Nanotechnology Perceptions
660-6795
www.nano-n ISSN 1tp.com

Intelligent Clinical Decision Support For Diabetes Using Ai Technologies

[1*]Ritika Pandey, [1]Jasmine Minj, [1]Pooja Patre, [2]Advait Khare

[1]Vishwavidyalaya Engineering College, Ambikapur, India
[2]Shri Shankara Institute of Professional Management and Technology, Bhilai, India
[*]Address correspondence to: ritikappandey@gmail.com

Diabetes, a chronic metabolic disorder, affects millions worldwide and poses a significant global health challenge due to its rising prevalence, driven by sedentary lifestyles, unhealthy diets, and aging populations. As a leading cause of death, diabetes imposes substantial burdens on healthcare systems and economies, compounded by diagnostic complexity, high patient-to-doctor ratios, and delayed clinical interventions.

To address these challenges, this study proposes an AI-based Clinical Decision Support System (CDSS) leveraging machine learning algorithms for early detection and personalized management of diabetes. Utilizing data from reputable healthcare repositories such as PhysioNet, the methodology involves rigorous pre-processing, including feature selection and handling of missing values, to ensure robustness and reliability. Comparative analysis of 14 machine learning models, facilitated through PyCaret, aims to identify the most effective algorithm for improving diagnostic precision and clinical workflows.

Furthermore, explainable artificial intelligence (XAI) techniques, including LIME and SHAP, are integrated to ensure transparency and trustworthiness in model predictions. By generating local and global explanations, the system provides insights into variables influencing decision-making, enabling clinicians to interpret and act on model outputs effectively. Experimental results demonstrate the potential of merging machine learning with XAI, achieving an accuracy of 86% on test data, and highlighting the strengths and limitations of different interpretable models.

The expected outcomes include enhanced diagnostic accuracy, real-time evidencebased recommendations, and streamlined clinical workflows, leading to reduced complications, optimized resource allocation, and improved patient outcomes. This research underscores the transformative role of AI and XAI in managing complex diseases like diabetes, paving the way for innovative, transparent, and patient-centric healthcare solutions.

1 Introduction

Diabetes is a serious chronic disease caused by either insufficient insulin production or the body's inability to use insulin effectively. This condition leads to high blood glucose levels and is associated with significant health complications, including cardiovascular diseases, kidney damage, and vision loss. According to the World Health Organization (WHO), diabetes is among the leading causes of death globally, and its prevalence continues to rise due to factors such as sedentary lifestyles, unhealthy diets, and ageing populations [1]. The growing burden

of diabetes presents a challenge to healthcare systems worldwide, demanding innovative solutions for effective management.

The management of diabetes is hindered by several factors. First, diagnosing diabetes involves complex clinical evaluations and tests, which can lead to inconsistencies in early detection. Second, the high ratio of patients to healthcare providers, especially in resourcelimited settings, reduces the time and attention available for personalized care. Third, delayed clinical interventions often result in the progression of complications that could have been mitigated with timely action [2]. These challenges underscore the need for advanced tools to improve diabetes detection, monitoring, and treatment.

Machine learning (ML) techniques offer a promising avenue for addressing these challenges. ML has gained popularity in the medical and health sciences due to its ability to analyze large datasets and identify patterns that are not easily discernible through traditional methods [3]. Examples of ML models, such as Support Vector Machines (SVM), Random Forest (RF), and Logistic Regression (LR), have been effectively applied in predicting diseases like diabetes (American Diabetes Association, 2020). Moreover, explainable artificial intelligence (XAI) has emerged as a critical subfield of AI that enhances trust and transparency by elucidating how ML models arrive at their predictions. XAI methods such as Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) allow clinicians to understand and interpret model outputs, fostering confidence in AI-driven decisions [4, 5].

This study aims to develop an AI-based Clinical Decision Support System (CDSS) for diabetes management. The proposed system leverages data from trusted healthcare repositories, such as PhysioNet, to train machine learning models. A total of 14 algorithms, including Random Forest, Support Vector Machines, Gradient Boosting, and Extreme Gradient Boosting, will be evaluated using PyCaret, an open-source machine learning library [6]. The methodology involves pre-processing data to handle missing values, selecting relevant features, and training the models. The models will be assessed and compared to identify the one with the highest predictive accuracy for diabetes diagnosis. Additionally, explainability techniques such as LIME and SHAP will be integrated to ensure that the CDSS provides interpretable and actionable insights [7]. By incorporating XAI, the system aims to address the critical need for transparency and reliability in AI-driven healthcare solutions.

The findings of this research will contribute to improving clinical workflows by integrating real-time, evidence-based recommendations into the decision-making process. This study not only underscores the transformative potential of AI and XAI in managing complex diseases like diabetes but also highlights their utility in enhancing early detection, optimizing resource allocation, and improving patient outcomes.

2 Related work

Diabetes is a pressing issue in both developed and developing nations, significantly impacting public health [8]. Pancreatic dysfunction is a major cause of diabetes, leading to severe health complications such as cardiovascular diseases, renal failure, and neuropathy [9, 10]. Explainable Artificial Intelligence (XAI) techniques like Individual Conditional Expectation (ICE) plots and SHAP have been utilized to analyze medical datasets, demonstrating their utility in elucidating the factors influencing medical insurance premium costs. These insights

benefit decision-makers, insurers, and consumers by aiding in better policy selection [11]. Machine learning (ML)-based Clinical Decision Support Systems (CDSS) have shown the potential to enhance clinical decision-making [12]. However, their adoption requires collaboration between stakeholders and the incorporation of performance evaluations and external validation. In the context of Alzheimer's disease, XAI methods such as LIME, SHAP, GradCAM, and LRP have been employed to classify models into conceptual frameworks, providing insights from local to global interpretations [13, 14].

For chronic kidney disease (CKD) prediction, researchers have applied XGBoost classifiers enhanced with SHAP analysis to identify key biomarkers such as hemoglobin and albumin, improving model interpretability and facilitating clinical understanding [15, 16]. Similarly, LIME and SHAP have been used in another study to visualize the impact of clinical features on CKD prediction models, enhancing transparency and aiding physicians in comprehending model reasoning [17]. Numerous diabetes prediction algorithms have also been developed, employing techniques like Linear Discriminant Analysis, Naive Bayes, Random Forest, and XGBoost. These models have integrated processes such as feature selection, data normalization, and dimensionality reduction to optimize prediction accuracy [18–23]. Comparative analyses have shown that Random Forest often outperforms other models, demonstrating high sensitivity, specificity, and diagnostic accuracy [24].

