Received 16 March 2024, accepted 1 April 2024, date of publication 8 April 2024, date of current version 19 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3386220

RESEARCH ARTICLE

An Improved Concatenation of Deep Learning Models for Predicting and Interpreting Ischemic Stroke

SAPIAH SAKRI[1], (Member, IEEE), SHAKILA BASHEER[1], ZUHAIRA MUHAMMAD ZAIN[1], NURUL HALIMATUL ASMAK ISMAIL[2], DUA' ABDELLATEF NASSAR[1], GHADAH NASSER ALDEHIM[1], AND MAIS AYMAN ALHARAKI[1]

[1]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[2]Department of Computer Science and Information Technology, Applied College, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding author: Nurul Halimatul Asmak Ismail (NHIsmail@pnu.edu.sa)

This work was supported in part by the Deputyship for Research and Innovation, Ministry of Education, Saudi Arabia, under Project RI-44-0213.

ABSTRACT Early detection of stroke warning symptoms can help reduce the severity of ischemic stroke, the leading cause of mortality and disability worldwide. This study aims to develop a model to predict the disease by leveraging machine learning-based models. A model that concatenates a convolutional neural network and a long short-term memory was developed as the proposed model. Seven other classifiers were treated as the baseline models: logistic regression, random forest, extreme gradient boosting, k-nearest neighbor, artificial neural network, long short-term memory, and convolutional neural network. All models were trained using a healthcare dataset of 5110 patients' health profiles. A synthetic minority oversampling technique was deployed to balance the data. Metrics such as accuracy, precision, F1-score, recall, area under the curve, and confusion metrics were used to evaluate the models' performance. With a 95.9% accuracy, the proposed model outperformed the models employed in this study and improved the accuracy of prior studies that used the same dataset. The Shapley Additive Explanations method was applied to explain the result obtained by the best model. The proposed model was created to predict ischemic stroke. It considers each patient's profile, allowing for personalized decision-making in resource-constrained settings.

INDEX TERMS Ischemic stroke prediction, SHAP method, hybrid deep learning model, machine learning.

I. INTRODUCTION

Stroke, also known as a cerebrovascular accident, is a significant global health problem and one of the leading global causes of death. The World Health Organization (WHO) [1] defines stroke as a severe, worldwide disruption or malfunction of the blood arteries supplying the brain that leads to limb paralysis, severe morbidity, and coma. Symptoms of a stroke may persist for more than 24 hours and often end in mortality within 3 to 10 hours [2]. Annually, fifteen million individuals have a stroke, and one person dies every four to five minutes [3]. Ischemic and hemorrhagic strokes are the two most common types of strokes [4]. According to the American Heart Association [5], ischemic strokes (IS) occur when a blood clot obstructs or stops a brain-supplying blood vessel.

In contrast, 15% of strokes [5] are hemorrhagic, which occurs when a weakened blood vessel ruptures or bleeds. In the previous five decades, the incidence of stroke-related mortality in developing countries has been ten times greater than in the West [6]. The United States has seen permanent disability as a consequence of this disease. The World Health Organization [7] states stroke has the 84th-highest estimated global mortality rate. Each year, this sickness affects around 700,000 people. In the past several decades, numerous research methods have identified nonmodifiable

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

risk indicators (genetic factors, male gender, older age) and modifiable risk markers (hypertension, cigarette smoking, diabetes mellitus) [8]. Figure 1 depicts risk variables associated with strokes. With an increasing worldwide population, the disease's mortality rate and the number of people it affects rise. However, early treatment and diagnosis may decrease this mortality rate.

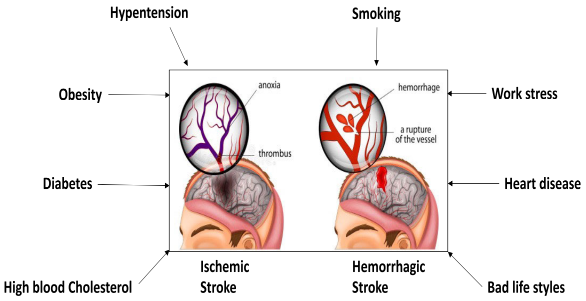


FIGURE 1. Risk factors for stroke [8].

Previously, conventional statistical techniques for developing prediction models, including "Cox proportional hazards regression analysis" and "logistic regression," were employed. However, due to the data type limitations inherent to these methods, these approaches have a limited impact on data mining when dealing with clinical data, which has imbalanced and high-dimensionality issues. With the advent of powerful computer technology and competent medical care, machine learning (ML) and deep learning (DL) have been used to predict sickness risk. The capacity of a computer to deduce patterns from data regularities is crucial to the development of artificial intelligence [5], [6]. Previous research reveals that ML and DL algorithms beat traditional statistical methods in predicting stroke with more accuracy and lower costs [7] due to their capacity to produce predictions based on large datasets. Thus, it reflects the inferiority of traditional approaches compared to prediction models constructed using ML and DL algorithms [8]. DL may outperform conventional ML algorithms (without needing feature selection) if the hyperparameters are well-controlled.

In this study, the performance of the proposed hybrid DL model (a convolutional neural network (CNN) and a long short-term memory (LSTM)) was compared to the baseline models (other ML and DL models) and the previous studies that used similar datasets. The Shapley Additive Explanations (SHAP) method explains the features revealed by the best-performing model.

A. CONTRIBUTION

Based on the gaps identified from the review of the related work in section II, below are the following key contributions of this study:

- A novel and robust hybrid DL model (CNN+LSTM) is proposed.
- Synthetic minority oversampling technique (SMOTE) was deployed to balance the imbalanced dataset.
- The SHAP method is used to discover which model attributes are most important.
- The experimental results compare the proposed model's performance to the baseline and models from previous studies.

The rest of the paper is organized into four sections, where Section II represents related work. The proposed materials and method of this paper for predicting stroke disease are explained in Section III. The result and analysis are discussed in detail in Section IV. Section V highlights some of the study's limitations. Finally, Section VI concludes this study.

II. RELATED WORK

This section reviews recently published studies on ischemic stroke prediction from 2017 onwards from various Internet sources. This study compares the deployed dataset, data preprocessing or resampling method, deployed classifiers, and the performance metrics attained by the study.

A. AN OVERVIEW OF PREVIOUS PREDICTION OF ISCHEMIC STROKE APPROACHES

Numerous ischemic stroke prediction models based on regression or other statistical methods have been created thus far. Nevertheless, because these models by [9] only incorporate small factors, their clinical value is sometimes severely limited. A study by [10] found that multivariable logistic models performed well, with an area under the receiver operating characteristic (AUROC) curve of 0.71-0.74. These models were developed using clinical and retinal characteristics (20 variables) based on 332 patients. Utilizing ML and DL approaches, along with large amounts of real-world patient-level data from EHR, can increase the number of features captured, allowing for the construction of more accurate prediction models [11].

B. STATE-OF-THE-ART PREDICTION OF ISCHEMIC STROKE APPROACHES

In 2017, [12] gathered the "Cardiovascular Health Study" (CHS) datasets. Three datasets were created, containing 212 strokes and non-strokes. The completed collection has 357 properties and 1,824 entities, with 212 stroke occurrences. The suggested method employs C4.5 decision tree methods for feature selection and principal component analysis (PCA) for dimension reduction. Following the reduction, a classification model was developed using Artificial Neural Network (ANN), resulting in an accuracy classification model of 94.7%.

Stroke risk diagnosis is a laborious and intricate process, as explained by [13]. The biomedical examination uncovered six features linked to the significant risk factors. In addition, they suggested a novel feature selection approach that combines support vector machines (SVM) with "glow-worm swarm optimization" and is based on the standard deviation of features. The accuracy that the suggested model produced was 82.58%. This model will be considered in the research as it improved the accuracy of the described unique technique.

