DOI: https://doi.org/10.52756/ijerr.2023.v30.004 *Int. J. Exp. Res. Rev.*, **Vol. 30: 32-45 (2023)**

Original Article | *Peer Reviewed* | *Open Access*

International Journal of Experimental Research and Review (IJERR)

ISSN: 2455-4855 (Online)
www.iaph.in
ISSN 2455-4855

GLSTM: A novel approach for prediction of real & synthetic PID diabetes data using GANs and LSTM classification model

Sushma Jaiswal and Priyanka Gupta*

Department of Computer Science & Information Technology, Guru Ghasidas Vishwavidyalaya, Bilaspur, India

E-mail/Orcid Id:

SJ, jaiswal1302@gmail.com, https://orcid.org/0000-0002-6253-7327; *PG,* priyanka13666@gmail.com, https://orcid.org/0000-0001-8643-6857

Article History:

Received: 27th Jan., 2023
Accepted: 04th Mar., 2023
Published: 30th Apr., 2023

Keywords:
GLSTM, GAN, ML, PID, TVAE

Abstract: Generative Adversarial Network (GAN) is a revolution in modern artificial systems. Deep learning-based Generative adversarial networks generate realistic synthetic tabular data. Synthetic data are used to enhance the size of a relatively small training dataset while ensuring the confidentiality of the original data. In this context, we implemented the GAN framework for generating diabetes data to help the health care professional in more clinical applications. GAN is used to validate the Pima Indian Diabetes (PID) Dataset. Various preprocessing techniques, such as handling missing values, outliers and data imbalance problems, enhance data quality. Some exploratory data analyses, such as heat maps, bar graphs and histograms, are used for data visualisation. We employed hypothesis testing to examine the resemblance between real data and GAN-generated synthetic data. In this study, we proposed a GAN-Long Short-Term Memory (GLSTM) system, in which GAN is used for data augmentation, and LSTM is used for diabetes classification. Additionally, various GAN models such as CTGAN, Vanilla GAN, Coupula GAN, Gaussian Coupula GAN, and TVAE GAN are used to generate the synthetic dataset. Experiments were conducted on real data, synthetic data, and by combining real and synthetic data. The model that used both real and synthetic data obtained a substantially better accuracy of 97% compared to 92% when only real data was used. We also observed that synthetic data could be used in place of real data, as the mean correlation between synthetic and real data is 0.93. Our study's findings outperformed when compared to state-of-the-art methodologies.

Introduction

The worldwide frequency of diabetes cases (Saeedi et al., 2019) in 2019 was 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 frequency of individuals diagnosed with diabetes tends to rise with increasing age. In low- and middle-class families, 3 out of 4 people suffer from this problem all over the world. Many such people are living with diabetes (1 in every 2 adults, approx. 240 million), but they do not know they have it. Type 2 diabetes affects 90% of the population, whereas type 1 diabetes affects around 10% of the population. Around one in four adults aged 60 to 65 and beyond have diabetes, equating to over 25% of this population. According to the WHO's data assessment report, the number of diabetics, which was 108 million in 1980, increased to 422 million in 2014. The World Diabetes Federation estimates that 537 million adults (aged 20 to 79) worldwide had diabetes as of 2021, equal to 1 in 11 adults. The Centres for Disease Control and Prevention (CDC) states that 463 million adults worldwide have diabetes. In this regard, we present a machine learning-based application to forecast the risks of DM risk factors. In light of these considerations, the framework was created to yield more precise findings when contrasted with other research, which revealed that

***Corresponding Author:** priyanka13666@gmail.com

4.2 million fatalities occurred in 2019, making it among the principal contributors to global mortality.

Diabetes is an incurable condition that causes blood glucose levels to rise over time. Disturbances in the quantity of sugar in the blood cause this deadly condition. It is a condition in which the pancreas does not function properly; as a result, the body has a variety of challenges, including eyesight, heart, nerve, and kidney disease. (Aruna Kumari et al., 2022) Diabetes causes many biological tissue damage in people's bodies due to elevated blood glucose levels. People across the globe are impacted by diabetes, which is grouped into three types: Type 1, Type 2, and gestational diabetes. Insufficient insulin production by the pancreas leads to Type 1 diabetes, and it must be injected into the body from outside sources to keep glucose levels stable. This kind of diabetes is more frequent among teenagers. Type 2 diabetes results from the body's incapability to adequately utilize the insulin that is produced by the pancreas. It appears when the body's metabolic mechanism cannot digest food completely, leading to increased blood sugar levels. One of the causes of this type of diabetes could be hereditary. This type of diabetes is most frequent in adults over 45. Hormonal changes and increased insulin production during pregnancy cause gestational diabetes.

Reportedly, Machine learning (ML) techniques are used extensively in many areas, such as the diagnosis of diabetes and cancer (Anil et al., 2022; Monirujjaman et al., 2022; Rufo et al., 2021), Covid-19 (Meraihi et al., 2022), meningitis (Mentis et al., 2021), coronary heart disease (Akella and Akella, 2021; Heo et al., 2022), and hypertension (Islam et al., 2022).

A broad spectrum of models, based on data mining and machine learning techniques, has been developed by researchers to accurately predict the risk of diabetes mellitus occurrence. But the results obtained from these models could be better. So, the predictions are not trustworthy. In this regard, we present a deep learning-based application to forecast the risks of DM risk factors.

To the extent of our knowledge, we proposed GLSTM architecture for diabetes prediction on the PID dataset for the first time in the literature. GAN is employed for data augmentation, while LSTM is used for diabetes classification. We have developed the diabetes classification model using only real data. Then, another model is built using the combined real and synthetic data. Various GAN models, such as CTGAN, Vanilla GAN, Coupula GAN, Gaussian Coupula GAN, and TVAE GAN, are used to generate the synthetic dataset. To prepare the data, relatively new techniques arrange being used. For example, the log transformation technique eliminates outliers, and the MICE imputation technique fills in the gaps left by missing values. To create a balanced dataset, we investigated data-level techniques that modify the data distribution to equalize the frequency of events that belong to two distinct categories. The proposed framework provides better accuracy, sensitivity, specificity, ROC-AUC, and F1 score than sophisticated methodologies.

ML and DL classifiers, such as Decision Trees (Saxena et al., 2021), Artificial Neural Networks (Jaiswal & Gupta, 2021), Support Vector Machines (SVMs), Bayesian Networks, ensemble techniques (Jaiswal and Gupta, 2022), and Convolutional Neural Networks have been widely employed in PID dataset for diabetes prediction. However, these approaches need a large amount of data to be trained appropriately. Data selection is often complex and challenging since most data is private and subject to robust privacy regulations. There are probably many personal details about a patient's health in their medical records at the health center (Zhu et al., 2020), Regardless of whether the names and ID numbers are removed before the data is revealed. It is possible that specific persons can be recognized using a combination of other criteria such as age, gender, height, and weight. Because of the possibility of re-identification, several laws have been put in place to limit the usage of personal information of the patient using various databases. One approach to overcome this problem would be constructing a sufficiently accurate synthetic data set based on the original dataset. Diabetes mellitus may be predicted using GAN models. GANs have recently demonstrated impressive performance on a variety of tasks, including producing realistic pictures (Zhou et al., 2023), synthesizing electronic health records (Baowaly et al., 2019), and forecasting financial time series data (Takahashi et al., 2019).