To better assess diabetes risk, researchers have developed classification models using algorithms like Decision Tree, ANN, and SVM, emphasizing non-invasive and cost-effective detection strategies [25, 26]. Studies have also compared traditional ML approaches with deep learning methods, concluding that Random Forest consistently delivers superior performance in diabetes prediction [24]. Furthermore, ensemble methods, such as WeightedVotingLRRFs, have shown promise in improving predictive accuracy by leveraging multiple supervised ML classifiers [27]. The role of XAI in healthcare extends beyond diabetes. Techniques like LIME and SHAP are widely used to interpret black-box models, providing feature importance scores that enhance model transparency and align with human intuition [13, 14]. For example, OptiLIME offers a trade-off between explanation stability and adherence to the underlying model, optimizing interpretability for practical use [28]. Other frameworks, like ExMed, enable domain experts to execute XAI analytics without extensive programming knowledge, expanding the accessibility of interpretability tools [29].

Applications of XAI in Electronic Health Records (EHRs) have demonstrated the utility of tools like SHAP and LIME in classifying patient data. For instance, SHAP has been employed to highlight clinically relevant features in diabetes and other health conditions, assisting clinicians in decision-making processes [30, 31]. In the fight against pandemics such as COVID-19, XAI methods have proven invaluable. By integrating interpretability techniques, medical professionals can identify critical biomarkers, improve early diagnosis, and enhance treatment strategies, thereby addressing pressing public health challenges effectively [32, 33].

In conclusion, the integration of ML and XAI techniques across various domains, including diabetes, CKD, and Alzheimer's disease, highlights their transformative potential in improving healthcare outcomes. These approaches not only enhance prediction accuracy but also foster trust and understanding among clinicians, paving the way for broader adoption of AI-driven healthcare solutions.

3 Methodology

The proposed approach is depicted in Figure 1 as a model diagram, illustrating the steps undertaken in this study to build and evaluate machine learning models using PyCaret. The methodology involves several stages, including data preprocessing, model training, evaluation, and explainability.

Pre-processing involves transforming raw data into a clean and structured format suitable for machine learning algorithms. The data preparation phase includes tasks such as cleaning, integration, transformation, reduction, and handling missing values. Once preprocessing is completed, the dataset is split into training and testing subsets in an 80:20 ratio to ensure robust evaluation.

Using PyCaret [6], various machine learning models are trained and compared to identify the best-performing algorithm based on metrics such as accuracy, precision, recall, and Area Under the Curve (AUC). PyCaret simplifies the process of model evaluation by automating feature selection, hyperparameter tuning, and performance comparison across multiple models. The binary classification task predicts whether an individual has diabetes (pre-diabetes or diabetes) based on input features.

Explainability is a key focus of this study. To make the model's predictions interpretable, we leverage LIME and SHAP methodologies. These tools provide local and global explanations, helping clinicians understand the factors contributing to each prediction. Such transparency enhances trust and usability in clinical settings. This study discusses the strengths and limitations of LIME and SHAP to guide future researchers in selecting suitable explainability techniques for healthcare applications.

3.1 Dataset

The dataset used in this study is the "diabetes binary health indicators BRFSS2015.csv," containing 253,680 survey responses from the Centers for Disease Control and Prevention (CDC). The target variable, "diabetes binary," is binary, where 0 represents no diabetes or pre-diabetes, and 1 indicates either condition. The dataset includes 21 feature variables,

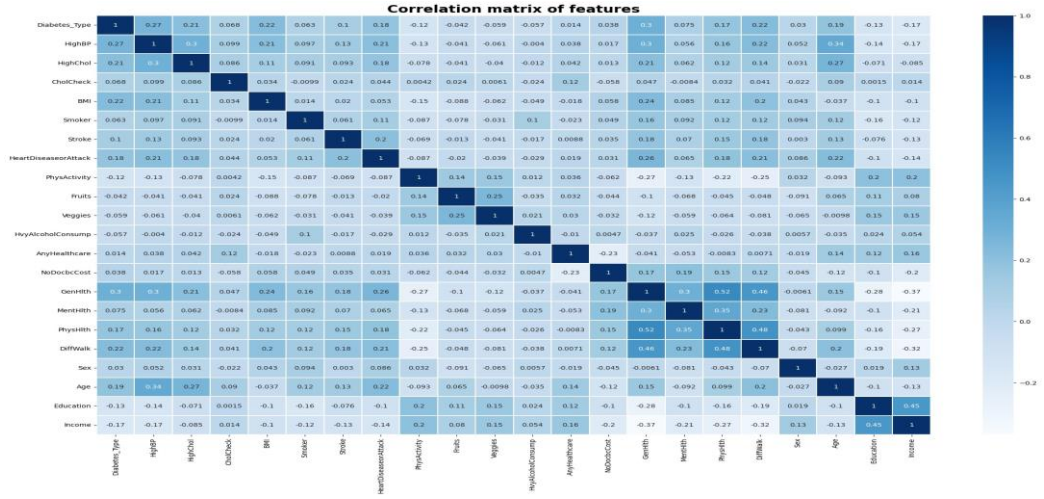


Figure 1: Correlation Heatmap of Diabetes Dataset Features and it is slightly imbalanced.

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual health survey conducted by the CDC, involving over 400,000 participants across the United States. Since its inception in 1984, the BRFSS has collected data on preventive health services, risky behaviors, and chronic health conditions. This dataset provides a rich source of information for predictive modeling.

Early diagnosis of diabetes can significantly benefit individuals by encouraging lifestyle changes, such as weight reduction, healthy eating, regular exercise, and medical consultations. Predictive models for diabetes risk are vital tools for public health experts, as they facilitate early interventions and personalized care.

Figures 1 and 2 illustrate various features and correlations in the diabetes dataset, including a heatmap showing feature interactions. Table 1 provides a detailed description of the dataset attributes, offering insights into the underlying data structure and variables used for model building.