According to [14], the 2019 study aimed to test ML-based modeling methodologies. In this work, they developed the concept of recognizing the type of stroke, hemorrhagic or ischemic, and predicting the condition's future repercussions. In conjunction with monitoring technologies, they can identify the kind of stroke within minutes following the emergency. Both the a) stroke diagnosis and the b) death prediction datasets had seven predictors and two objective variables, respectively. A total of seven algorithms were examined and assessed. The model with the best performance was the Random Forest (RF) one, with average values of 0.93±0.03.

In 2019, [15] suggested and implemented five stroke prediction approaches: SVM, ANN, PCA+ANN, DT+ANN, and DT+PCA+ANN. Feature selection was done using only the C4.5 and the Decision Tree (DT) model. The PCA algorithm was used to reduce dimensionality, improving accuracy while decreasing run time. Classifiers such as ANN and SVM were used as the baseline. Finally, among the several approaches used, the DT, PCA, and ANN composite approach produced the best results.

According to [16], the DNN model had a considerably higher AUC than the ASTRAL (statistical method) score (0.888 against 0.839; $P<0.001$); however, neither the RF (0.857; $P = 0.136$) nor the LR models' areas under the curves were significantly greater than the ASTRAL score. Not even when limited to just the six components used by the ASTRAL score did the ML models' performance deviate significantly from the ASTRAL score,

Conducting an early brain stroke prediction in 2020 [17] could generate additional data. Several machine learning approaches were applied in this investigation, such as ST, MT, CT, LR, LSVM, QSVM, and ANN classifiers. Lastly, the ANN model finished with the highest score of 95.3%.

RLR, SVM, and RF were the ML techniques supplied by [18] in 2020. The RF model successfully achieved a maximum accuracy of 78 percent. Due to the fact that the data source was quite unequal, ROS, RUS, and SMOTE were applied to bring it into balance.

In 2021, [19] built five distinct models for reliable prediction using machine learning techniques such as LR, DT, RF, kNN, SVM, and NB Classification. These algorithms were utilized to take into account a variety of physiological parameters. It was determined that the algorithm that performed this work the most effectively was NB, which achieved an accuracy of roughly 82 percent.

In 2021, [20] created ML models to predict the likelihood of a stroke in the brain. This study employs a variety of physiological metrics and machine learning methods, including LR, DT, RF, and Voting Classifier (VC), to train four distinct models for accurate prediction. RF was the best-performing algorithm for this task, with an accuracy of around 95.7%.

In 2021, [21] began creating algorithms to identify high-risk patients for targeted therapies, as well as improving predictors of 30-day readmission following an ischemic stroke. They utilized patient-level information from electronic health records (EHR). The five ML algorithms are RF, GBM, XGBoost, SVM, and LR. The methods that were utilized included data-driven feature selection and adaptive sampling. Out of all the tested algorithms, XGBoost with ROSE sampling had the highest area under the curve (AUC), whereas LR with ROSE sampling and feature selection achieved the highest sensitivity.

In 2022, [22] developed the model, and a ''brain stroke dataset'' was utilized. Data can be standardized through the use of a procedure called standardization. RF, SVM, and DT classifiers are employed simultaneously in training and testing rounds. Accuracy, sensitivity (SEN), error rate, false-positive rate (FPR), false-negative rate (FNR), root mean square error, and log loss were some of the characteristics that were utilized to assess the performance of each classifier being tested. The findings indicated that the RF classifier achieved a maximum accuracy of 95.30%.

In 2022, [23] aims to enhance stroke prediction by conducting an in-depth analysis of the multiple components that make up electronic health records (EHR). They use statistical and principal component analysis methods to identify the most critical stroke prediction factors. The RF, SVM, and DT models were deployed so that they could be trained on the EHR. It was determined that RF was the model that performed the best, with an accuracy of 95.3 percent.

Table 1 presents the summary and the accomplishments of previous studies reviewed in this section.

III. MATERIALS AND METHODS

This study follows a systematic research methodology, as shown in Figure 2, which encompasses six main steps. By complying with this method, researchers and healthcare practitioners can develop robust and clinically relevant stroke prediction methods that improve patient outcomes and reduce the burden of stroke-related morbidity and mortality. The steps include data acquisition (as explained in sub-section A), exploratory data analysis (as elaborated in sub-section B), data preprocessing (as discussed in sub-section C), model classification (as analyzed in sub-section D), model evaluation (as clarified in sub-section E), and model interpretation (as described in sub-section F).

A. DATASET AND ACQUISITION

This study deployed the healthcare stroke dataset [24] to train the proposed and baseline models. Each row of the data set contains the patient's vital statistics. The data aims to assess historical estimations and forecast whether or not the patient will have a stroke. No personally identifying information, such as a patient's name, address, or Social Security number, is included in the dataset. Therefore, there is no risk that the dataset used in the experiment would endanger patient privacy. Here, the main significant characteristics of the data set are summarized.

TABLE 1. A review summary of previous studies' accomplishments.

Ref.	Authors (Year)	Data Preprocessing	Deployed Classifiers	Accuracy/AUC Performance
[12]	Singh & Choudary (2017)	PCA for dimensionality reduction, DT for feature selection	DT, ANN	ANN = 94.7%
[13]	Zhang et al. (2018)	Developed a novel feature selection method	SVM+ glow-worm swarm optimization techniques	82.5%
[14]	Garcia Terriza et al. (2019)	Using z to normalize the data	RF	93%
[15]	Singh et al. (2019)	DT and C4.5 for feature selection, PCA for dimensionality reduction	PCA+ANN, ANN, SVM, DT+ANN, PCA+ANN+DT	ANN + PCA = 95.2%
[16]	Heo et al. (2019)	Missing value handling	LR, DNN, RF and ASTRAL (statistical method)	DNN (AUC = 0.88)
[17]	Govindarajan et al. (2020)	Using tagging and maximum entropy methodologies for feature selection	ST, MT, CT, LR, LSVM, QSVM, and ANN	ANN = 95.3%
[18]	Wu et al. (2020)	ROS, RUS, and SMOTE (balancing)	RLR, SVM, RF	RF =78%
[19]	Sailasya et al. (2021)	Missing value handling, label encoding, and imbalanced data using ROS	LR, DT, RF, kNN, SVM and NB	NB = 82%
[20]	Tazin et al. (2021)	Utilized SMOTE for data balancing	LR, DT, RF, and VC	RF = 95.7%
[21]	Darabi et al. (2021)	RStudio Baruta for feature selection, ROSE (Random over-sampling) is used to balance the dataset.	RF, RB, XGBoost, SVM, LR,	LR+ROSE = 64%
[22]	Akter et al. (2022)	Train-Test Split Method	RF, SVM, DT	RF = 95.3%
[23]	Dev et al. (2022)	Feature selection using PCA	DT, RF, ANN	ANN with PCA = 75%