In order to address this discrepancy, we will examine the application of Generative Adversarial Networks (GANs) in the simulation of synthetic tabular data. The LSTM model is used for the classification of diabetes mellitus and determines if a person has the disease or not. The arrangement of the article is in the following manner: Section 2 presents prior studies and commonly used classification methods for diabetes. Section 3 describes the methodology used, which includes defining the problem, data preprocessing, and models, with a particular emphasis on the suggested GAN-LSTM algorithm. The findings are presented in Section 4, and Section 5 contains a discussion. Lastly, Section 6 concludes the article and offers recommendations for future research.

DOI: https://doi.org/10.52756/ijerr.2023.v30.004

According to the study, traditional models are not as effective, and accuracy in predicting diabetes diagnosis remains a complex topic to be studied further. In light of this, deep learning models are data intensive. Data is limited in the medical field, and finding large amounts of data on specific diseases is challenging. Also, the datasets available for diabetes prediction are inadequate for proper processing. The researchers face challenges due to the small size, missing values, outliers, and lack of specific information in the diabetes datasets. Available solutions are expensive and require complicated mathematical models for immediate diagnosis. Cost-effective solutions should be developed for all individuals (diabetics, pre-diabetics, and healthy people).

Materials and Methods

The proposed model evaluates both real and fake data generated by the GAN model and combined data containing both real and fake data. The flow diagram of the research work is shown in Figure 1. In this diabetes prediction model, visualization and simulation are accomplished using several libraries, including NumPy, pandas, scikit, matplotlib, and sea born. Gathering relevant patient information is essential; first, selecting and preprocessing the PID dataset. Preprocessed data is input to the GAN model. For classification, preprocessed real data, GAN-generated data, and a combination of both are fed into the LSTM network. The output of the LSTM model included accuracy, precision, recall, F1-score, and Log-loss.

Real Data Set (PID)

In this article, the PID dataset (Albahli, 2020) is addressed and obtained from the UCI Machine Learning repository. PID dataset contains information about whether a person has a diabetic problem. It incorporates 768 samples and 9 features such as Diabetic Pedigree Function, Pregnancies, Glucose, Insulin, Skin_Thickness, Blood_Pressure, BMI, Age, and Outcomes. One can determine whether a patient has diabetes (Yes) or not premised on the outcome variable (No). This demonstrates the need to address the issue of binary categorization in this scenario. The initial features are listed in Table 1, along with a brief description of each.

Synthetic Dataset (GAN generated)

The number of features is identical between the original data and the data generated by the GAN model, but there are significantly more records. 768 samples and 9 features are present in the original data, while 2000 samples and 9 features are present in the data produced by GAN.

Proposed GLSTM based model for generating PID synthetic data

GAN

The GAN architecture was first described in 2014 by Ian Good Fellow et al. (2020). GANs are generative models useful for synthetic data creation in data science. GAN models are implicit density unsupervised models that supervise themselves. The model learns from the statistical properties of the data. There are two networks in the model, one network called a generator which works to generate data, and a second network called a discriminator. The discriminator classifies the data generated by the generator, and after classifying it, it gives feedback to the generator so that generator updates the model using the back propagation technique. We keep the generator constant during the discriminator training phase and vice versa. Both generator and discriminator use neural networks.

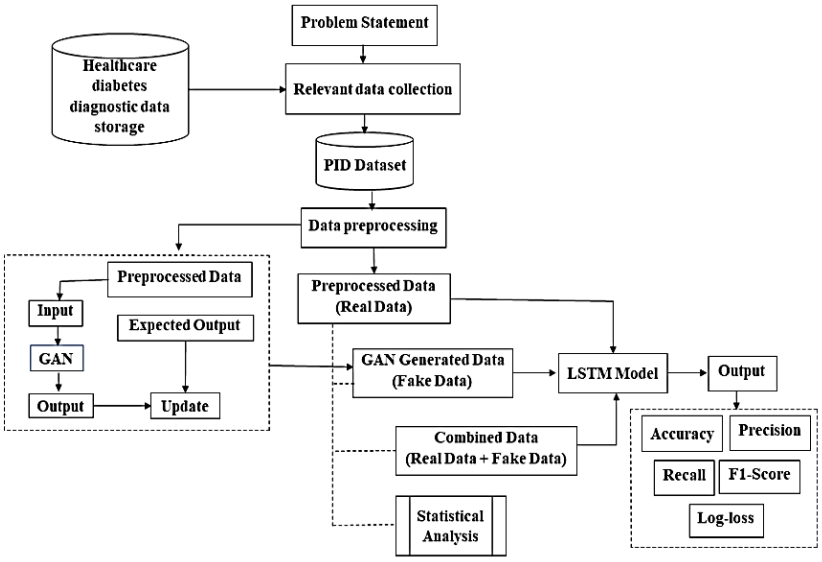


Figure 1. Block diagram of GLSTM-based system for diabetes prediction.

DOI: https://doi.org/10.52756/ijerr.2023.v30.004

Generative

Generative models generate the data. It means it creates new fake instances.

Adversarial

Adversarial model classifies the data generated by the generator and gives feedback to it.

Network

It is a neural network that takes input and, after the calculation, gives output.