Attributes	Summary
File Name	diabetes binary health indicators BRFSS2015
Description	The dataset includes information from over 400,000 Americans on their use of preventive services, engagement in risky behaviors, and chronic health conditions related to diabetes.
Source of Data	The BRFSS is an annual health-related telephone survey conducted by the CDC. This dataset is publicly available on Kaggle.
Overview	Diabetes is a rapidly spreading global health issue, affecting people across all age groups, including children, teenagers, young adults, and seniors. Its long-term complications can lead to severe outcomes such as organ failure—including the liver, kidneys, heart, and stomach—and may ultimately result in death.
Total Records	253,681 rows
Number of Features	22 columns
Features	Includes variables such as Diabetes binary, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDisease or Attack, Physical Activity, Fruits and Vegetables, Heavy Alcohol Consumption, Any Healthcare, No Doctor's Visit Cost, General Health, Mental Health, Physical Health, DiffWalk, Sex, Age, Education, and Income.
Class	The target variable, Diabetes binary, has two categories: 0 represents no diabetes, while 1 indicates prediabetes or diabetes.

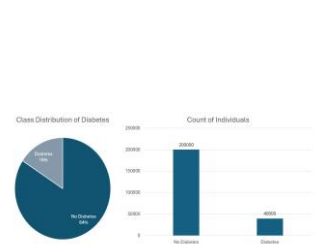
Table 1: Characteristics and Attributes of the Dataset

3.2 Multilayer Perceptron (MLP) ML Model

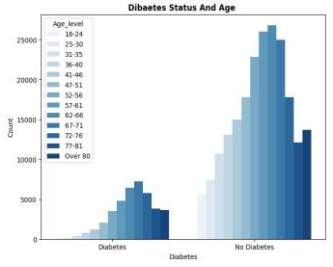
The widely used machine learning Multi-Layer Perceptron (MLP) model is utilized for binary classification issues. This neural network-based supervised learning algorithm leverages multiple layers of interconnected nodes to learn complex patterns in data and make predictions. According to the input features, MLP creates a decision boundary by optimizing weights through backpropagation to divide data points into two groups. The following are the primary steps in developing an MLP model:

- Data Collection: Each instance in our labeled dataset consists of a set of input features and a corresponding binary outcome indicating class membership (0 or 1). The dataset is carefully curated to ensure sufficient representation for both classes.

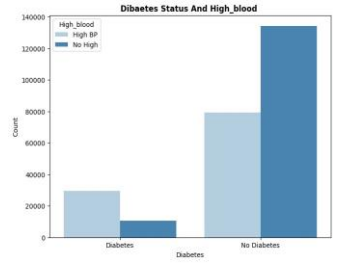
- Data Preparation: Essential preprocessing steps are carried out, including filling in missing values, removing outliers, and applying necessary transformations to ensure



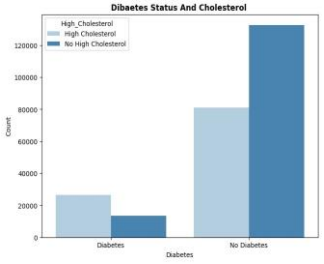
(a)



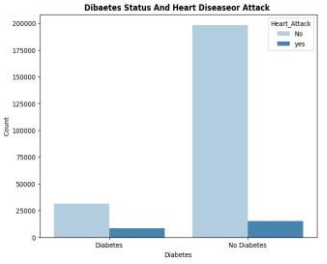
(b)



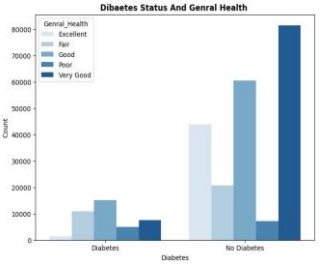
(c)



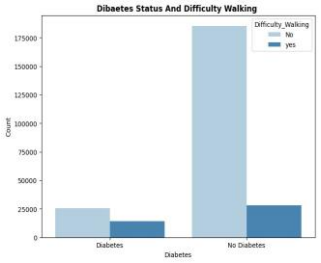
(d)



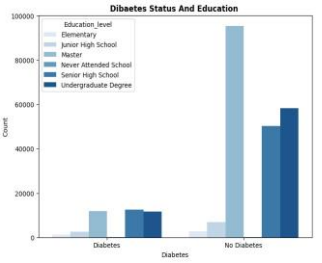
(e)



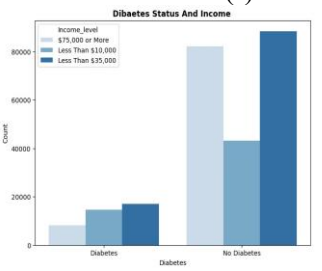
(f)



(g)



(h)



(i)

Figure 2: Feature Plots Representing Dataset Characteristics. data quality and consistency.

- Train-Test Split: The dataset is split into two subsets—80% for training and 20% for testing. The training set is used to build the Multi-Layer Perceptron (MLP) model, while the test set evaluates the model's performance on unseen data.

- Feature Scaling: Input features are scaled to ensure uniformity and facilitate model training. Common methods include standardization (subtracting the mean and dividing by the standard deviation) and normalization (scaling data to a range between 0 and 1).

- Model Training: The MLP model is trained using the training data. The model adjusts its weights through backpropagation by minimizing a cost function, typically binary cross-entropy loss, using optimization algorithms like stochastic gradient descent or Adam.

- Model Evaluation: The trained model is evaluated using the test set. Performance metrics such as accuracy, precision, recall, and F1 score are employed to gauge the model's ability to correctly classify instances in the test set.

- Prediction: After training and evaluation, the model can be used to make predictions on new, unseen data. The MLP employs the activation function in the output layer (e.g., sigmoid for binary classification) to compute the probability of class membership. A predefined decision threshold is then used to assign class labels based on these probabilities.

-

Class	Recall	F1-score	Precision	Support
No Diabetes	0.858	0.986	0.917	34230
Diabetes	0.613	0.122	0.204	6359
Macro avg	0.735	0.554	0.561	40589
Weighted avg	0.820	0.850	0.806	40589
Accuracy		0.8506		

Table 2: Shows the Precision, Recall, F1-score, Support, and Accuracy metrics of the MLP model for diabetes prediction without any resampling.