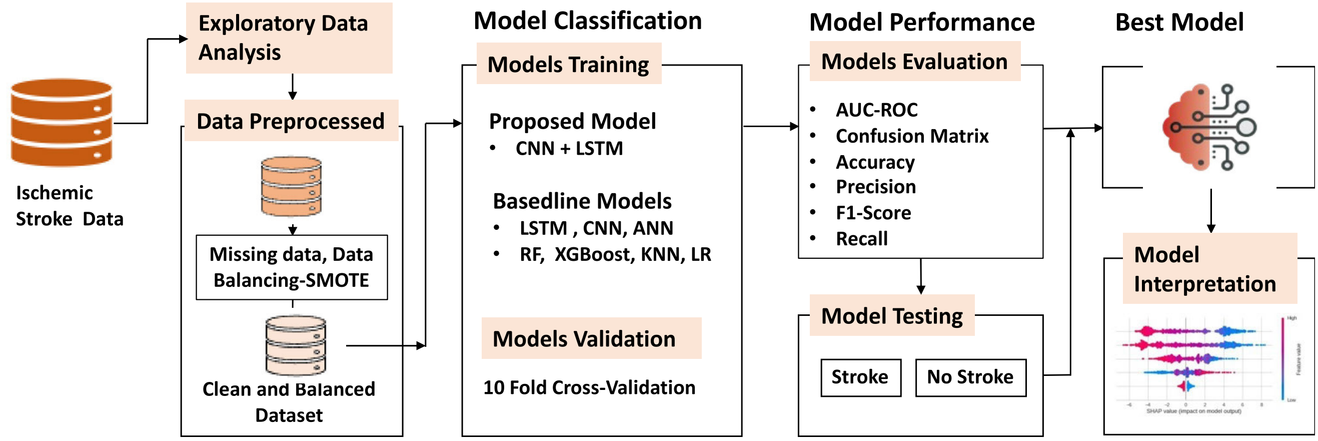


FIGURE 2. Proposed methodology of Ischemic stroke prediction.

- The dataset has 5110 samples.
- Each sample has 11 features and one target value.
- The target includes “0”’, which refers to stroke cases, and “1”, which refers to no stroke cases.
- Between the classes, 249 samples were with stroke, and 4861 samples did not.
- Handle missing data and removing “unknown” samples. The cleaned dataset has 179 stroke cases and 3,247 non-stroke cases for 3,426 samples.

A short description of those features is defined in Table 2.

B. EXPLORATORY DATA ANALYSIS (EDA)

EDA presents the characteristics of the deployed dataset. Figure 3 illustrates the distribution plot of the two numerical features. Figure 4 shows the scatter plot of six selected categorical features. Figure 5(a) shows the correlation between the features. Figure 5(b) illustrates the FreeViz plot of all

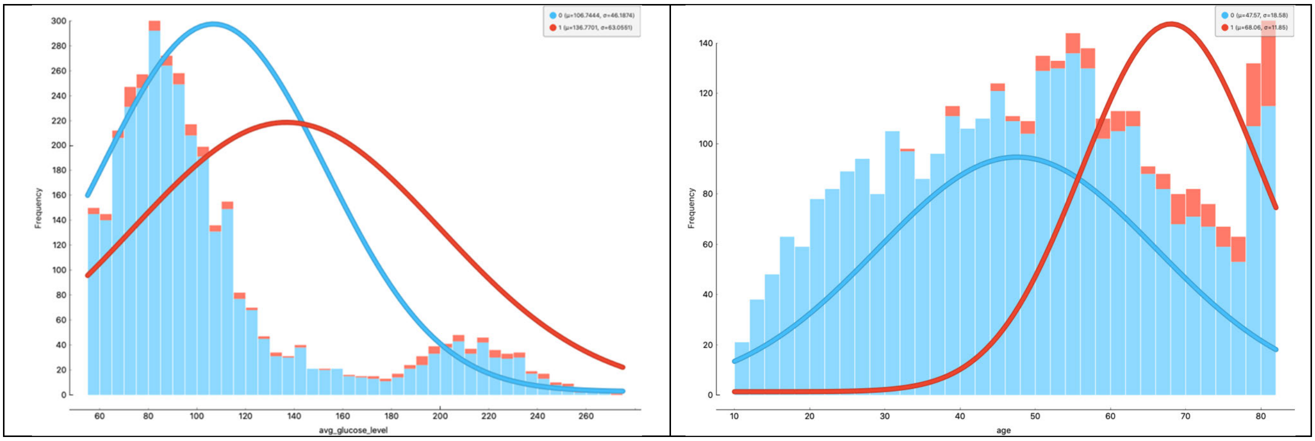


FIGURE 3. Distribution analysis of the "avg_glucose_level" and "age" variables.

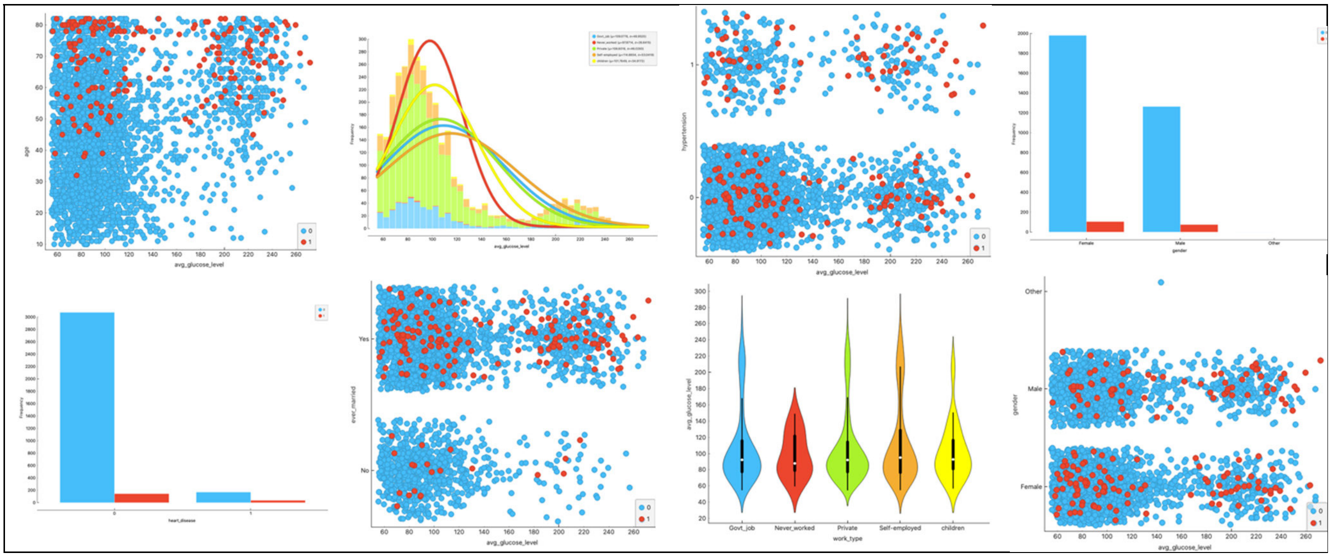


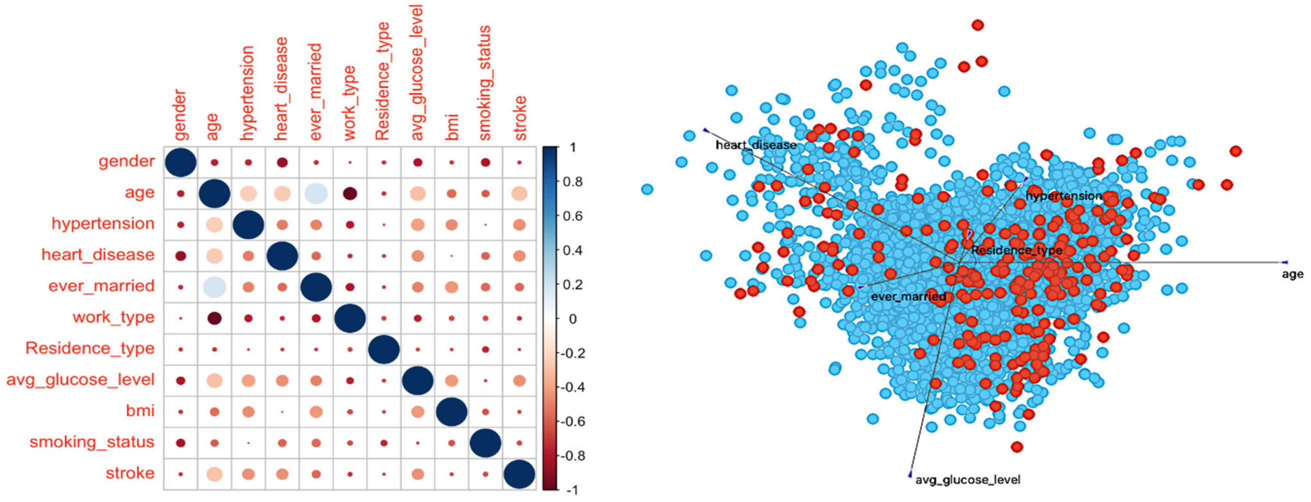
FIGURE 4. The characteristics of all variables.

variables. The red color denotes the occurrence of a stroke, and the blue color denotes non-stroke occurrence. Figure 3 and Figure 4 indicate that the data distribution is not balanced between the stroke and non-stroke cases. The data reported that 95% contained 'no stroke' data and only 5% with stroke occurrence.