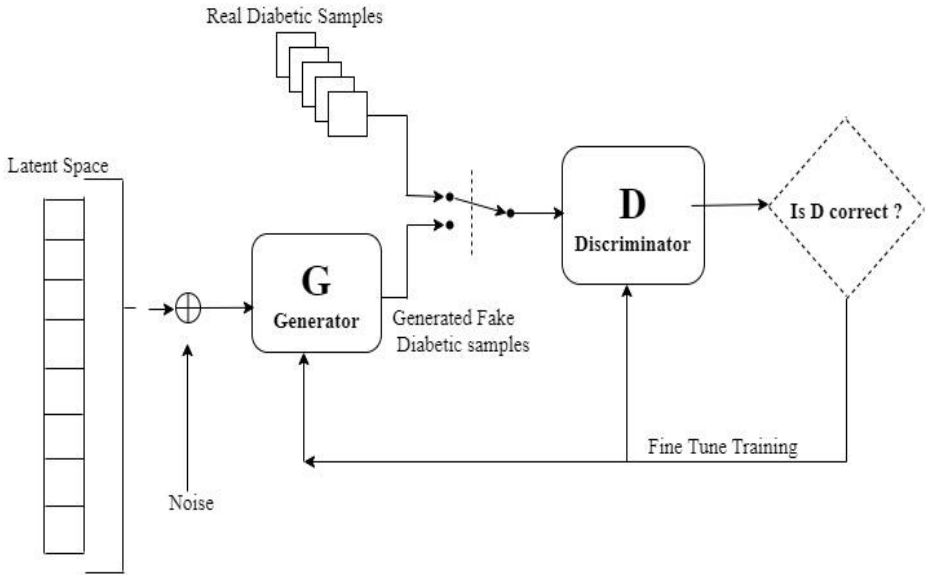


Figure 2. The Framework of Generative Adversarial Network

The following sections explain the process & functionality of GAN

GAN is a combination of two models. There are two networks called generator and discriminator; both networks are neural networks. They have an input layer, a hidden layer and an output layer. The GAN model's generator part learns from latent space's probability distribution. It uses Gaussian distribution techniques for creating synthetic data. The synthetic data pass through the discriminator as an input, and the discriminator utilizes a conventional classification technique to classify the data generated by the generator. Discriminator identifies the similarity between real data and synthetic data, and based on this they classify them as real or fake. After the discriminator's classification, the generator has to update the model with the back propagation technique. Based on the discriminator's feedback, the generator creates new samples. The iteration will perform until the generator fools the discriminator, so the discriminator classifies the fake data as real data. The generator and discriminator both act as competitors of each other. Once the discriminator predicts fake samples as real, then the training is complete.

Here the generator endeavors to scale back the following characteristic while the discriminator makes an effort to maximize it. The formulation in equation 1, derives from the cross entropy between the real and fake distribution.

$$\min_G \max_D V(D,G) = E_{x\sim P_{data}(x)}[\log D(x)] + E_{z\sim P_z(z)}[\log(1 - D(G(z)))] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

- D(x) is the discriminator's estimate of the probability that real data instances x are real
- E_x is the predicted value overall real data instances
- $E_{x\sim Pdata(x)}$ is the probability distribution of real data
- G(z) is the generator's produced value when given noise z
- $E_{z\sim Pz\,(z)}$ is the Probability distribution of the noise
- D(G(z)) is the discriminator's estimated probability that a fake instance is real
- E_z is the predicted value over the entire random inputs to the generator

The framework of the GAN model is illustrated in Figure 2, where the discriminator and generator are denoted as D and G, respectively. $E_{x\sim P_{data}(x)}$ is a real sample, and $E_{z\sim P_z(z)}$ is generated synthetic samples of random Gaussian distribution. It is a probability distribution of the latent space. When x is generated from P_{data}, the Discriminator classifies the generated data as real data. G (z) is a generator's output when given G (z) as input to the Discriminator. The Discriminator aims to maximize [1-D{G(z)}] to ensure that D [G (z)] becomes 0. In contrast, the Generator wants to minimize [1-D{G(z)}] to force the probability of D[G(z)] to 1, causing the Discriminator to make an error in identifying generated data as real. Therefore, instead of minimizing

DOI: https://doi.org/10.52756/ijerr.2023.v30.004

log [1−D (G (z)], the Generator G can be trained to maximize log D [G (z)] to overcome this issue.

Exploratory Data Analysis & Data preparation

Data Visualization

Information is presented visually through data visualization to make it easier to comprehend. This section displays the information as a bar chart, box plot, heat map, etc. The bar graph analysis reveals the proportion of people who have diabetes. A box plot can detect an outlier whether or not it is present in the dataset. The statistics show the values' first quartile, median, and third quartile ranges as well as their highest and lowest values. The heat map demonstrates the variables' correlation, ranging from -1 to +1.

Data Wrangling

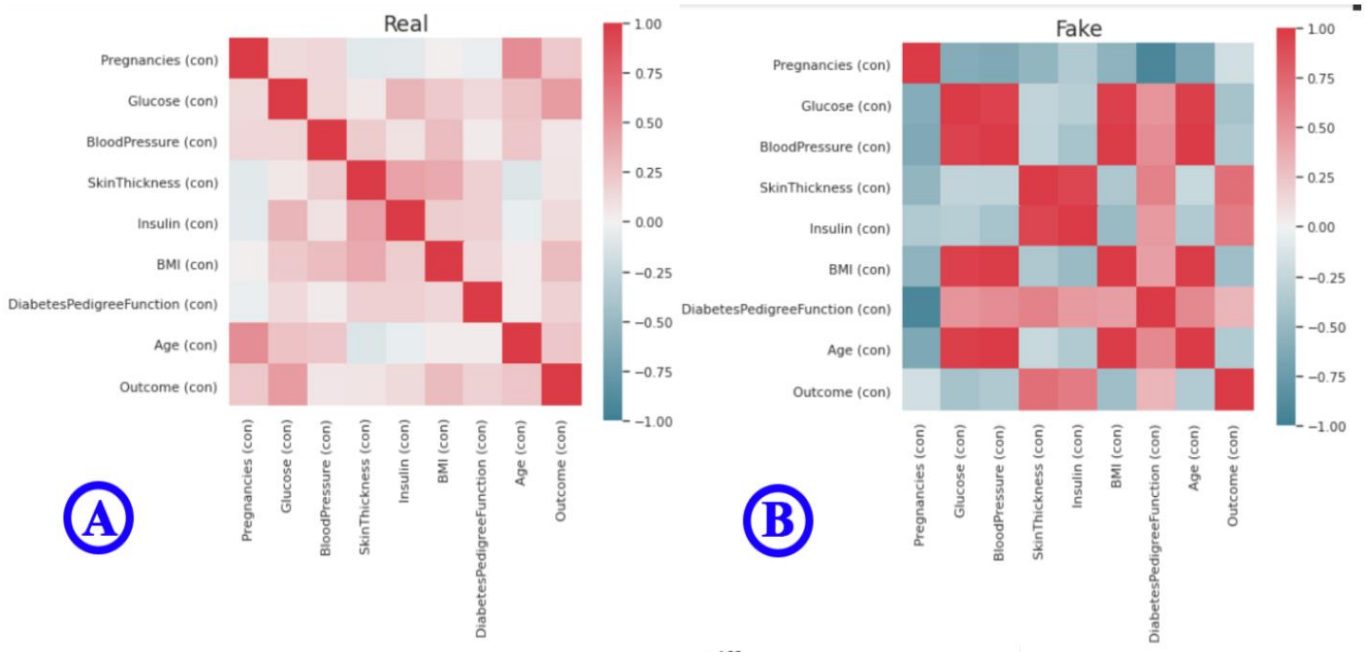
Data wrangling is the process of reducing redundant values, data discretization, and feature selection. It must be improved in order to be suitable for data analysis and machine learning tasks. The data standardization and outlier removal are included in this section. A model was developed using the data that had been processed. There are 500 people without diabetes and 268 patients with diabetes in the PID data set, which includes 768 patients. The dataset can be balanced in two ways: under-sampling and oversampling. We employ the Proximity Weighted Synthetic Oversampling (PROW) technique, which does not depend on detecting k-nearest neighbors, to over

