Hybrid Ensemble and Deep Learning Models for Improving Diabetes Prediction

Cuu-Duong Dang
Faculty of Information Technology
HCMC University of Technology and Education
Ho Chi Minh City, Vietnam
22110124@student.hcmute.edu.vn

Thi-My-Dung Le
Faculty of Information Technology
HCMC University of Technology and Education
Ho Chi Minh City, Vietnam
22110117@student.hcmute.edu.vn

Van-Dung Hoang
Faculty of Information Technology
HCMC University of Technology and Education
Ho Chi Minh City, Vietnam
dunghv@hcmute.edu.vn

Abstract - Diabetes mellitus, a major metabolic disorder caused by insulin dysfunction, is the third leading cause of death globally. Early diagnosis plays an important task for enhancing treatment effectiveness, preventing complications, and reducing healthcare costs. This study introduces a deep learning-based framework for diabetes prediction based on key clinical features such as BMI, age, and insulin levels. To address the class imbalance problem, there are many techniques applied and compared multiple resampling techniques: Generative Adversarial Networks (GAN), and the Synthetic Minority Over-sampling Technique for Nominal and Continuous variables (SMOTENC). These techniques were integrated with several predictive models, including Multi-layer Perceptron (MLP), Long Short-Term Memory (LSTM), and Stacking ensemble methods. Experimental results demonstrate that the proposed approach achieved high accuracy, ranging from 94% to 96%. The combination of SMOTENC and MLP outperformed other approaches, reaching a prediction accuracy of 96.29%.

Keywords—LSTM, MLP, STACKING, SMOTENC

I. INTRODUCTION

Diabetes mellitus is a chronic condition characterized by prolonged high blood glucose levels due to insufficient or impaired insulin function. Type 2 diabetes, the most common form, primarily affects adults and is rapidly increasing worldwide, especially in low- and middle-income countries (LMICs). This rise is driven by factors such as urbanization, unhealthy diets, and sedentary lifestyles. According to the World Health Organization (WHO), controlling diabetes is a key objective for 2025 to mitigate severe complications like cardiovascular diseases, neuropathy, kidney damage, and vision loss, which affect 55% of patients. Global studies reveal that diabetes prevalence has surged from 7% in 1990 to 14% today, with LMICs experiencing the sharpest increases. Despite this growth, treatment rates in these countries remain low compared to high-income nations, exacerbating disparities in care. Early diagnosis can reduce the risk of diabetes and its complications. Still, traditional diagnostic processes are limited in several ways, such as differences in physician expertise, limited access to healthcare, and underdeveloped medical infrastructure in some areas. The rapid advancement of artificial intelligence (AI) has opened up new approaches in medicine. The application of AI in predicting diabetes using machine learning (ML) and deep learning (DL) algorithms can leverage historical data to analyze patient data with higher accuracy and optimize diagnostic and treatment processes. This helps alleviate pressure on hospitals and healthcare centers, enabling more patients to be diagnosed early and receive timely treatment.

This study compared the performance of machine learning and deep learning models in diabetes prediction. The deep learning models used are Multi-layer Perceptron (MLP), Long Short-Term Memory (LSTM) and Stacking model under ensemble learning technique.

The amount of the dataset for model training is an important factor in determining the accuracy of the model. This study uses some data balancing techniques such as Generative Adversarial Networks (GAN), and Synthetic Minority Over-sampling Technique for Nominal and Continuous variables (SMOTENC) to increase the number of samples of the minimum class to help minimize the phenomenon of data imbalance.

II. RELATED WORKS

Machine learning techniques have facilitated diabetes prediction. In recent years, many research groups have considered the problem of diabetes prediction using various machine learning and deep learning models, as evidenced by the contributions recorded in the reference. This section will provide an overview discussing the advances, advantages and disadvantages of the cleaning prediction techniques used.