In Figure 5(a), it was observed that "age" and "stroke" are strongly positively correlated. "avg_glucose_level," "ever_married," "heart_disease," and "hypertension" are also associated with stroke. At the same time, the "residence_type," "bmi," and "gender" seem not to be correlated to stroke. The data shows that "age" and "avg_glucose_level" correlate with stroke. Further evaluation of the relationship between stroke and all the variables is shown in Figure 5(b). The general overview from this plot indicates that the variables containing stroke data were much less than those without stroke. The concentration of stroke data was found between the "age" and "avg_glucose_level" variables quadron.

C. DATA PREPROCESSING

Data preparation is essential to address the issue of missing, noisy, and unknown data, which may degrade the experimental data quality and thus affect prediction accuracy. SMOTE [25] was used to adjust for the imbalance between the stroke and non-stroke classes. The output of the balancing is shown in Figure 6. The underrepresented sample was oversampled to guarantee accurate representation. Since there were no missing values, neither deletion nor imputation of data was undertaken. In ML, "hyperparameters" are parameters with predetermined values [26]. It influences how machine learning models behave. Without it, model failure is more likely to occur. "GridSearchCv," an implementation



(a) Correlation analysis of all the variables

(b) FreeViz analysis of all the variables

FIGURE 5. Correlation and FreeViz analysis of all the variables.

TABLE 2. Dataset feature description.

Feature	Description
ID	Unique identifier for a person
Gender	Male, Female, others
Age	Age of the patient
Hypertension	0 (zero) = no hypertension, 1 (one) = has hypertension
Heart disease	0 (zero) = no heart disease, 1 (one) = has heart disease
Ever married	Patient's marital status
Work Type	Patient's work type
Residence type	Patient's residence type
Avg. glucose level	The average glucose level in the blood
BMI	Patient's body mass index
Smoking status	Patient's smoking status: formerly smoking/never smoked/ smoked
Stroke (Target)	0 (zero) = no stroke, 1 (one) = has stroke

of grid search with cross-validation, is used in this study to fine-tune hyperparameters. This method evaluates the model for each conceivable combination of the dictionary's input values. Consequently, the ideal model is selected, and the highest level of precision is achieved for all hyperparameter values.

D. MODEL CLASSIFICATION

1) THE PROPOSED MODEL (CNN+LSTM)

We proposed a model that concatenates CNN with LSTM to be trained using the source dataset in predicting ischemic stroke. This model (Figure 7) harnesses both the excellent capabilities of CNN and LSTM models. During the implementation, the CNN layer will collect ischemic stroke data as input to learn and extract the most significant properties from the data and use 'ReLu' as the activation function. This adds non-linearity to the network and enables the model to learn more rapidly and effectively. Each CNN layer generates a feature map that activates a separate kernel by sliding it over the stroke data. CNN then convoluted the extracted features and compacted the stroke data for the classification tasks. Then, this data is transferred to the LSTM layer, which uses past and future knowledge to enable more exact classification and more accurate prediction. Thus, it created a fully connected neural network to learn long-term dependencies. 'ReLu' is then used as an activation function to categorize two classes of patients into stroke and no stroke. Later, to keep the model from overfitting, a dropout layer of 0.5 is applied; also, an attention layer is employed to give weights to the significant features while disregarding the unimportant ones. Finally, a sigmoid activation function is applied to the dense layer. Then, using binary cross-entropy as the loss function and 0.001 as the learning rate, Adam is utilized as an optimizer to minimize the loss function. The ischemic stroke prediction mechanism is improved when all of these layers are combined. The proposed model leverages the best hyperparameters, as stated in Table 4.

2) THE BASELINE MODELS

The justification for choosing the seven models (as the baseline models) was based on the findings of the related work, which delineate the model's performance as stated in Table 1. CNN and LSTM models were included to observe if there is any performance against the hybrid of these two models. The following paragraph describes the models:

1) Random Forest (RF): This model can be used for classification and regression problems and as an ensemble learning technique [27]. Their strategy depends on the extensive training of a forest of decision trees. When used for classification problems, the RF output is the

TABLE 3. Hyperparameter search space and best hyperparameters.

Hyperparameter	Search Space	Value
Optimizer	Adam, adadelta, rmsprop, sgd	Adam
Activation functions (Hidden layers	ReLU, tanh, elu	ReLU
Dropout rate	0.1-0.5	0.3
Epoch	10, 20, 40, 60	40
Batch size	16, 32, 64	64

TABLE 4. Description of best-selected hyperparameters of deep learning models.

Parameters	ANN	LSTM	CNN	CNN+LSTM
Number of units	64	64	64	64
Number of layers	3	1	2	3
Number of fully connected units	64, 32, 16	16	16, 8	64, 32, 16
Number of fully connected layers	1-3	1-1	1-2	1-5
Activation Function	ReLU	ReLU	ReLU	ReLU
The last layer's activation function	Sigmoid	Sigmoid	Sigmoid	Sigmoid
Learning rate	1e-4	1e-4	1e-4	1e-4
Loss function	Adam	Adam	Adam	Adam
Number of epochs (in each fold)	40	40	40	40
Optimizer	Adam	Adam	Adam	Adam
Dropout	0.3	0.2	0.3	0.3

class most trees choose. A method that employs several classifiers to address complicated issues and enhance the efficacy and precision of a model is known as ensemble learning. As its name suggests, the RF classifier "combines a large number of decision trees on diverse subsets of a given dataset and calculates an average to increase the accuracy of its predictions."

2) K-Nearest Neighbor (KNN): KNN [28] is a kind of slow learning in which all classification computations are maintained, and there is no distinct preprocessing phase. This data categorization approach determines decisions based on the closeness of training data points on the feature map. Using the "Euclidean distance measure," the KNN classifier provides predictions about the target class. The dataset determines the ideal value of the classifier's performance control parameter, k. Following an analysis of the consequences, the excellent value is determined. Our research used a K value of 3.

3) Extreme Gradient Boosting (XGBoost):XGBoost implements the gradient augmentation approach**well** [29]. The gradient gain alternative may be rigorously developed for precision and optimization, even if no mathematical breakthroughs exist in this specific instance. A linear representation is used, and the newborn tree may be a strategy that utilizes multiple artificial intelligence algorithms to assess whether a susceptible rookie would result in a trustworthy rookie to increase the model's accuracy.

4) Logistic Regression (LR): LR [30] is relatively prominent among supervised learning ML approaches. It is a technique for predicting a categorical dependent variable by evaluating several contributing factors. The primary distinction between logistic and linear regression is in their respective uses. Linear regression addresses regression difficulties, while logistic regression addresses classification issues. Any multicollinear data may be studied using ridge regression, a model-tuning approach. Through this method, L2 regularization is accomplished.