Table 1. PID dataset description

Sl. No.	Attributes	Mean	SD	Min./max.	Missing Values
1.	Number of pregnancies	3.8	3.4	1/17	0
2.	The level of plasma glucose	120.9	32	56/197	5
3.	Diastolic_BP	69.1	19.4	24/110	35
4.	Skin_Thickness(mm)	20.5	16	7/52	227
5.	2-Hour serum insulin level	79.8	115.2	15/846	374
6.	BMI (kg/m2)	32	7.9	18.2/57.3	0
7.	Family history of diabetes	0.5	0.3	0.0850/2.32	0
8.	Age	33.2	11.8	21/81	0
9.	Class	-	Yes/No	Diabetic/Non-diabetic	-

Correlation measures the strength between two continuous variables. In figure 3-4, we can see how close the synthetic data is to real data. The generated synthetic data is very close to the original data. The mean correlation is 0.9359 between fake and real columns. We could also explore real and synthetic data with a visualization plot.

sample the minority class. The significance weights are instead distributed among all the minority examples according to how far apart they are from the majority occurrences. In contrast to SMOTE, PROW generates samples along line segments between instances of the



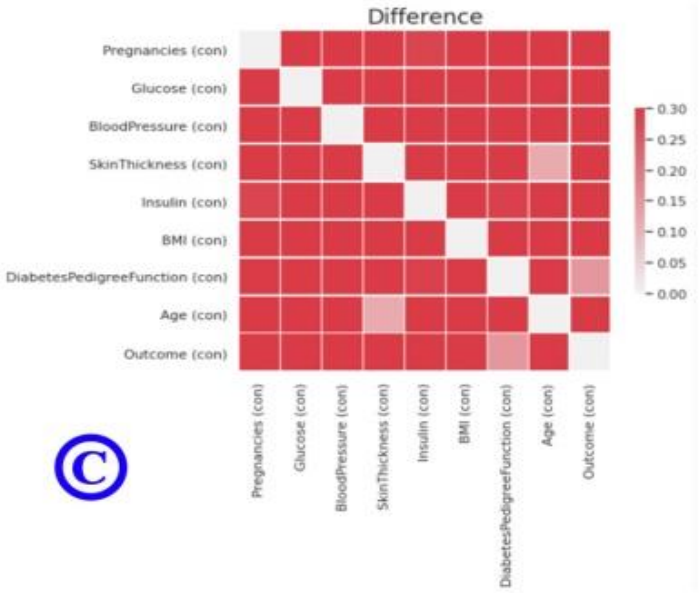


Figure 3. Correlation of the features in Real data, Fake data and their differences

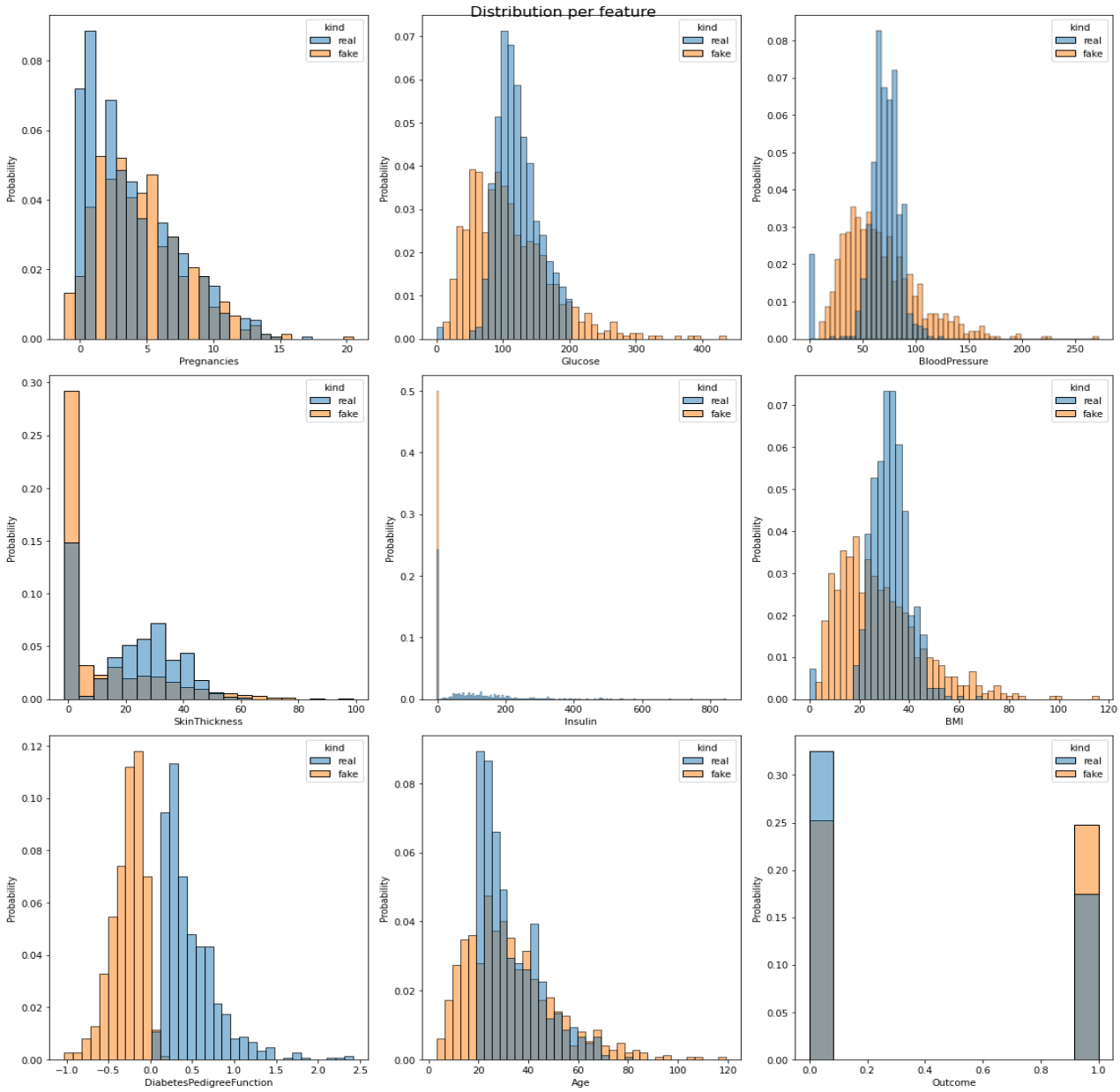


Figure 4. Synthetic data and Real data Visualization

DOI: https://doi.org/10.52756/ijerr.2023.v30.004

minority class that are relatively far apart, resulting in a wider distribution of additional occurrences throughout the minority class's manifold.

The measure of missing values for each feature is presented as a proportion as 48.56% for Insulin, 29.58% for Skin Thickness, 4.58% for Blood Pressure, 1.43% for BMI, and 0.65% for Glucose. There is no missing value in the remaining features. To fill this value, we are using MICE multivariate imputation technique. A MOUSE is an acronym for the Multivariate Imputation by Chained Equations algorithm. Univariate methods such as mean, median, mode, frequent data, and constant frequently do not provide reliable information for missing values because they impute missing values in that column using that particular column. The multivariate method estimates the best prediction for each missing value in a dataset by gathering information from other columns.