Research groups focus on improving the comparison of models through traditional machine learning algorithms. Soni & Varma [5] proposed a prediction model using Random Forest with high accuracy compared to other traditional models. The paper [6] proposed a model using K-means clustering to preprocess the data and then applied Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes (NB) machine learning models for classification. Xue et al. [7] explored machine learning algorithms and gave good results on SVM. Talukder et al. [8] presented predictive models such as Random Forest, XGBoost and Decision Tree. The paper [9] compared Nearest Neighbors (KNN) and Naive Bayes in predicting diabetes from basic health attributes, the method is simple, easy to implement and suitable for low-resource environments. Shamim Ahmed et al. Shamim Ahmed et al. [10] presented a study using Logistic Regression combined with two XAI techniques, LIME and SHAP, the techniques have difficulty with high-dimensional data and have not solved the dependency relationship between features well. Rashi

2025 International Workshop on Intelligent Systems (IWIS) | 979-8-3315-0257-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/IWIS66215.2025.11142400

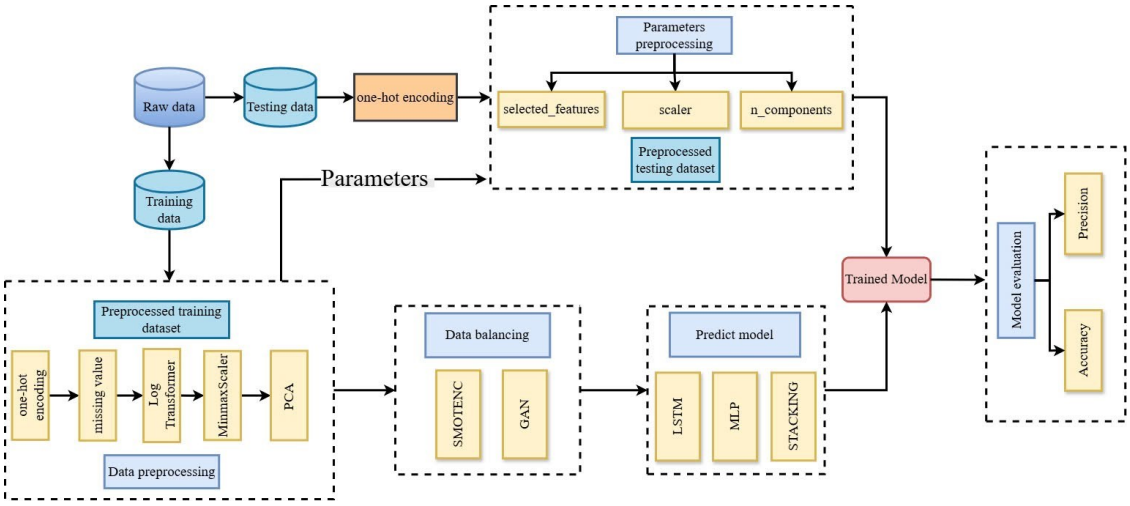

Fig 1: Diabetes prediction model architecture

Rastogi & Mamta Bansal [11] proposed a predictive model using data mining techniques, applying Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR) and Naive Bayes (NB) classification algorithms.

Some studies presented advanced machine learning models. The paper [12] presents a combination of SMOTE + RUS data balancing technique combined with Optima parameter optimization for the LightGBM model. Tasin et al. [13] proposes to use ADASYN + XGBoost technique to build the model, integrating LIME and SHAP to make the model transparent to explain the prediction results. Other studies also focus on improving the results with deep learning models. Zarghani [14] uses LSTM to take advantage of time series relationships in the dataset, but the results show that LightGBM and XGBoost achieve higher overall performance. Alex et al. [15] presents a study on deep LSTM combined with SMOTE to solve data imbalance. Srinivasu et al. [16] proposes an artificial intelligence system using CNN combined with Bi-LSTM to classify blood glucose levels through XAI integrated spectra to explain important features. The paper [17] introduces the self-designed deep learning model 2GDNN (Twice-Growth Deep Neural Network) along with RF and SVM machine learning models combined with Polynomial Regression nonlinear data processing. The method for paper [18] presents the data balancing SMOTE, ADASYN, SMOTEENN combined with a variety of models such as MLP, Random Forest, Gradient Boosting and factor analysis for each gender. In addition to the above studies, the paper [19] uses the AdaptDiab technique to select features in an ensemble style to improve the prediction performance regardless of the model combined with NOVA F-score, Fisher score, Variance threshold, Point biserial and gives synthesis results based on Mutual Information. Usama Ahmed et al. [20] proposed the Fused Machine Learning for Diabetes Prediction (FMDP) model which is a combination of Artificial Neural Network (ANN) and Support Vector Machine (SVM) merging the results using fuzzy logic system to improve accuracy by exploiting the strengths of each single model. Umair Muneer Butt et al. [21] proposed a model to classify diabetes and predict future blood sugar levels based on machine learning algorithms with LSTM, MLP combined with a simulated IoT monitoring system using BLE sensors and real-time processing using Kafka + MongoDB.