5) Artificial Neural Network (ANN): An ANN [31] is a distributed, massively parallel processor composed of fundamental processing units with an innate propensity to retain and make experimental data available. Because ANN is so good at resolving multivariate and non-linear modeling problems, such as function approximations and classification, it is often used as a surrogate or response surface approximation model. ANN is a data processing technique inspired by the organic nervous system of the human brain. In addition to neurons, it contains input, output, hidden, and activation layers. Figure 8 displays the ANN structure.

6) Long short-term memory (LSTM): The LSTM [32] method is an RNN subset. Traditional RNNs fail to make sense of data sets with ten or more time steps, their primary shortcoming. If necessary, LSTM may prioritize and retain data for an extended time. Data separated by over a thousand-time interval may still be connected. Each node in an LSTM is provided with the input text, output Ht 1, and bias Ct = 1 (cell state of the previous node). Long-term, the condition of the cell hides vital information. Data flows via the It (input gate), Ft (forget gate), and Ot (output gate) gates of the LSTM (output gate). These gates enable the LSTM node to retain or discard the initial cell state to generate the subsequent outputs. Figure 9 depicts the LSTM structure.

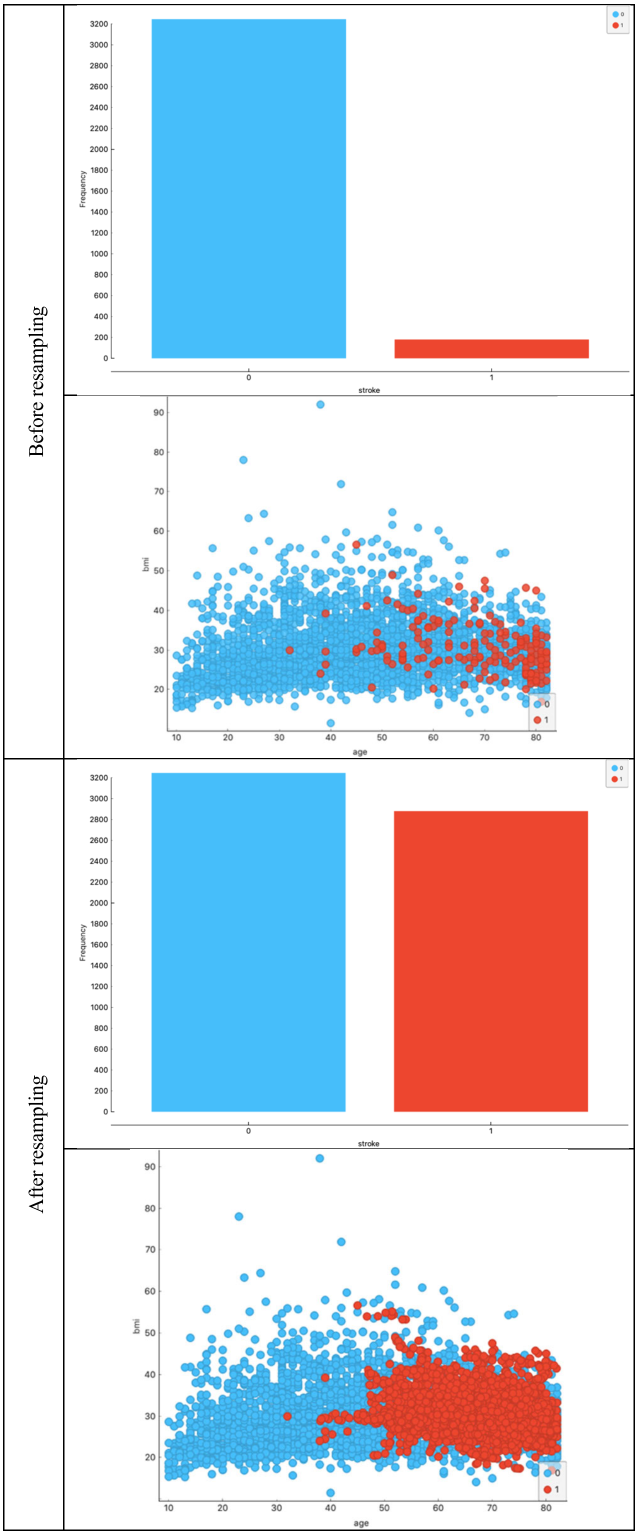


FIGURE 6. The distribution before and after resampling with smote.

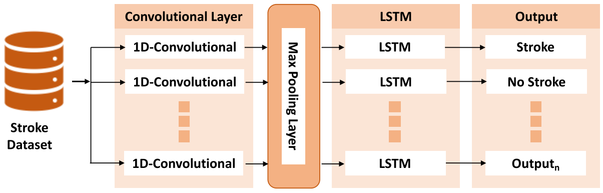


FIGURE 7. Illustrates the implementation workflow of the proposed model (CNN+LSTM).

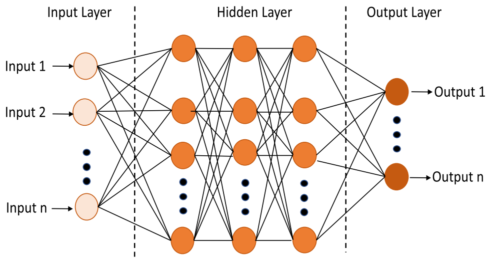


FIGURE 8. The structure of ANN.

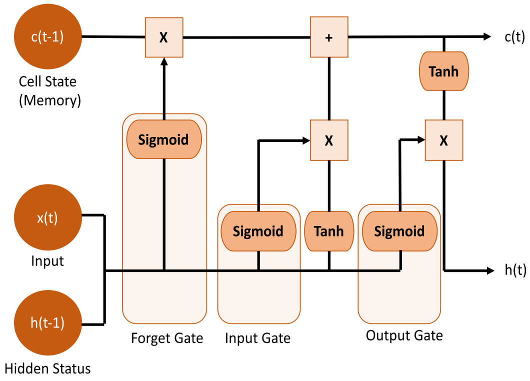


FIGURE 9. The structure of LSTM.

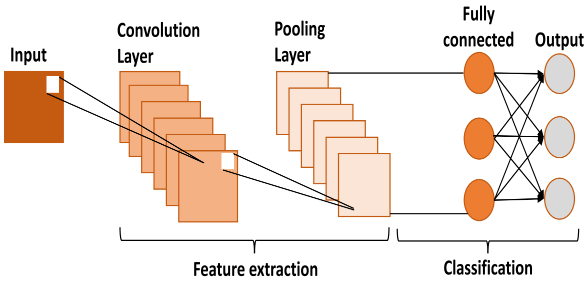


FIGURE 10. The structure of CNN.

7) Convolutional Neural Network (CNN): CNN-based models depend primarily on "convolutional" processes. CNNs can learn and extract the most significant properties from data due to their capacity to do these tasks [32]. Each CNN layer generates a feature map that activates a separate kernel by sliding it over the input. The fundamental advantage of CNNs over DNNs is their ability to reduce the computational cost of each successive layer. These models' convoluted features and compact input data representations may be utilized in downstream tasks such as classification. Figure 10 illustrates the core design of a CNN.