Class	Recall	F1-score	Precision	Support
No Diabetes	0.947	0.680	0.792	34230
Diabetes	0.316	0.796	0.453	6359
Macro avg	0.632	0.738	0.622	40589
Weighted avg	0.848	0.699	0.739	40589
Accuracy		0.7245		

Table 3: Shows the Precision, Recall, F1-score, Support, and Accuracy metrics of the MLP model for diabetes prediction with downsampling.

In this work, we provide a Multi-Layer Perceptron (MLP) learning-based ML model for diabetes prediction. The MLP model leverages a neural network architecture to predict the probability of a discrete outcome given a set of input variables. The binary output of the MLP model can be either true or false, or one of two other possible classes, based on the decision boundary defined by the model's learned parameters.

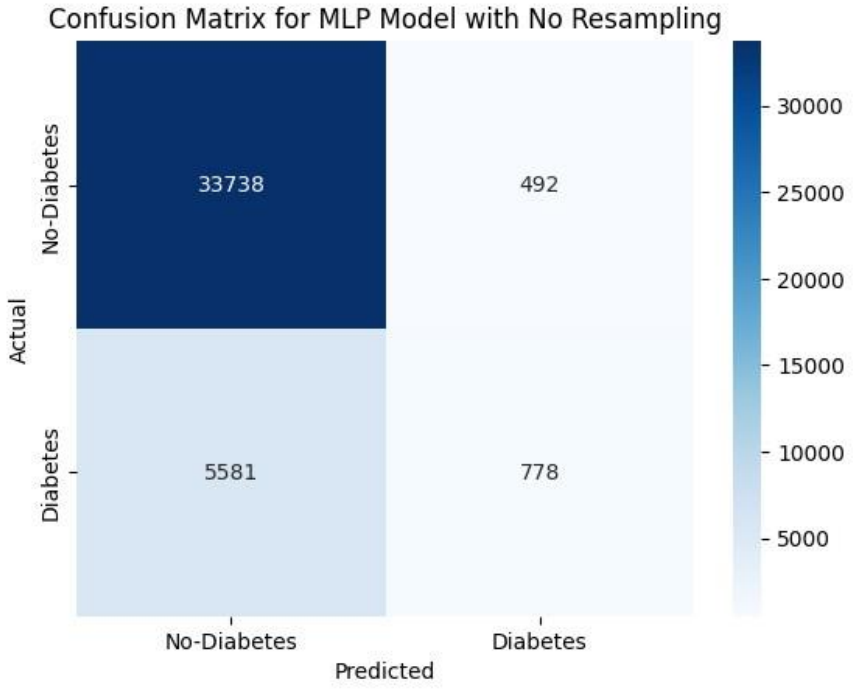
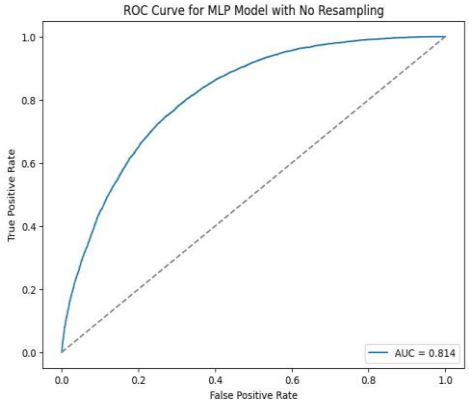
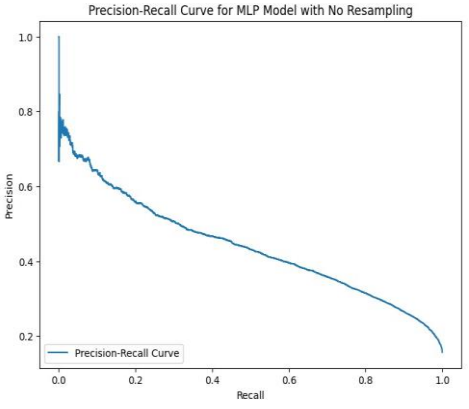


Figure 3: Confusion matrix comparing actual and predicted labels with no sampling.





(a) (b)

Figure 4: (a) Displays the AUC curve, and (b) illustrates the recall and precision curve of the MLP model.

MLP is an advanced technique for solving classification problems, particularly those involving non-linear relationships between features. It can effectively determine whether a new sample belongs to a particular group by utilizing multiple interconnected layers of neurons. MLP models are powerful tools for handling complex datasets, offering robust

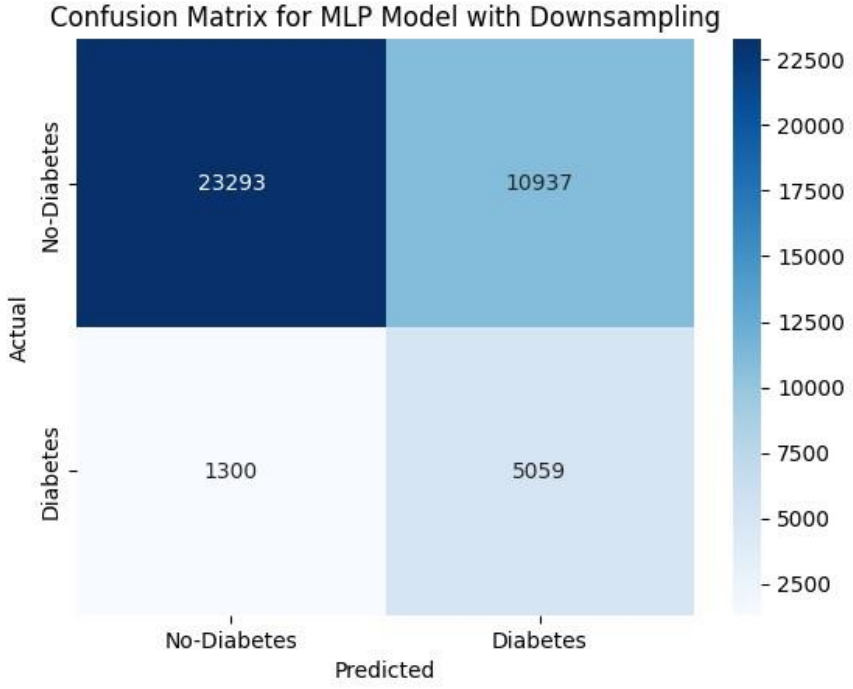
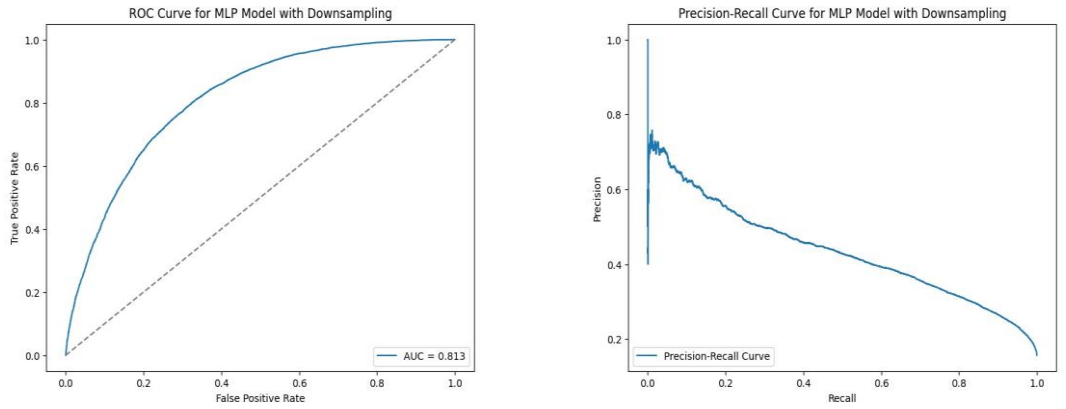