PID dataset's features contain outliers. Outliers are the values that are distinct from the rest of the data. It causes a problem during model fitting. When we feed these features into the model as is, one of them may have a significantly more prominent influence on the outcome due to its higher value. To give equal weight to each feature, feature scaling is required; we use the log transformation technique to accomplish these values. It reduces data skewness and can aid in transforming a non-linear model into a linear one.

We can understand more about data imbalance from figure 5. Here, we can see that diabetics' records are significantly higher than non-diabetics. It has been observed that the imbalanced data gives poor performance.

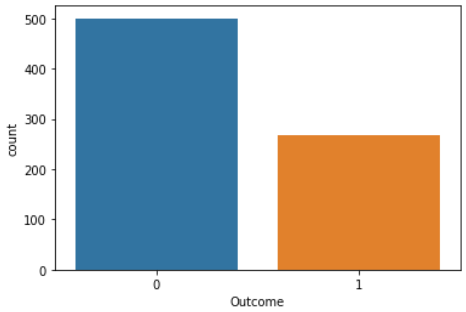


Figure 5. Distribution of the dependent features

Table 2. Statistical Significance Test Results

Features	Vanila GAN	Coupula GAN	Gausian Coupula GAN	CTGAN	TVAE Model
BMI	0.259814	0.17321	0.24611	0.12101	0.31345
Blood_Pressure	0.26678	0.12911	0.251879	0.10234	0.32811
Skin_Thickness	0.280832	0.18924	0.2742	0.01201	0.36281
Glucose	0.000062	0.000003	0.000041	0.000001	0.00251
Diabetes_Pedigree_Function	0.788961	0.28921	0.75321	0.16721	0.88235
Insulin	0.560453	0.17821	0.54329	0.13985	0.67021
Pregnancies	0.152257	0.02331	0.13257	0.01278	0.18921
Age	0.78002	0.28902	0.65439	0.21784	0.89209
Outcome	0.269629	0.16021	0.25481	0.02349	0.321099

Hypothesis Testing

A statistical method known as testing for hypotheses determines whether a relationship between real data and artificial data is statistically significant or more likely to be the product of randomness. We calculated the P-value to measure the difference between the two distributions (Probability Score).

Our Null Hypothesis (H0), shown in equation 2, suggests that the mean of a specific variable in the actual dataset (Mean 1) is equivalent to the mean of the same variable in the simulated dataset (Mean 2). If the p-value is less than 0.05, we reject H0 and infer that the two means have a significant difference.

Ideally, we want the two means to be the same; hence, we do not want to reject H0 for as many variables as possible. In general, a high p-value indicates that the null hypothesis is more likely to be true; a low p-value indicates that the null hypothesis is more likely to be rejected. Therefore, the two means are considered different if the p-value is greater than 0.05.

$$Ho: Mean1 = Mean2\ VS\ H1: Mean1 \neq Mean2 \ldots (2)$$

3.6 Machine learning classification model

Long- Short Term Memory (LSTM)

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

DOI: https://doi.org/10.52756/ijerr.2023.v30.004

Table 3. Hyperparameters of the GLSTM framework

Hyperparameters	Range	GAN model	LSTM model
Learning rate	0.001, 0.01, 0.1, 0.002…..	0.002	0.001
Batch Size	8, 16,32, 64, 128, 512	32	128
Number of epochs	10, 20, 30, 40, 50	30	20
Optimizer	NAdam, SGD, Adam, Adagrad	-	Adam
Generator Optimizer	SGD, Adam, RMSprop, Adagrad	SGD	-
Discriminator Optimizer	SGD, Adam, RMSprop, Adagrad	SGD	-
Number of hidden layers	1-10	3	5
Activation Function	ReLU, Leaky ReLU, tanh, sigmoid, Linear	sigmoid	Leaky ReLU
Loss function	Binary Cross-Entropy, Wasserstein loss, Mean squared error	Binary Cross Entropy	Binary Cross Entropy

multiplicative units: input gate, forget gate and output gate.

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

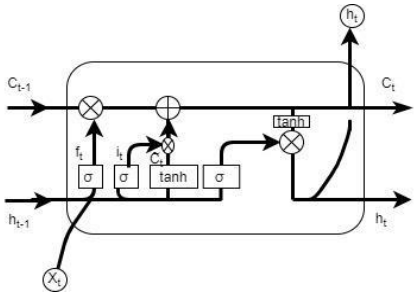


Figure 6. The architecture of LSTM cell

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

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

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

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

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

$$\widetilde{C_t} = \tanh(W_c . [h_{t-1}, x_t] + b_c), \ldots\ldots\ldots . (5)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}, \ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

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

$$O_t = \sigma(W_o . [h_{t-1}, x_t] + b_o), \ldots\ldots\ldots\ldots\ldots\ldots .. (7)$$

$$h_t = O_t * (\tanh(C_t)), \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . (8)$$

In mathematical statement equation 1 to equation 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 is correlated with weight matrices. Tanh 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.

The LSTM model has been acquired for diabetes disease binary classification. In the LSTM 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 equation (9). The balanced records are reshaped in equation (10). It complies with the input requirement of the deep neural network framework.

$$input = \sum_{i=1}^{8} Y_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . (9)$$

$$Tensor_{input} = reshape(Y_i) \ldots\ldots\ldots\ldots (10)$$

Hyperparameter Tuning

Hyperparameter's values are used to control the learning process, which is described in table 3. In building Machine learning models, hyperparameter tuning is defined as selecting the appropriate collection of values. Among the hyper-parameters are batch size, learning rate, generator, number of epochs, discriminator optimizer, number of units in a dense layer, number of layers, loss function, activation function, and certain properties like the dropout layer keep probability and batch normalization momentum. Stochastic Gradient Descent is used to train GAN; one difficulty is carefully determining the learning rate and batch size.

Table 4. Comparative analysis of Balanced and Imbalanced Data

Classifiers (LSTM)	Imbalanced Real Data	Balanced Real Data (using PROW)
Accuracy	85%	92%
Precision	82%	92%
Recall	81%	92%
F1-Score	81%	92%
Log-loss	0.23	0.14
AUC	0.81	0.92

DOI: https://doi.org/10.52756/ijerr.2023.v30.004

Table 5. Comparison of Fake data with different GAN's Outcomes

Classifiers (LSTM)	Fake Data (Vanila GAN)	Fake Data (Coupula GAN)	Fake Data (Gausian Coupula GAN)	Fake Data (CTGAN)	Fake Data (TVAE Model)
Accuracy	89%	86%	82%	74%	**92%**
Precision	89%	87%	80%	78%	**93%**
Recall	89%	86%	82%	74%	**92%**
F1-Score	89%	86%	81%	73%	**92%**
AUC	0.89	0.86	0.82	0.74	**0.94**

Given that the GAN is trained using SGD, using back propagation, the model's weights are updated by calculating the error gradient for the current state. Weight updates occur during the training period. An extreme value could result in an early convergence. A too-small value, on the other hand, can create excruciatingly delayed convergence. A larger batch size offers a more precise estimation of the error gradient. The likelihood that increased performance will result from tweaking the model weights is higher. Another option is to employ a small batch size, which results in a less precise estimation. To address the over fitting problem, we use a 10-fold cross-validation technique. The grid search tuning for the hyperparameters is shown in Table 3 as a key-value pair.