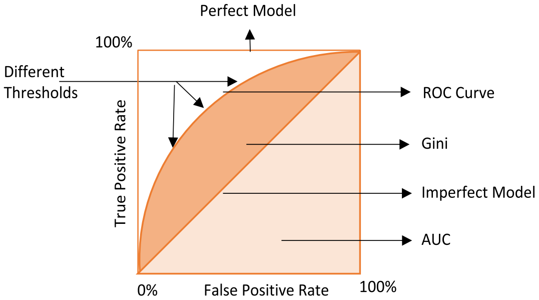
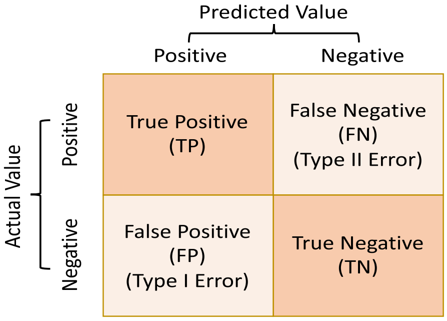


FIGURE 11. AUCROC plot description.



		Predicted Value	
		Positive	Negative
Actual Value	Positive	True Positive (TP)	False Negative (FN) (Type II Error)
	Negative	False Positive (FP) (Type I Error)	True Negative (TN)

FIGURE 12. Confusion matrix description.

E. MODEL EVALUATION

Accuracy, precision, specificity, recall, and F1 score are used to assess the effectiveness of models. The formula of the metrics is shown in Equation 1 – Equation 4. The efficacy of the proposed deep learning model is determined by comparing the predicted and actual results. The true positive (TP) and true negative (TN) scores represent the accuracy with which the classifier model can determine whether or not a patient has had a stroke (TN). False positives (FP) and false negatives (FN) reflect inaccurate model predictions (FN). Accuracy is the proportion of genuine positives relative to the total number of positives. Recall measures the number of accurate predictions, while specificity counts the number of false negatives. The function measure determines average recall and precision.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

$$\text{Precision} = \frac{(TP)}{(FP + TP)} \tag{2}$$

$$\text{Recall} = \frac{(TP)}{(FN + TP)} \tag{3}$$

$$\text{F1-Score} = 2\,x\,\frac{(Precision\ x\ Recall)}{(Precision + Recall)} \tag{4}$$

The function of ROC Figure 11 depicts the link between the true positive rate (TPR) and the false positive rate (FPR) via the receiver operating characteristic (ROC) curve (FPR). The area under the receiver operating characteristic (ROC) curve demonstrates a classifier's ability to differentiate between its two classes (AUC). When AUC is substantial, model predictions are accurate.

The confusion matrix shown in Figure 12 is a tabular depiction of the proportion of correct and wrong predictions produced by the classifier. It is a metric for assessing a classification model's performance. It is feasible to calculate and use measures such as accuracy, precision, recall, and F1-score to evaluate the utility of a classification model. Confusion matrices are preferable to classification accuracy as an indicator of a model's performance.

F. MODEL INTERPRETATION

Lundberg and Lee introduced the SHAP (Shapley Additive Explanations) [33] value to assess an individual's contribution to a collective effort. The objective was to distribute the prizes of victory according to each contestant's contribution to the final score. Since Shapley values comply with the principles of local precision (additivity), consistency (symmetry), and nonexistence, they may be utilized to compensate all parties fairly or appropriately (null effect). In prediction work, Shapley values may be rationalized as a realistic distribution of feature importance given a specific model output. Shapley values account for the magnitude and direction of each feature's effect on the model's performance or prediction. The Shapley value quantifies a feature's relevance (contribution size) and orientation (sign). Some personality qualities influence activity prediction positively, whereas others influence the prediction of inactivity negatively (i.e., a negative contribution to activity prediction). Specifically, the Shapley value in Equation 5 describes the significance of the feature:

$$\phi_i = \frac{1}{|N|!}\sum_{\mathrm{S\subseteq N\setminus\{i\}}} |S|!\,(|N| - |S| - 1)!\,|f\,(S \cup |i|) - f\,(S)| \tag{5}$$

The ML model's output, f (S), can be described by a collection of S features, but the set of all potential features, N, cannot. The Shapley value of "feature I" is computed by averaging its contributions over all possible feature set permutations. Therefore, features are added to the set one at a time, and the change in the output indicates their importance. Considering the ordering of features known to alter the reported changes in a model's output when correlated features are present, this method is especially advantageous. Since ML and DL models are interpretable, their conclusions may be valid. If the logic behind the predictions of a complex model could be comprehended, its black-box aspect may be minimized or removed.

SHAP's additive feature attribution is distinct, mathematically sound, and precise. It pulls from other sources, such as an enlargement of the feature's relevance, game theory, and in-depth localization. The final conditional prediction is the sum of all the qualities of the model's outcomes. Response reliability is assured using game-theoretic arguments and an average of all feasible feature orderings. The primary advantage of this method is that it evaluates the relevance of various sample characteristics by giving each one a score (a SHAP value). SHAP values may be estimated using the following techniques:

- Substitute a random value for a subset of sample characteristics.
- Determine the forecast for each variant sample.
- Use the SHAP Kernel to calculate SHAP values.

Increasing the sample size will decrease the estimate's variance since the original estimate was based on a random sample. It also presupposes the autonomy of individual traits. If we substitute x1 with random values when x1 and x2 are highly correlated in our training data, our SHAP value estimate will suffer since it will be based on fewer examples of x1 and x2 in the training set. Each prediction in a given dataset is analyzed using a linear model in a Kernel SHAP Explainer. The loss Equation 6 that this kernel seeks to optimize is as follows:

$$L\left(f, g, \pi_{\mathrm{x}}\right) = \sum \left[f\left(h_x\left(z'\right)\right) - g\left(z'\right)\right]^2 \pi_x\left(z'\right) \tag{6}$$

Predictions that use less or virtually all qualities get more weight in the x term, which is a compliance weighting strategy. Only tree-based methods may be utilized with the SHAP Tree Explainer. Instead of random sampling, trees simulate missing data by bypassing the critical decision paths. Therefore, Tree Explainer produces results that are both predictable and unaffected by context. Complexity is reduced from O(T L2M) to O(T LD2) by pushing all varieties through the tree concurrently rather than iterating over each conceivable feature combination (or subset thereof), where M is the number of features, T is the number of trees, L is the maximum number of leaves, and D is the maximum tree depth.

Since no linear models are used, calculating SHAP values using this approach is more efficient. The SHAP values are calculated by looking at how the conditional expectation of all features changes when specific characteristics are changed. This approach estimates the shift based on the conditional expectation given this subset sample, ignoring the nodes of features that aren't present. SHAP is a novel method for gaining a deeper understanding of a projected occurrence by shedding light on the interdependence of numerous components in a predictive model. SHAP is an effective strategy to separate the effects of the drivers and break down the forecast on the impact of its component features. When using ML methods for stroke classification, feature selection is crucial. Modelers may dissect any prediction into its constituent parts by adding up the SHAP value and explaining how each feature value contributed to the result.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this subsection, we presented the experimental results, which aim to evaluate the model's performance and explain the results using the SHAP method.

A. PREDICTION MODEL PERFORMANCE ANALYSIS

Model classification for both ML and DL classifiers was first performed. All models were then validated using 10-fold cross-validation and evaluated based on performance metrics. The Python programming language was used to execute all experiments, and extensions were created for both Python and the scikit-learn ML library. The hyperparameter search space and best selected hyperparameters are shown in Table 3, Table 4, and Table 5, respectively. The experimental results, which comprise prediction models' performance results, models' AUCROC results, models' confusion matrix results, and models' training and testing time results, are shown in Table 6, Figure 13, Figure 14, Figure 15, and Figure 16, respectively.