Figure 5: Confusion matrix comparing actual and predicted labels with downsampling.



(a) (b)

Figure 6: (a) Displays the AUC curve, and (b) illustrates the recall and precision curve of the MLP model with downsampling.

performance in binary classification tasks. By using non-linear activation functions and backpropagation for training, MLP overcomes the limitations of simpler models like logistic regression, especially for datasets with non-linearly separable classes.
Table 2 and Table 3 displays the MLP model's Precision, Recall, F1-score, Support,

and Accuracy for predicting diabetes with no resampling and down sampling.

$$P(Y = 1|X) = \text{Softmax}(WX + B) \quad (1)$$

Where:

- P(Y = 1—X) represents the probability of the dependent variable being 1, given the input variables.
- W is the weight matrix of the MLP model. • B is the bias vector applied to the neurons.
- X represents the input features or predictors.
- Softmax is the activation function in the output layer, converting the outputs into probabilities.

Unlike logistic regression, which directly uses the sigmoid function for binary classification, the Multi-Layer Perceptron (MLP) employs multiple layers of neurons to learn complex patterns in the data. Each neuron in the hidden layers applies a non-linear activation function (e.g., ReLU, tanh) to the weighted sum of its inputs, enabling the model to capture non-linear relationships.

The weights (W) and biases (B) are learned during the training process using optimization algorithms such as stochastic gradient descent (SGD) or Adam. The objective is to minimize

a cost function, typically binary cross-entropy for binary classification tasks, to improve the model's predictive accuracy.

The final probabilities are computed using the softmax function (or sigmoid for binary outputs in some configurations) in the output layer, making MLP a flexible and powerful alternative to logistic regression for complex datasets.

Figures 3 and 5 present the data in a matrix format, with the actual classes represented on the Y-axis and the predicted classes on the X-axis. Figures 4 and 6 illustrate the Precision-Recall Curve and ROC AUC score of the MLP model without and with downsampling, which evaluates the rank correlation between predictions and actual targets. This helps demonstrate the model's effectiveness in ranking forecasts, allowing for informed decisions regarding the precision-recall trade-off.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{5}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{FP + TN} \tag{7}$$

AUC is determined as the Area Under the Curve of Sensitivity (True Positive Rate, TPR) versus (1 - Specificity) (False Positive Rate, FPR).

Evaluation metrics are essential for assessing the performance of machine learning models in classification tasks. To evaluate our model, we utilized several metrics, including Precision, Recall, F1-score, Sensitivity, Specificity, and AUC. Using logistic regression, our machine learning model achieved an accuracy of 85% on the diabetes dataset, demonstrating its effectiveness in predicting diabetes in patients.

4 LIME and SHAP Explanation Techniques

The importance of model interpretability in data science cannot be overstated. Gaining insight into a model's inner workings is valuable for various reasons, including building trust in predictions, meeting regulatory requirements, debugging models, and ensuring model safety. Tools like LIME and SHAP play a significant role in enhancing model interpretability across a wide range of machine learning models, including Naive Bayes, Logistic Regression, Linear Regression, Decision Tree, Random Forest, Gradient Boosted Tree, SVM, Neural Networks, and more.

LIME (Local Interpretable Model-Agnostic Explanations) approximates any black-box machine learning model with a local, interpretable surrogate model to explain individual

predictions [13]. This technique can be applied to diverse data types, including images, text, tabular data, and video. By generating explanations near a specific instance of interest, LIME provides localized insights into model behavior. Its versatility makes it a powerful tool for supervised learning models across various machine learning domains. In the field of explainable AI (XAI), LIME is especially notable for its applicability to text, graphical, and tabular data, offering a flexible and extensible approach.

SHAP (SHapley Additive exPlanations), developed by Lundberg and Lee (2017), interprets individual predictions using Shapley values derived from cooperative game theory. Shapley values provide theoretically optimal explanations by calculating the average marginal contribution of each feature value across all possible coalitions of features [14]. SHAP offers a robust framework for understanding the influence of input features on predictions.

InterpretML, an open-source Python library developed by H. Nori et al. [50], integrates multiple machine learning interpretability techniques into a unified package. This library is user-friendly and versatile, enabling the training of glass-box interpretable models that can explain machine learning predictions. InterpretML also provides interactive dashboards that allow users to filter data, form cohorts, and visualize model performance across different dataset variations. Its primary focus is to help users comprehend how models derive their predictions, offering a valuable resource for enhancing transparency and trust in machine learning systems.