III. METHODOLOGY

The proposed approach consists of three main components: data preprocessing, data balancing, and classification. In the preprocessing phase, the raw data is cleaned and augmented using various techniques to address class imbalance. For the classification task, multiple machine learning models with different hyperparameter settings are trained to predict the diabetes status of patients. The detail of the proposed architecture is illustrated on Fig.1.

Preprocessing: First, the team treated the NULL value with the average value of the column and encoded the categorical variables into numeric form. Then the team continued to normalize the data to help the model converge faster and give more stable results. Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data, remove noisy data, and redundant information. Finally, imbalance handling techniques such as GAN, and SMOTENC were applied to create more samples for the minority class, helping the dataset to be balanced.

Model training: In this research, team-built prediction algorithms such as LSTM, MLP, and Stacking model built from XGBoost, adaboost, and Logistic Regression algorithms. The algorithms represent a range of algorithms from recurrent neural networks, which are ideal for time series or time series data, to feedforward neural networks that can capture non-linear relationships, as well as techniques that combine multiple machine learning models to improve performance. This study also tweaked some of the parameters of the algorithms to improve the accuracy of the classification problem.

A. Preprocessing model

Before being fed into the main predictive model, the data is passed through a data preprocessing model with the pipeline shown in Algorithm 1. The model optimizes the parameters of the preprocessing techniques in the model and applies those parameters to the incoming test datasets. The model is built with one-hot encoding data processing steps that encode the “gender” column with “Female” 0, “Male” 1

and change the value of the "smoking_history" column into dummy variables or binary 0 and 1. The value 0 is added when the minimum values caused by one-hot encoding appear. Log Transformer is used to transform the data, making the data more suitable for machine learning models. In this problem, the logarithm function is investigated to apply, which is a variant of the natural log function, defined as log(x+1). Adding 1 to the input value helps avoid the case where the input value is 0, because the log of 0 is not defined. MinmaxScaler normalizes the data, bringing the column values to the same range [0,1] to ensure that the contribution of the features to the model is the same. PCA retains 99% of the variance to avoid affecting the model performance due to the small number of features and reduces the possibility of overfitting the model.

Algorithm 1: Pipeline for preprocessing

DATA PREPROCESSING: Prepare data for training and testing

INPUT: Raw data

OUTPUT: Clean data

```
Procedure Preprocess_Data(raw_data):
  Step 1: Begin
    data ← raw_data;
  Step 2: Convert categorical features to numerical
    data["gender"] ← Map 'Male' to 1, 'Female' to 0;
    data←OneHotEncode(data, column=
          "smoking_history", drop_first=True);
  Step 3: Log Transformation
    log_transformer ← Function to apply log(x + 1)
    data ← log_transformer(data);
    Fill missing values with 0;
  Step 4: MinMaxScaler
    If training mode:
      Fit MinMaxScaler on data;
      data ← Transform using fitted scaler;
    Else:
      data ← Transform using previously fitted scaler;
  Step 5: Principal Component Analysis (PCA)
    If training mode:
      Fit PCA (preserve 99% variance);
      data ← Transform using fitted PCA;
 Else:
      data ← Transform using previously fitted PCA;
  Step 6: End
    Return data as clean_data;
End Procedure
```

B. Dealing with data imbalance

To address the common issue of class imbalance in medical datasets, techniques such as SMOTENC, and GAN are employed. This method of balancing processing is only used on the training data set, not applied to the test data set to avoid affecting the evaluation price of the results.