Table 6. Comparison of combined data with different GAN's Outcomes

Classifiers (LSTM)	Combined Data (Vanila GAN)	Combined Data (Coupula GAN)	Combined Data (Gausian Coupula GAN)	Combined Data (CTGAN)	Combined Data (TVAE Model)
Accuracy	95%	90%	94%	84%	**97%**
Precision	94%	91%	94%	88%	**97%**
Recall	94%	92%	94%	84%	**96%**
F1-Score	94%	92%	94%	83%	**96%**
AUC	0.96	0.92	0.94	0.84	**0.97**

Results

Experimental Setup

In this experimental purpose, a set of programs that are run and carried out on a workstation system that meets the following requirements to demonstrate all experiments on the Python 3.11.1 development environment. Intel(R) Core (TM) i5, RAM 8 GB, and the system type is the 64-bit operating system, x64-based processor.

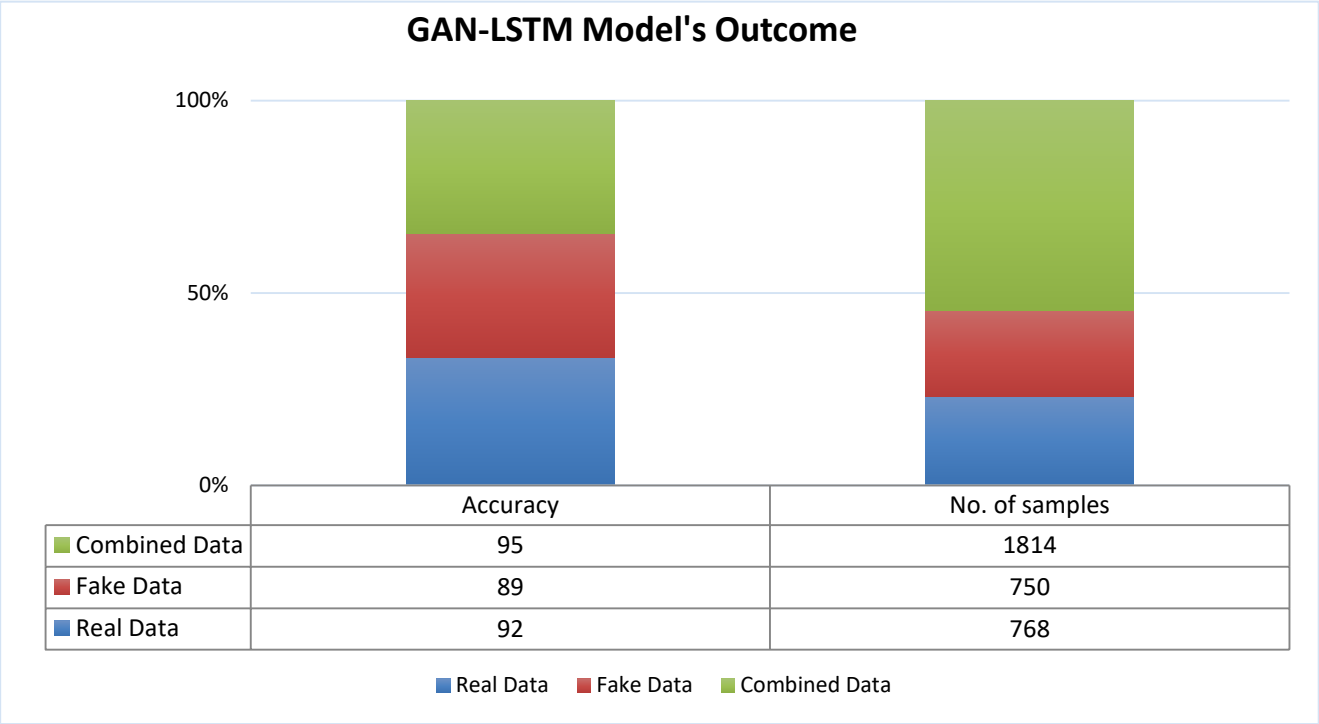
After completing the data preprocessing task, the preprocessed data is fed into the GAN framework for data augmentation called augmented data. Then we have real data, which is preprocessed data. Augmented data is GAN-generated data and a combination of real and augmented data. The dataset was partitioned into training and testing sets and the deep learning model LSTM algorithms were employed to classify the data. In order to achieve the most favorable results for the dataset, a combination of cross-validation methods and hyperparameter tuning was executed.

Table 4 represents the result of balanced and imbalanced data. Balanced data gives better results as compared to imbalanced data.

Tables 5 and 6 represent the classification reports of different GAN results. The method exhibited impressive performance characterized by high prediction accuracy, AUROC, and other metrics, including precision, recall, and F1-Score.

The simulation result shows in the above graph. Figure 7 represents the GAN-LSTM models' performance. Real data contains 768 instances and achieves 92% accuracy, fake data includes 750 cases and 91% accuracy, while combined data has 1413 instances and achieves 97% accuracy. The Roc curves for the real, synthetic, and combined data are shown in Fig. 8 and are 0.92, 0.94, and 0.97, respectively.