4.1 Model Interpretation using LIME

Popular methods for interpreting models in machine learning include LIME, which is particularly valuable for understanding and explaining predictions generated by complex models like Multi-Layer Perceptrons (MLP). LIME provides local explanations by approximating the behavior of the MLP model around a specific instance or prediction. As a modelagnostic approach, LIME can be applied to any machine learning model without requiring insight into its internal workings. By bridging the gap between complex neural networks like MLP and human interpretability, LIME enhances trust and understanding of predictions. However, it is essential to note that LIME is one of many interpretation techniques, and its utility may vary based on the model's characteristics and the specific use case.

The forecast probabilities for the two classes — "0 = No diabetes" and "1 = Have prediabetes or diabetes"—are shown in the figure's leftmost box. The central chart displays the main contributing features with their boundary values, while the actual feature values corresponding to the observation are shown in the rightmost table. This explanation is applied to the dataset's eighth row, focusing on interpreting LIME's forecast for this specific instance. The MLP model, referred to as model mlp, is passed to LIME. LIME uses the predict proba function to analyze the prediction results, allowing it to explain the model's behavior for the selected instance.

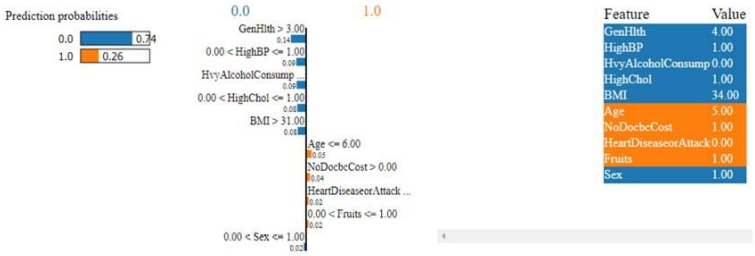


Figure 7: LIME was used to generate an explanation for a prediction involving ten features, interpreting the forecast for the 8th row. The prediction likelihood provided by LIME is as follows: no diabetes with a probability of 0.74, and prediabetes or diabetes with a probability of 0.26.

Finally, the dataset's characteristics and labels are defined, with the number of features set to 10 and the top label probability calculated as 0.74. According to LIME's prediction, the instance has a 0.74 probability of "No diabetes" and a 0.26 probability of "Prediabetes or diabetes," as shown in Figure 7. The rules contributing to the prediction are displayed: negative contributors (left) include "GenHlth > 3.00," "HighBP <= 1.00," and "BMI > 31.0," while positive contributors (right) include "Age <= 6.00," "NoDocbcCost > 0.00," and "Fruits <= 1.00."

Figure 8 demonstrates how to interpret LIME's prediction for the dataset's 10th row using 12 attributes. According to LIME's prediction probabilities, the instance has a 0.76 likelihood of "No diabetes" and a 0.24 likelihood of "Prediabetes or diabetes," showcasing how LIME effectively elucidates the contribution of individual features in the MLP model's predictions.

4.2 Model Interpretation using SHAP

Another popular method for interpreting machine learning models, including Multi-Layer Perceptrons (MLPs), is SHAP (SHapley Additive exPlanations). SHAP assigns importance scores to features based on their contributions to the prediction, providing explanations for specific outcomes. SHAP is grounded in cooperative game theory, specifically Shapley values, which measure each player's contribution in a collaborative setting. This makes SHAP a flexible and model-agnostic interpretability technique that can be applied to various machine learning models, including deep learning architectures like MLPs.

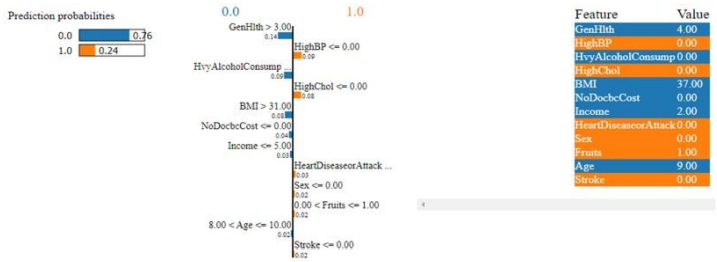


Figure 8: LIME was utilized to explain a prediction involving 12 features, interpreting the forecast for the 10th row. The prediction likelihood provided by LIME is as follows: no diabetes with a probability of 0.76, and prediabetes or diabetes with a probability of 0.24.

The algorithm takes the MLP model and the instance to be explained as inputs. It outputs the SHAP values ϕ for each feature, which represent the contribution of each feature to the prediction for the given instance . The SHAP algorithm is detailed in Algorithm 2.

It is important to note that interpreting complex models like MLPs remains an evolving area of research. Depending on the specific model and dataset, different interpretability techniques, such as LIME and SHAP, may yield varying results.

Figure 9 demonstrates the feature importance derived from SHAP for a trained MLP model predicting diabetes. It shows the average impact of each feature on the model's output. For instance, "HighBP" was identified as the most critical feature, altering the probability of diabetes by an average of 6.4 percentage points (0.064 on the X-axis). "GenHlth" was the second most significant feature, changing the probability by 6.0 percentage points (0.06 on the X-axis). Unlike permutation feature importance, which assesses relevance by measuring the decline in model performance, SHAP's importance is based on the magnitude of feature attributions. While feature importance plots are valuable, they lack additional context about the feature impacts.

Figure 10 presents a summary plot that combines feature relevance and their impacts. Each point on the summary plot corresponds to a Shapley value for an instance and a feature. The x-axis represents the Shapley value, while the y-axis corresponds to the feature. The color of the points indicates the value of the feature, ranging from low to high. Jittering overlapping points along the y-axis reveals the distribution of Shapley values for each feature. Features are displayed in order of relevance.