+ *SMOTENC:* SMOTENC which is a technique that combines numerical data and data analysis type through the "penalty point" mechanism, selecting the most common value in the neighborhood. This is a suitable technique for diabetes prediction dataset with tabular data with diverse data

+ *GAN:* The GAN model in this research is built with a simple architecture consisting of two main components: Generator and Discriminator. The Generator is designed to generate fake data with the same dimension as the real data, including the following layers: a Dense layer with 32-neurons using ReLU activation function and random noise vector input, followed by three Dense layers with 64, 128 and input_dim neurons, interspersed with Batch Normalization layers for stabilization and speed-up. The final layer uses a linear activation function to ensure the output has a continuous value. Meanwhile, the Discriminator is configured as a binary classifier consisting of four Dense layers with the number of neurons decreasing from 128 to 32 and using the ReLU activation function, ending with a Dense output layer with 1 neutron using a sigmoid function to predict the accuracy of the real data. The Discriminator model is compiled with a binary_crossentropy loss function and a priority equal to Adam (learning rate 0.0001). The resulting GAN model can be constructed by connecting the Generator and the Discriminator into a sequential model, where the number of Discriminators is frozen to ensure that the Generator has been updated during the joint training phase. GAN helps to expand and diversify minority class data to create more quality data for learning models.

C. Predictive model

+ *Stacking:* The stacking model in this paper is built by combining two machine learning models including XGBClassifier and AdaBoostClassifier in the base model layer, with the parameter random_state = 42 to ensure stability and reproducibility of results. The predictions of these models are based on the model in the last layer (Logistic Regression, random_state=42) to combine and produce prediction results. During the training process, the model applies the 5-fold cross-validation technique (cv=5) to minimize overfitting. In addition, the passthrough=True option allows retaining the original features when training the final model, helping to maximize the information from the original data.

+ *LSTM*: In this paper, the LSTM model is built with two LSTM layers. The first LSTM layer has 128 hidden units, uses the tanh activation function and recurrent activation sigmoid, with a dropout rate of 0.4 and returns the result sequence. The second LSTM layer has 64 hidden units, also uses the tanh activation function and recurrent activation sigmoid, with a dropout rate of 0.4 but does not return the result sequence. Following the two LSTM layers is a Dense layer with 32 neurons using the ReLU activation function, and finally a Dense output layer with 1 neuron using the sigmoid activation function suitable for binary classification problems like this one. The model is compiled with Adam optimizer, learning rate = 0.0001 and uses binary_crossentropy as the loss function for the model.

+ *MLP:* The MLP model is built with three Dense layers, with the number of neurons in each layer decreasing from

64 - 32 -16, using the ReLU activation function and dropout rate 0.3 to reduce overfitting across layers. Similar to the LSTM model, the dense output layer is also built with 1 neuron and sigmoid activation function, compiled with the Adam optimizer, learning rate 0.0001 and the binary_crossentropy loss function.

In the study, we used two data balancing techniques GAN and SMOTENC combined with LSTM, MLP and Stacking algorithms. The selected models and techniques have a variety of processing methods. The study evaluates all combinations to objectively verify the impact of each balancing method on each model in the diabetes prediction problem

IV. Material and Data Processing

The research focused on finding and building a model to predict diabetes with higher accuracy. This study evaluated various classification algorithms and data processing methods to achieve this goal. The dataset is separated to 70% for training and 30% for testing.

A. Dataset Description

In this paper, we used the Diabetes prediction dataset. The dataset is a collection of medical and demographic data from patients and their diabetes's status (positive or negative). It consists of 100,000 patient data with 8 input features and 1 output variable column "diabetes" indicating whether the patient has diabetes or not. The data includes several features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level which can be considered as basic indicators to predict diabetes.

Table 2: Dataset description

	Count	Mean	STD	Min	Max
gender	100000	N/A	N/A	N/A	N/A
age	100000	41.88	22.5	0.08	80.0
hypertension	100000	0.07	0.26	0.00	1.00
heart_disease	100000	0.04	0.19	0.00	1.00
smoking_history	100000	N/A	N/A	N/A	N/A
bmi	100000	27.32	6.63	10.01	95.69
HbA1c_level	100000	5.52	1.07	3.50	9.00
blood_glucose_level	100000	138.05	40.7	80.0	300.0
diabetes	100000	0.08	0.27	0.00	1.00

Based on Table 2, the values in the count column are all at 100 000, indicating that this dataset does not have NULL values. The numerical features show a fairly clear level of data dispersion. The age has an average of 41.88 and a high standard deviation which is 22.51, reflecting the diversity of ages. BMI and blood glucose level also have relatively large standard deviations, indicating that the dataset may contain outliers. The mean value of the diabetes column shows that the proportion of people with the disease in the data is very low, leading to a serious imbalance between the group of people with and without diabetes. Gender and smoking_history are not summarized with mean and standard deviation, as they are categorical value so these variables will need appropriate encoding techniques.