DOI: https://doi.org/10.52756/ijerr.2023.v30.004



	Accuracy	No. of samples
Combined Data	95	1814
Fake Data	89	750
Real Data	92	768

Figure 7. GAN-LSTM model results for Pima Indian Diabetes Dataset

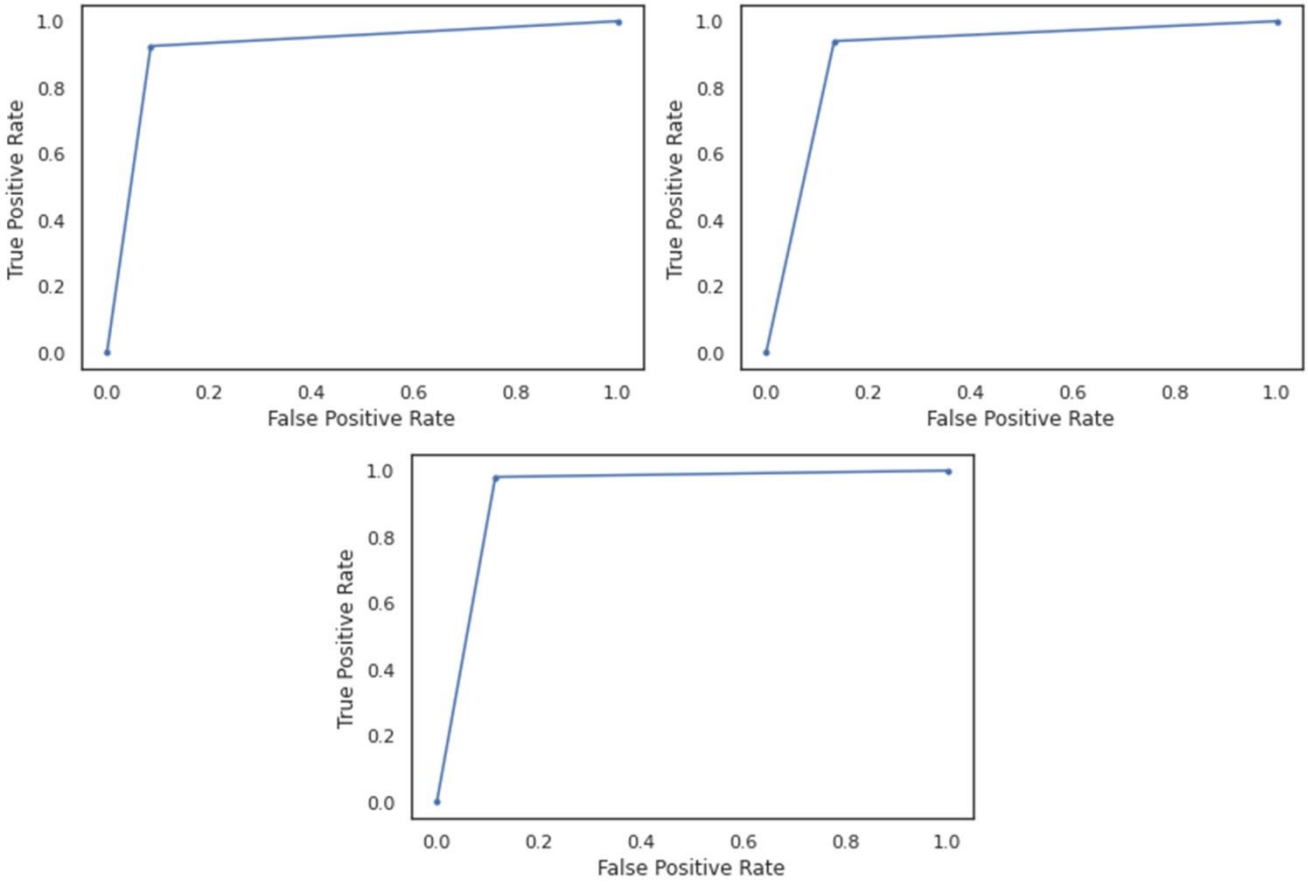


Figure 8. Roc curve for Real data, Fake data and combined data

DOI: https://doi.org/10.52756/ijerr.2023.v30.004

Discussion

Furthermore, comparison with benchmarking classifier, our findings outperformed the results presented by other authors. We also used the same dataset and performance metrics that were employed in previous studies. Table 5 lists the number of writers who have used various deep learning techniques, including MLP and LSTM, as well as more conventional methods like SVM, logistic regression, naive bayes, decision trees, and KNN. Both of which produce superior results, but our result is superior in terms of precision.

Table 7 summarises discussions from various researchers' perspectives on developing non-invasive real-time systems and the technique for detecting diabetic illness. The table below indicates that the performance of the GLSTM model surpasses that of the current state-of-the-art methods.

created synthetic dataset. The MICE Imputing approach is applied to balance the classes on the real and combined data sets. We integrated the GLSTM architecture to improve the overall accuracy of diabetes prediction. We produced numerous synthetic data sets using multiple GAN models and compared their results to discover the best synthetic data.

The proposed GLSTM framework is effective and efficient, with a success rate of 97% when simulated on the test PID dataset. Additionally, we evaluated the system against cutting-edge techniques, and our system outperformed them all. This comparison demonstrated the potential of a DNN-based diabetes prediction system for achieving significant benefits. The presented DNN method's superiority could be seen through performance evaluation of LSTM models utilising several performance measures, including

Table 7. Comparison of benchmarking classifiers with our proposed work

Sl. No.	Authors	Techniques	Classification Accuracy
1	Ramezani et al., 2018	LANFIS intelligent system	88.5%
2	Daanouni et al., 2020	ANN	82%
3	Saxena et al., 2022	KNN, RF, DT, MLP	79.8%
4	Mahboob Alam et al., 2019	ANN	75.7%
5	Kaur and Kumari, 2022	SVM-linear, RBF, k-NN, ANN and MDR	89%
6	Kumari et al., 2021	LR, K-NN, SVM, NB, DT, RF, Soft voting classifier, Adaboost, Bagging, GB, XGBoost, Catboost	79.08%
7	Rajendra and Latifi, 2021	LR, DT, SVM, K-NN, NB, Ensemble Model	77.8%
8	Li et al., 2023	(GA-Kmeans, GA-PSO- Kmeans, HR-Kmeans) with K-nn	91.65%
9	Bukhari et al., 2021	ABP-SCGNN	93%
10	Butt et al., 2021	MLP, LSTM	87.26%
11	Alaa Khaleel and Al-Bakry, 2021	LR, NB, and KNN	94%
12	Roopa and Asha, 2019	PCA-LRM	82.1%
13	Azad et al., 2021	Genetic Algorithm and Decision Tree	82.12%
14	Alex et al., 2022	NB, LR, SVM, LSTM, DCNN	86.29%
15	Our Proposed Work	GAN-LSTM	97%

Conclusion

In this article, we've discussed techniques that make healthcare more accessible. The findings of this study might be helpful to doctors, researchers, students, and medical professionals working on research and development. We investigated frequent issues encountered when training GANs. For GAN training, much integration of hyperparameters were utilized. To identify which combination works best, we used hypothesis testing to assess the similarity of the real and accuracy, specificity, sensitivity, accuracy, and F1 score. In the future, we will examine other network designs and data sets with varying characteristics.