Figure 11 illustrate force plots that demonstrate how individual features influence the MLP model's prediction for specific observations. These plots provide explanations of

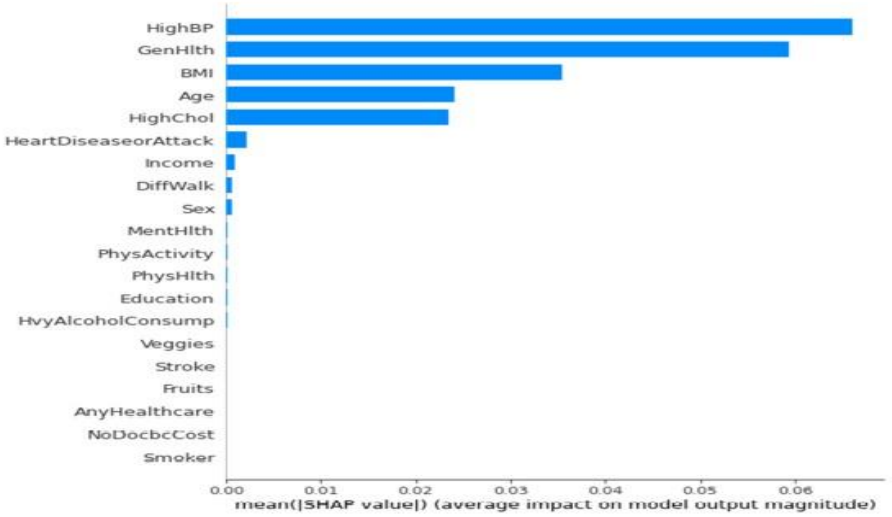


Figure 9: The average absolute Shapley value is used to measure the importance of SHAP features. High blood pressure emerged as the most significant factor, increasing the likelihood of absolute diabetes by an average of 6.4 percentage points (0.064 on the X-axis).

the model's decision-making process for particular instances. The binary target variable "Diabetes binary" has two classifications: 0 for no diabetes and 1 for prediabetes or diabetes. For example, the model's score for a specific observation might be 0.03, indicating a low likelihood of diabetes. Features that increased the score are shown in red, while features that decreased the score are shown in blue. The proximity of a feature to the redblue boundary reflects its impact, and the length of the bar represents the magnitude of its contribution.

The color map in the force plot includes two hues: one for positive SHAP values (e.g., "HighBP" and "GenHlth") and another for negative SHAP values. Various visualization options are available, such as arranging samples by similarity or by output value, allowing for deeper insights into the MLP model's interpretability.

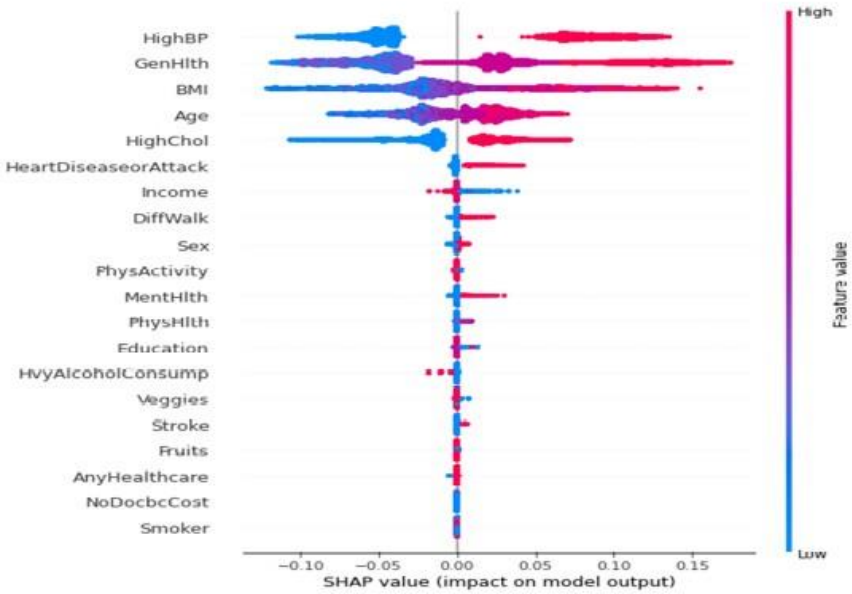


Figure 10: A lower level of high blood pressure reduces the risk of diabetes, whereas higher levels increase the risk. These effects describe the model's behavior and its influence on the output.

4.3 Evaluation Metrics of LIME and SHAP

We have employed several metrics and factors to evaluate the effectiveness of LIME and SHAP as explainability techniques for the Multi-Layer Perceptron (MLP) model. The first metric is fidelity, which measures how accurately the explanation reflects the behavior of the underlying MLP model. For LIME, fidelity can be assessed by comparing predictions made

by the original MLP model with those of the local surrogate model near the instance being explained. Similarly, SHAP provides a unified indicator of explanation quality for all instances.

The second metric is stability, which evaluates the consistency of explanations for similar instances or when the input data is perturbed. A stable explanation technique should produce similar explanations for comparable cases. For LIME, stability can be assessed by measuring the variability in explanations when perturbing the instance under consideration. SHAP, due to its foundation in Shapley values, inherently provides stability in its explanations.
The third metric is consistency, which examines how explanations change in response

Figure 11: The SHAP Force plot highlights the features that most significantly influenced the model's prediction for a single observation. The binary target variable indicates 0 for no diabetes and 1 for prediabetes or diabetes. In this instance, the model score is 0.03.

to modifications in the data or model. Consistency is crucial for understanding the robustness of explanations across different models or datasets. For both LIME and SHAP, consistency can be evaluated by comparing explanations generated from models trained on related datasets or under different configurations.

The fourth metric is comprehensibility, which measures how easily humans can understand and interpret the explanations provided by the technique. Assessing comprehensibility often involves subjective evaluations, such as user research or expert feedback, to gauge the utility and clarity of the explanations produced by LIME and SHAP.

Overall, the performance of LIME and SHAP in explainable artificial intelligence (XAI) is assessed using quantitative criteria like fidelity, stability, and consistency, alongside qualitative evaluations of comprehensibility. When choosing between LIME and SHAP for a specific application, it is essential to weigh the trade-offs between interpretability, computational efficiency, and scalability to make an informed decision.