As shown in Fig. 2, the dataset is severely imbalanced with only 8.5% of the infected class (Class 1) being the only one. This results in a biased training model towards the non-infected class, resulting in a low recall for the positive class. The team applied several data balancing techniques to address this issue.

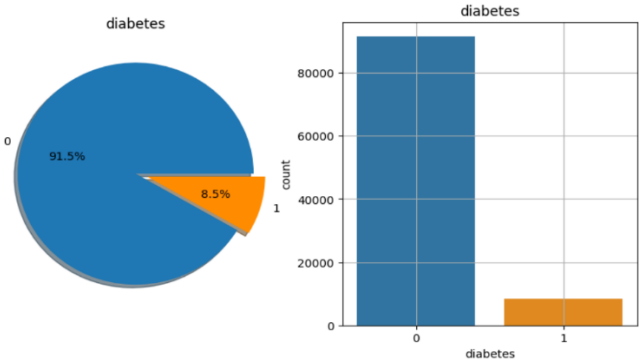

Fig 2: Statistics of the number of negative and positive classes in the column "diabetes"

B. Data Preprocessing

Data plays an important role in building a diabetes prediction model. A clean, complete and well-processed dataset will help improve the accuracy of the model. Before being put into the model, data needs to go through many cleaning steps, each step will have different functions and tasks. Before being put into training, the data needs to be carefully processed through the steps of handling missing values, selecting important features, balancing data, normalizing data and reducing data dimensionality. These steps not only help remove noise and improve data quality, thereby optimizing model performance, helping the model learn effectively and produce better prediction results.

V. Experiment, Evaluation and Discussion

A. Performance evaluation

In the process of developing machine learning models, evaluating the quality of the model plays an important role in determining and selecting the optimal model that best suits the problem. In this study, we evaluated the performance of the models through the following metrics: Root Mean Squared Error (RMSE) to assess data balancing methods, Accuracy and Precision for prediction models. Below is a presentation of the metrics used in this paper to evaluate the models:

RMSE measures the average difference between the predicted values and the actual values:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

Accuracy is calculated as the ratio of correctly predicted samples to the total number of samples:

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN}$$

Precision is the ratio of correctly predicted positive samples to the total number of samples predicted as positive:

$$Precision = \frac{TP}{TP + FP}$$

B. Results

After splitting data for training and testing, the training data was heavily imbalanced with 64,050 samples for Label 0 and only 5,950 samples for Label 1. After applying the methods, GAN generated 60,000 samples for the minority class, helping the class have several samples almost equal to the number of the majority class and achieved the lowest RMSE of 0,19, showing high efficiency in generating new data while still ensuring quality. SMOTENC achieved a perfect balance of 64050 samples for both classes, but the RMSE was still high (0.45).

Table 3: Compare the output and results of the data balancing method

	GAN	SMOTENC
Label 0	64050	64050
Label 1	65950	64050
RMSE	0,19	0,45

In this paper, the models were trained in 100 epoches with a batch size of 32. Through invisible training and testing with data imbalance handling methods such as GAN, SMOTENC combined with deep learning models including LSTM, MLP, STACK based on two main criteria: Precision and Accuracy.

Table 4: The precision and accuracy of prediction results

BALANCE METHOD	PREDICT MODEL	Precision	Accuracy
GAN	LSTM	98.08%	95.37%
	MLP	100%	94.87%
	STACKING	90.22%	96.08%
SMOTENC	LSTM	94.21%	95.19%
	MLP	99.31%	96.29%
	STACKING	66.73%	94.80%

According to table 4, overall, the result on the dataset reaching high performance and accuracy with nearly 95% to over 96%.

The GAN balancing method shows outstanding performance with predictive models. Specifically, the LSTM model achieves Precision 98.08% and Accuracy 95.37% shows that the data generated from GAN still retains the deep hidden time series features consistent with the strengths of LSTM. While the MLP achieves absolute Precision 100%, but Accuracy is only 94.87%. In particular, the Stacking model when combined with GAN achieves the highest Accuracy of 96.08%, although the Precision is only 90.22%. This shows that GAN helps models maintain high Precision and Accuracy, and Stacking can make good use of balanced data from GAN to improve overall accuracy.