Conflict of interest

Nil

References

Akella, A., & Akella, S. (2021). Machine learning algorithms for predicting coronary artery

disease: Efforts toward an open source solution. *Future Science OA*, *7*(6). *7*(6), FSO698. https://doi.org/10.2144/FSOA-2020-0206

Alaa Khaleel, F., & Al-Bakry, A. M. (2021). Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings*. https://doi.org/10.1016/J.MATPR.2021.07.196

Albahli, S. (2020). Type 2 Machine Learning: An Effective Hybrid Prediction Model for Early Type 2 Diabetes Detection. *Journal of Medical Imaging and Health Informatics*, *10*(5), 1069–1075. https://doi.org/10.1166/JMIHI.2020.3000

Alex, S. A., Nayahi, J. J. V., Shine, H., & Gopirekha, V. (2022). Deep convolutional neural network for diabetes mellitus prediction. *Neural Computing and Applications*, *34*(2), 1319–1327. https://doi.org/10.1007/S00521-021-06431-7/FIGURES/3

Anil Kumar, C., Harish, S., Ravi, P., Svn, M., Kumar, B. P. P., Mohanavel, V., Alyami, N. M., Priya, S. S., & Asfaw, A. K. (2022). Lung Cancer Prediction from Text Datasets Using Machine Learning. *BioMed Research International*, *2022*. https://doi.org/10.1155/2022/6254177

Aruna Kumari, G. L., Padmaja, P., & Suma, J. G. (2022). A novel method for prediction of diabetes mellitus using deep convolutional neural network and long short-term memory. *Indonesian Journal of Electrical Engineering and Computer Science*, *26*(1), 404–413. https://doi.org/10.11591/IJEECS.V26.I1.PP404-413

Azad, C., Bhushan, B., Sharma, Rohit, Shankar, A., Krishna, ·, Singh, K., Khamparia, A., Sharma, R., & Singh, K. K. (n.d.). *Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus*. *1*, 3. https://doi.org/10.1007/s00530-021-00817-2

Baowaly, M. K., Lin, C. C., Liu, C. L., & Chen, K. T. (2019). Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association : JAMIA*, *26*(3), 228. https://doi.org/10.1093/JAMIA/OCY142

Bukhari, M. M., Alkhamees, B. F., Hussain, S., Gumaei, A., Assiri, A., & Ullah, S. S. (2021). An Improved Artificial Neural Network Model for Effective Diabetes Prediction. *Complexity*, *2021*. https://doi.org/10.1155/2021/5525271

Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. *Journal of Healthcare Engineering*, *2021*. https://doi.org/10.1155/2021/9930985

Daanouni, O., Cherradi, B., & Tmiri, A. (2020). *Type 2 Diabetes Mellitus Prediction Model Based on Machine Learning Approach*. 454–469. https://doi.org/10.1007/978-3-030-37629-1_33

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Heo, J. N., Yoo, J., Lee, H., Lee, I. H., Kim, J. S., Park, E., Kim, Y. D., & Nam, H. S. (2022). Prediction of Hidden Coronary Artery Disease Using Machine Learning in Patients With Acute Ischemic Stroke. *Neurology*, *99*(1), E55–E65. https://doi.org/10.1212/WNL.0000000000200576

Islam, S. M. S., Talukder, A., Awal, M. A., Siddiqui, M. M. U., Ahamad, M. M., Ahammed, B., Rawal, L. B., Alizadehsani, R., Abawajy, J., Laranjo, L., Chow, C. K., & Maddison, R. (2022). Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries. *Frontiers in Cardiovascular Medicine*, *9*, 762. https://doi.org/10.3389/FCVM.2022.839379/BIBTEX

Jaiswal, S., & Gupta, P. (2021). MLP-DTP: Performance Evaluation of Diabetes Class Prediction. *IEMECON 2021 - 10th International Conference on Internet of*

Everything, Microwave Engineering, Communication and Networks. https://doi.org/10.1109/IEMECON53809.2021.9689183

Jaiswal, S., & Gupta, P. (2022). *Ensemble Approach: XGBoost, CATBoost, and LightGBM for Diabetes Mellitus Risk Prediction*. pp. 1–6. ttps://doi.org/10.1109/ICCSEA54677.2022.9936130

Kaur, H., & Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, *18*(1–2), 90–100. https://doi.org/10.1016/J.ACI.2018.12.004/FULL/PDF

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40–46. https://doi.org/10.1016/J.IJCCE.2021.01.001

Li, X., Zhang, J., & Safara, F. (2021). Improving the Accuracy of Diabetes Diagnosis Applications through a Hybrid Feature Selection Algorithm. *Neural Processing Letters*. pp. 1-17. https://doi.org/10.1007/S11063-021-10491-0

Mahboob Alam, T., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Imtiaz Baig, T., Hussain, A., Malik, M. A., Raza, M. M., Ibrar, S., & Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, *16*, 100204. https://doi.org/10.1016/J.IMU.2019.100204

Mentis, A. F. A., Garcia, I., Jiménez, J., Paparoupa, M., Xirogianni, A., Papandreou, A., & Tzanakaki, G. (2021). Artificial Intelligence in Differential Diagnostics of Meningitis: A Nationwide Study. *Diagnostics (Basel, Switzerland)*, *11*(4), 602. https://doi.org/10.3390/DIAGNOSTICS11040602

Meraihi, Y., Gabis, A. B., Mirjalili, S., Ramdane-Cherif, A., & Alsaadi, F. E. (2022). Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey. *SN Computer Science*, *3*(4). https://doi.org/10.1007/S42979-022-01184-Z

Monirujjaman Khan, M., Islam, S., Sarkar, S., Ayaz, F. I., Ananda, M. K., Tazin, T., Albraikan, A. A., & Almalki, F. A. (2022). Machine Learning Based Comparative Analysis for Breast Cancer Prediction. *Journal of Healthcare Engineering*, *2022*. https://doi.org/10.1155/2022/4365855

Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Programs in Biomedicine Update*, *1*, 100032. https://doi.org/10.1016/J.CMPBUP.2021.100032

Ramezani, R., Maadi, M., & Khatami, S. M. (2018). A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. *Alexandria Engineering Journal*, *57*(3), 1883–1891. https://doi.org/10.1016/J.AEJ.2017.03.043

Roopa, H., & Asha, T. (2019). A Linear Model Based on Principal Component Analysis for Disease Prediction. *IEEE Access*, *7*, 105314–105318. https://doi.org/10.1109/ACCESS.2019.2931956

Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). *Diagnostics (Basel, Switzerland)*, *11*(9). https://doi.org/10.3390/DIAGNOSTICS11091714

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., Ogurtsova, K., Shaw, J. E., Bright, D., & Williams, R. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice*, *157*. https://doi.org/10.1016/J.DIABRES.2019.107843

Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/3820360

Saxena, S., Mohapatra, D., Padhee, S., & Sahoo, G. K. (2021). Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms. *Evolutionary Intelligence*, *1*, 1–17. https://doi.org/10.1007/S12065-021-00685-9/FIGURES/17

Takahashi, S., Chen, Y., & Tanaka-Ishii, K. (2019). Modeling financial time-series with generative adversarial networks. *Physica A: Statistical Mechanics and Its Applications*, *527*, 121261. https://doi.org/10.1016/J.PHYSA.2019.121261

Zhou, X., Wei, Y., Xing, G., Feng, Y., & Song, L. (2023). *A Survey in Virtual Image Generation Based on Generative Adversarial Networks*. 137–143. https://doi.org/10.1007/978-981-99-1256-8_16

How to cite this Article:

Sushma Jaiswal and Priyanka Gupta (2023). GLSTM: A novel approach for prediction of real & synthetic PID diabetes data using GANs and LSTM classification model. *International Journal of Experimental Research and Review*, *30*, 32-45.

DOI : https://doi.org/10.52756/ijerr.2023.v30.004