5 Conclusion

Diabetes is a long-term metabolic disease marked by elevated blood sugar levels (hyperglycemia) brought on by insufficient insulin synthesis or an inefficient use of insulin by the body. It is a global health concern that affects millions of people worldwide. The diagnosis and treatment of diabetes are significantly improved by artificial intelligence and machine learning. Risk Prediction, Early Detection, Image Analysis, Glucose Monitoring, Personalized Treatment, Remote Monitoring, and Support are just a few ways AI and machine learning help with diabetes treatment. It's important to note that while AI and machine learning promise to improve diabetic diagnosis and management, they should complement, rather than replace, healthcare professionals. Medical expertise and human judgment are crucial for interpreting

results and making informed decisions. In summary, multilayer perceptron is an excellent machinelearning technique that successfully forecasts the course of diabetes. By combining this algorithm with interpretable models such as LIME and SHAP, we can gain valuable insights into the factors driving the predictions and increase the transparency and trustworthiness of the model. In addition to achieving precise predictions, using LIME and SHAP interpreters in conjunction with logistic regression provides insightful information about the underlying connections between characteristics and the target variable. This combination of accuracy and interpretability is crucial in healthcare, where understanding the reasoning behind predictions is paramount for medical professionals and patients alike.

References

1. Saeedi P, Salpea P, Karuranga S, et al. Mortality attributable to diabetes in 20–79 years old adults, 2019 estimates: Results from the International Diabetes Federation Diabetes Atlas. Diabetes research and clinical practice 2020;162:108086.
2. Paul SK, Klein K, Thorsted BL, Wolden ML, and Khunti K. Delay in treatment intensification increases the risks of cardiovascular events in patients with type 2 diabetes. Cardiovascular diabetology 2015;14:1–10.
3. Alanazi A. Using machine learning for healthcare challenges and opportunities. Informatics in Medicine Unlocked 2022;30:100924.
4. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, and Acharya UR. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine 2022;226:107161.
5. Alsaleh MM, Allery F, Choi JW, et al. Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: A systematic review. International Journal of Medical Informatics 2023;175:105088.
6. Ali M. PyCaret: An open source, low-code machine learning library in Python. PyCaret version 1.0. 2020. URL: https://www.pycaret.org.
7. Salih AM, Raisi-Estabragh Z, Galazzo IB, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems 2024:2400304.
8. Goyal Y, Verma AK, Bhatt D, Rahmani AH, Dev K, et al. Diabetes: perspective and challenges in modern era. Gene Reports 2020;20:100759.
9. Burcelin R, Knauf C, and Cani PD. Pancreatic α-cell dysfunction in diabetes. Diabetes & metabolism 2008;34:S49–S55.
10. Dludla PV, Mabhida SE, Ziqubu K, et al. Pancreatic β-cell dysfunction in type 2 diabetes: Implications of inflammation and oxidative stress. World journal of diabetes 2023;14:130.
11. Orji U and Ukwandu E. Machine learning for an explainable cost prediction of medical insurance. Machine Learning with Applications 2024;15:100516.
12. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, and Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine 2020;3:17.
13. Viswan V, Shaffi N, Mahmud M, Subramanian K, and Hajamohideen F. Explainable artificial intelligence in Alzheimer's disease classification: A systematic review. Cognitive Computation 2024;16:1–44.
14. Sudar KM, Nagaraj P, Nithisaa S, Aishwarya R, Aakash M, and Lakshmi SI. Alzheimer's Disease Analysis using Explainable Artificial Intelligence (XAI). In: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). IEEE. 2022:419–23.

15. Raihan MJ, Khan MAM, Kee SH, and Nahid AA. Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. Scientific Reports 2023;13:6263.
16. Ghosh SK and Khandoker AH. Investigation on explainable machine learning models to predict chronic kidney diseases. Scientific Reports 2024;14:3687.
17. Reddy S, Roy S, Choy KW, et al. Predicting chronic kidney disease progression using small pathology datasets and explainable machine learning models. Computer Methods and Programs in Biomedicine Update 2024;6:100160.
18. Maniruzzaman M, Rahman MJ, Al-MehediHasan M, et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. Journal of medical systems 2018;42:1–17.
19. Reinhardt A and Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic acids research 1998;26:2230–6.
20. Kegl B. The return of AdaBoost. MH: multi-class Hamming trees. arXiv preprint´ arXiv:1312.6086 2013.
21. Tabaei BP and Herman WH. A multivariate logistic regression equation to screen for diabetes: development and validation. Diabetes Care 2002;25:1999–2003.
22. Jenhani I, Amor NB, and Elouedi Z. Decision trees as possibilistic classifiers. International journal of approximate reasoning 2008;48:784–807.
23. Hasan MK, Alam MA, Das D, Hossain E, and Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 2020;8:76516–31.
24. Yahyaoui A, Jamil A, Rasheed J, and Yesiltepe M. A decision support system for diabetes prediction using machine learning and deep learning techniques. In: 2019 1st International informatics and software engineering conference (UBMYK). IEEE. 2019:1–4.
25. Ahmed S, Kaiser MS, Hossain MS, and Andersson K. A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions. IEEE Access 2024.
26. Priyadarshini A and Aravinth J. Correlation Based Breast Cancer Detection using Machine Learning. In: 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT). IEEE. 2021:499–504.
27. Fazakis N, Kocsis O, Dritsas E, Alexiou S, Fakotakis N, and Moustakas K. Machine learning tools for long-term type 2 diabetes risk prediction. ieee Access 2021;9:103737– 57.
28. Visani G, Bagli E, and Chesani F. Optilime: Optimized lime explanations for diagnostic computer algorithms. arXiv preprint arXiv:2006.05714 2020.
29. Kapcia M, Eshkiki H, Duell J, Fan X, Zhou S, and Mora B. Exmed: An ai tool for experimenting explainable ai techniques on medical data analytics. In: 2021 IEEE 33rd international conference on tools with artificial intelligence (ICTAI). IEEE. 2021:841–5.
30. Ong JH, Goh KM, and Lim LL. Comparative analysis of explainable artificial intelligence for COVID-19 diagnosis on CXR image. In: 2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). IEEE. 2021:185–90.
31. Hu CA, Chen CM, Fang YC, et al. Using a machine learning approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan. BMJ open 2020;10:e033898.
32. Ye Q, Xia J, and Yang G. Explainable AI for COVID-19 CT classifiers: an initial comparison study. In: 2021 IEEE 34th international symposium on computer-based medical systems (CBMS). IEEE. 2021:521–6.
33. Wu H, Ruan W, Wang J, et al. Interpretable machine learning for covid-19: An empirical study on severity prediction task. IEEE Transactions on Artificial Intelligence 2021;4:764–77.