B. FEATURES IMPORTANCE

Figure 17 depicts the plots of feature significance matrices for binary classification problems in stroke prediction. We used the "Permutation Feature Importance Technique," illustrated by a bar chart, to determine the relative significance of several variables. The result indicated that the higher the rank, the more likely the hybrid CNN+LSTM model is to emphasize this trait. The result also provides insight into the most significant stroke prediction factors. With a better understanding of these features, stroke patients may benefit from more targeted rehabilitation programs and treatments.

C. MODEL INTERPRETATION

Here, we evaluate the effect of different parameters on the outcomes of the optimal hybrid DL model (CNN+LSTM classifier). Figure 18 illustrates the relationships between these factors and the two potential outcomes (1 = stroke, 0 = no stroke). The values of these parameters (the SHAP summaries) are presented in descending order, with the most critical qualities shown first. Red indicates high importance, and blue indicates low values for each parameter's impact on the model's output. The variables such as age, patients who stayed in urban residences, were married, never-smoked patients, and patients who worked in private firms substantially impacted stroke prediction. However, cardiovascular risk factors such as hypertension and heart disease contribute only marginally to the prognosis of stroke. However, being self-employed, being male, and working in the public sector had little influence on stroke prediction.

D. DISCUSSIONS

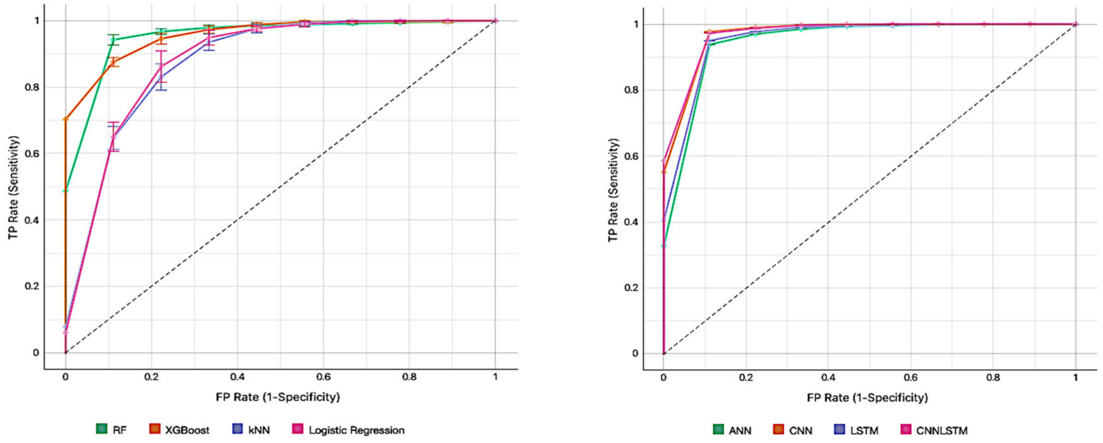
Since stroke is a devastating disease worldwide, it is critical to precisely detect the outcome early on so that patients can recover. As a result, this model was created to address the disease's restrictions. The EDA shows the dataset is highly imbalanced (Figure 6). Thus, we deployed the SMOTE resampling technique to balance the samples to avoid overfitting and increase the data quality. For the model classification, the results showed that the proposed CNN+LSTM model achieved the highest accuracy

TABLE 5. Description of best-selected hyperparameters of machine learning models.

Algorithms	Hyperparameter
RF	Deployed 300 trees, considering 10 attributes per split.
XGBoost	Each of the 100 trees had a learning rate of 0.300 and a maximum depth of 6.
kNN	Start k = 10, distance metrics = Euclidean distance.
LR	Regularization strength (C) = 0.01, 0.1, 1.0, 10.0, and higher. Regularization penalties = L2 (Ridge). Solver options = 'sag' and 'saga.' Max_iter = 100

TABLE 6. The results of the model classification analysis.

Classifiers	Accuracy	F1-Score	Precision	Recall
Machine Learning-based				
RF	0.9407	0.9357	0.9545	0.9177
XGBoost	0.8879	0.8747	0.9216	0.8323
kNN	0.8234	0.8335	0.7485	0.9403
LR	0.8142	0.8072	0.7880	0.8274
Deep Learning-based				
CNN+LSTM	**0.9587**	**0.9558**	**0.9617**	**0.9510**
CNN	0.9554	0.9525	0.9540	0.9500
LSTM	0.9251	0.9215	0.9077	0.9358
ANN	0.9158	0.9120	0.8961	0.9285



(a) ML-based Models of AUCROC Analysis

(b) DL-based Models of AUCROC Analysis

FIGURE 13. Results of AUCROC analysis.

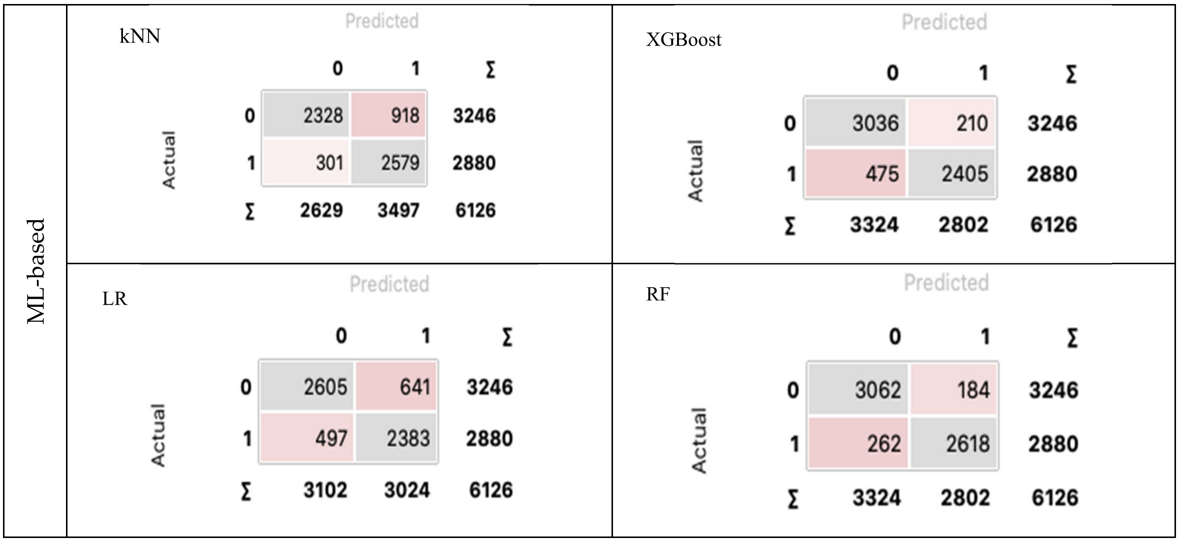
of 95.9% (Table 6). CNN+LSTM model also attains the highest AUCROC (Figure 13) of 98.9%. The confusion matrix analysis (Figure 14 and Figure 15) shows the proposed model can correctly classify the samples based on the number of instances predicted to suffer from stroke.

All of the experimental findings above supported this study's findings, which demonstrated that the proposed model is superior in predicting strokes. Regarding the time required to train and test the models, it was discovered that the suggested model requires 42.5 seconds to train and 0.3 seconds to test. This information can be found in Figure 16. kNN, on the other hand, is the model that can be trained the quickest, which takes 0.32 seconds, while LR has the shortest testing time (0.06 sec.). Based on the findings, the proposed model requires more processing time than the other models. The results might be considered a potential disadvantage for the proposed model.

To further understand the outcome of the best hybrid model, we used the Permutation Features Importance Technique to identify the critical features in the dataset. We used the SHAP method to measure the best-performing model's impact on the dataset's features. Feature "age" was identified as the most important feature in the stroke prediction, and CNN+LSTM highlighted the feature "age" as having a strong positive impact on the stroke prediction. The comparative analysis between the current and previous studies is shown in Table 7.

