1):100118.

[19] Yakut Ö. Diabetes prediction using colab notebook-based machine learning methods. *IJCESEN.* 2023;9(1):36-41.

[20] The Python Library Reference, Release 3.8.8, Python Software Foundation. Available online: https://www.python.org/downloads/release/python-388/ (accessed on 10 May 2023).

[21] Pima Indians Diabetes Database | Kaggle, https://www.kaggle. com/datasets/uciml/pima-indiansdiabetes-database/ Accessed 20 Nov. 2023.

[22] AlZubi, A. A., Alarifi, A., & Al-Maitah, M. (2020). Deep brain simulation wearable IoT sensor device based Parkinson brain disorder detection using heuristic tubu optimized sequence modular neural network. *Measurement, 161*, 107887. https://doi.org/10.1016/j.measurement.2020.107887

[23] Joshi, AP, Patel BV, Data preprocessing: The techniques for preparing clean and quality data for data analytics process. Orient. *J Comput Sci Technol.* 2021;13(0203):78-81.

[24] Ahsan MM, Mahmud MP, Saha PK, Gupta KD, Siddique Z. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies.* 2021;9(3):52.

[25] Yin, H., Mukadam, B., Dai, X., & Jha, N. K. (2019). DiabDeep: Pervasive diabetes diagnosis based on wearable medical sensors and efficient neural networks. *IEEE Transactions on Emerging Topics in Computing, 9*(3), 1139-1150. https://doi.org/10.1109/tetc.2019.2958946

[26] Prusty S, Patnaik S, Dash SK. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Front Nanosci.* 2022;4:972421.

[27] Yang, G., Liu, S., Li, Y., & He, L. (2023). Short-term prediction method of blood glucose based on temporal multi-head attention mechanism for diabetic patients. Biomedical Signal Processing and Control, 82, 104552.

[28] Ibrahim I, Abdulazeez A, The role of machine learning algorithms for diagnosing diseases. *J App Sci Techol Trends.* 2021;2(01):10-19.

[29] Koutny, T., & Mayo, M. (2022). Predicting glucose level with an adapted branch predictor. Computers in Biology and Medicine, 145, 105388.

[30] Belete DM, Huchaiah MD, Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int J Comput Appl.* 2022;44(9):875-886.