With SMOTENC method, the MLP model continues to outperform with Precision 99.31% and Accuracy 96.29%, the highest among models using SMOTENC since technique is suitable for tabular features with multiple data types compatible with the ability to learn nonlinear relationships representing complex relationships between features. The LSTM model achieved Precision 94.21% and Accuracy 95.19%, while Stacking had the lowest Precision (66.73%) but Accuracy still reached 94.80%. Thus, SMOTENC is most suitable for MLP, helping this model achieve optimal prediction efficiency, while stacking when combined with SMOTENC cannot maintain high Precision like the remaining methods.

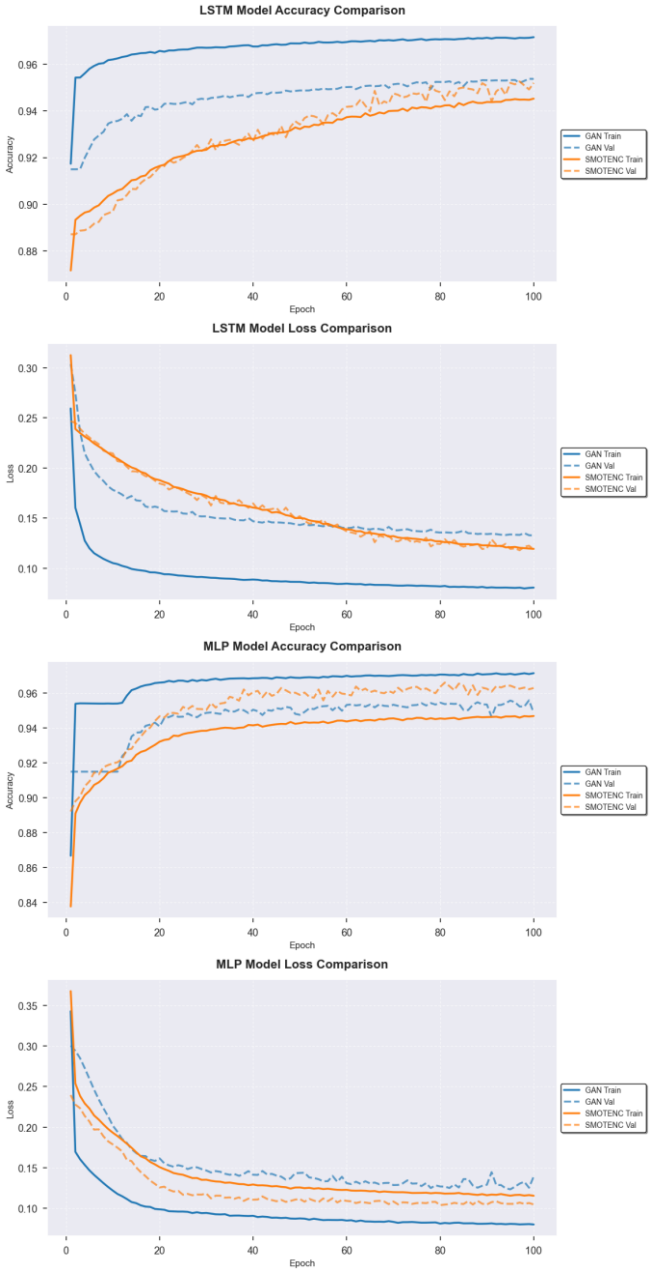

Fig 3: Performance progressing on the training and evaluation tasks

Fig 3 shows that the GAN models perform fine with high accuracy, fast convergence, and good performance throughout the training process, with a steady decrease in loss after each training iteration and a strong correlation between the training and validation sets with the losses of the training and validation sets being almost identical. The SMOTENC models also perform well, with performance gradually increasing to nearly 95%. The GAN models converge quickly after only about 10-15 epochs, the SMOTENC models converge longer than the GAN but still produce results quickly, with a decreasing loss trend. In particular, the LSTM

model combined with SMOTENC is considered the optimal choice, achieving a balance between performance and stability, suitable for deployment in real environments. Therefore, SMOTENC combined with LSTM is the recommended choice if stability and high generalization are the priorities.

VI. CONCLUSIONS

This study focuses on dealing with data imbalance using GAN and SMOTENC techniques combined with prediction algorithms such as LSTM, MLP and Stacking to predict diabetes based on the Diabetets_Prediction_Dataset dataset. The study delves into improving the performance of the model due to the feature imbalance in the disease prediction datasets. The results show that the model performs best when using the MLP algorithm combined with SMOTENC techniques for data augmentation with a performance of 96.29%. Accuracy for the above experiments all achieved high results above 94%, however, the recall and f1-score indices were relatively low. This happened because the imbalance in the dataset led to the difference in the test data set with the majority class being nearly 11 times larger than the minority class (18300 vs 1700). The model tends to favor the majority class, which reduces the ability to accurately detect cases in the minority class (the disease-positive class) - which is an important class in the disease diagnosis problem. When the model prioritizes optimizing overall accuracy, it may ignore cases that appear less frequently but have great clinical significance, such as patients at high risk of diabetes. The study shows great potential when applying advanced machine learning techniques and artificial neural network architectures in diabetes prediction. The model also shows great potential when using data generation techniques in diabetes prediction. GAN and SMOTENC are suitable for the classification problem, the data samples reflect the original data features well. Future studies need to collect diverse and balanced data in addition to improving predictive models to produce better results and avoid overfitting.

ACKNOWLEDGMENT

This work belongs to the project grant No: SV2025 - 13 funded by Ho Chi Minh City University of Technology and Education, Vietnam

REFERENCES

[1] Sharma, A., Singh, P. K., & Chandra, R. (2022). SMOTified-GAN for class imbalanced pattern classification problems. IEEE Access, 10, 30655-30665.

[2] Ratih, I. D., Retnaningsih, S. M., Islahulhaq, I., & Dewi, V. M. (2022, October). Synthetic minority over-sampling technique nominal continous logistic regression for imbalanced data. In AIP Conference Proceedings (Vol. 2668, No. 1). AIP Publishing.

[3] Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. Acta Numerica, 8, 143–195. doi:10.1017/S0962492900002919

[4] Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 9(09), 2278-0181.

[5] Edeh, M. O., Khalaf, O. I., Tavera, C. A., Tayeb, S., Ghouali, S., Abdulsahib, G. M., ... & Louni, A. (2022). A classification algorithm-based hybrid diabetes prediction model. Frontiers in Public Health, 10, 829519.

[6] Xue, J., Min, F., & Ma, F. (2020, November). Research on diabetes prediction method based on machine learning. In Journal of Physics: Conference Series (Vol. 1684, No. 1, p. 012062). IOP Publishing.

[7] Talukder, M. A., Islam, M. M., Uddin, M. A., Kazi, M., Khalid, M., Akhter, A., & Ali Moni, M. (2024). Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications. Digital Health, 10, 20552076241271867.

[8] Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. Procedia Computer Science, 216, 21-30.

[9] Ahmed, S., Kaiser, M. S., Hossain, M. S., & Andersson, K. (2024). A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions. IEEE Access.

[10] Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, 100605.

[11] Shao, H., Liu, X., Zong, D., & Song, Q. (2024). Optimization of diabetes prediction methods based on combinatorial balancing algorithm. Nutrition & Diabetes, 14(1), 63.

[12] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare technology letters, 10(1-2), 1-10.

[13] Zarghani, A. (2024). Comparative Analysis of LSTM Neural Networks and Traditional Machine Learning Models for Predicting Diabetes Patient Readmission. arXiv preprint arXiv:2406.19980.

[14] Alex, S. A., Jhanjhi, N. Z., Humayun, M., Ibrahim, A. O., & Abulfaraj, A. W. (2022). Deep LSTM model for diabetes prediction with class balancing by SMOTE. Electronics, 11(17), 2737.

[15] Srinivasu, P. N., Ahmed, S., Hassaballah, M., & Almusallam, N. (2024). An explainable artificial intelligence software system for predicting diabetes. Heliyon, 10(16).

[16] Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, 220, 106773.

[17] Paliwal, M., & Saraswat, P. (2022, October). Research on Diabetes Prediction Method Based on Machine Learning. In 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 415-419). IEEE.

[18] Natarajan, K., Baskaran, D., & Kamalanathan, S. (2025). An adaptive ensemble feature selection technique for model-agnostic diabetes prediction. Scientific Reports, 15(1), 6907.

[19] Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., ... & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. IEEE Access, 10, 8529-8538.

[20] 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(1), 993098