V. LIMITATIONS AND IMPLICATIONS

This section briefly discusses the limitations and implications of this study.



ML-based

kNN

Actual \ Predicted	0	1	Σ
0	2328	918	3246
1	301	2579	2880
Σ	2629	3497	6126

XGBoost

Actual \ Predicted	0	1	Σ
0	3036	210	3246
1	475	2405	2880
Σ	3324	2802	6126

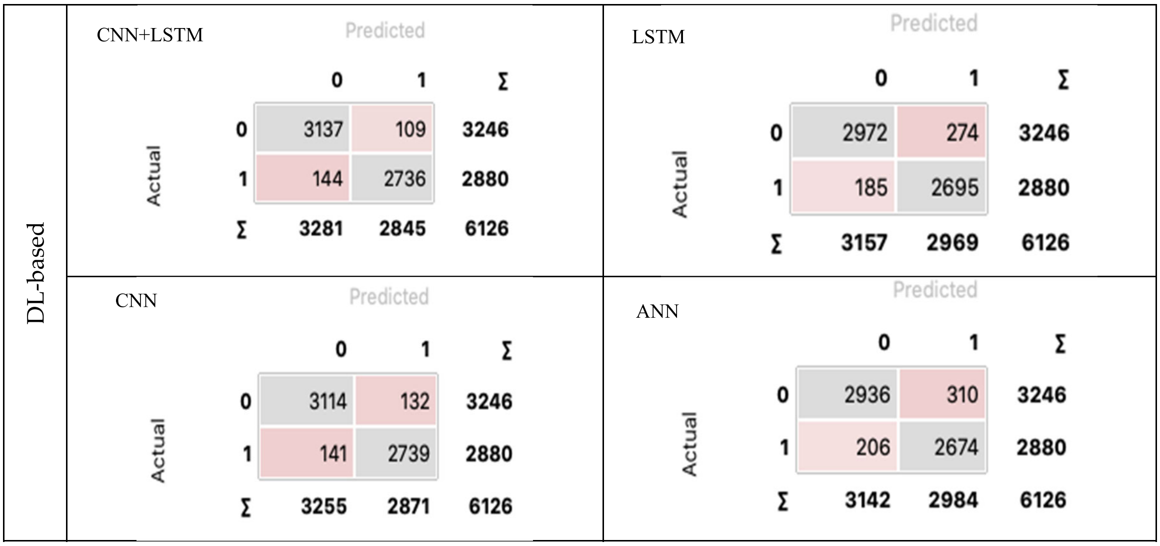
LR

Actual \ Predicted	0	1	Σ
0	2605	641	3246
1	497	2383	2880
Σ	3102	3024	6126

RF

Actual \ Predicted	0	1	Σ
0	3062	184	3246
1	262	2618	2880
Σ	3324	2802	6126

FIGURE 14. Results of confusion matrix analysis for ML-based models.



DL-based

CNN+LSTM

Actual \ Predicted	0	1	Σ
0	3137	109	3246
1	144	2736	2880
Σ	3281	2845	6126

LSTM

Actual \ Predicted	0	1	Σ
0	2972	274	3246
1	185	2695	2880
Σ	3157	2969	6126

CNN

Actual \ Predicted	0	1	Σ
0	3114	132	3246
1	141	2739	2880
Σ	3255	2871	6126

ANN

Actual \ Predicted	0	1	Σ
0	2936	310	3246
1	206	2674	2880
Σ	3142	2984	6126

FIGURE 15. Results of confusion matrix analysis for DL-based models.

A. LIMITATIONS

The limitations are:

- The stroke dataset utilized in this study is highly imbalanced and consists of sample values labeled "unknown." The preprocessed data is performed to increase its quality. However, this process could have performed similar methods and classifiers on several other stroke datasets for comparison. The performance could change significantly on various datasets. Nonetheless, the results are expected to be satisfactory even if the models utilized diverse datasets.
- The study assumes that the dataset used in the literature is precise, complete, and free of noisy cases. However, as in many circumstances, a few noisy events may impact our models' performance. To further improve the success of the results, different resampling techniques, feature selection methods, model validations, and different evaluations of error can be deployed. This means the hybrid model can be exhaustively processed to ensure its robustness, efficiency, and effectiveness in predicting stroke based on the patient's clinical data.
- This study only deployed one hybrid model, which could have possible hyperparameter tuning biases affecting the result.
- Predicting strokes involves analyzing various risk factors, including medical history, lifestyle factors, genetic predisposition, and clinical symptoms. While CNNs and LSTMs can capture complex patterns in data, predicting strokes accurately often requires considering a wide range of factors, some of which the model may not effectively capture.

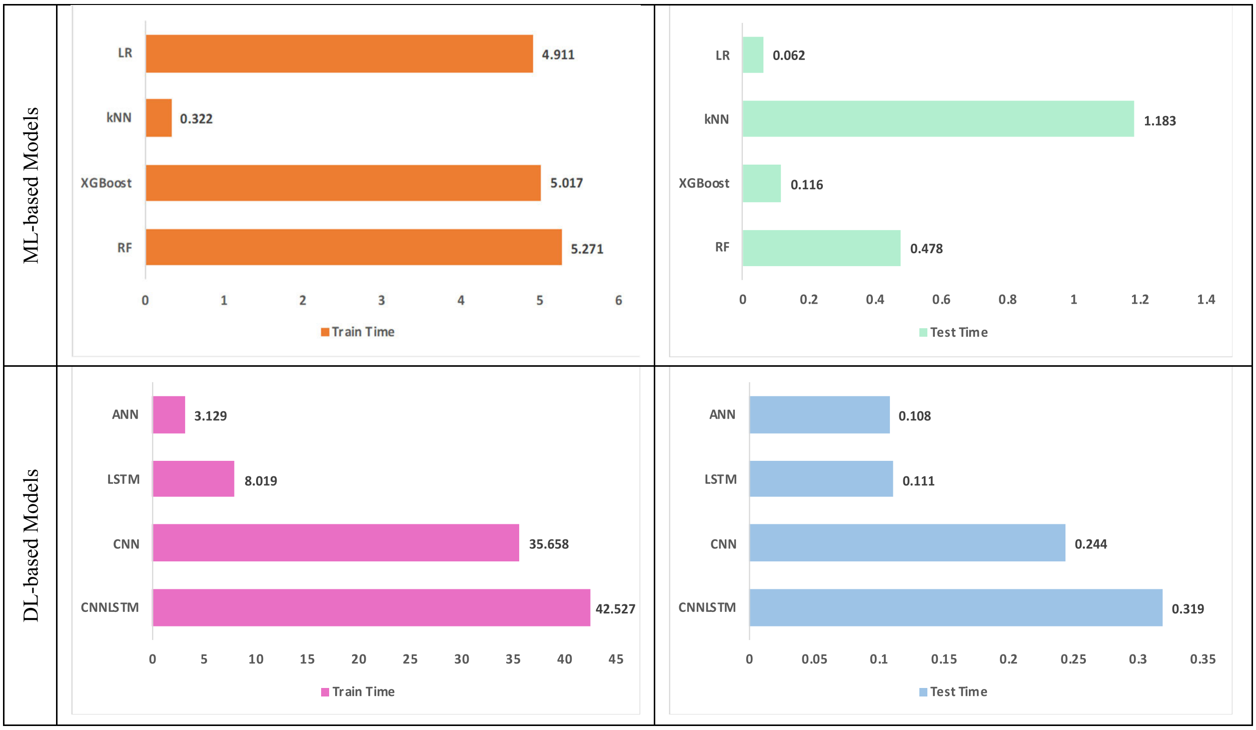


FIGURE 16. The results of training and testing time taken analysis.

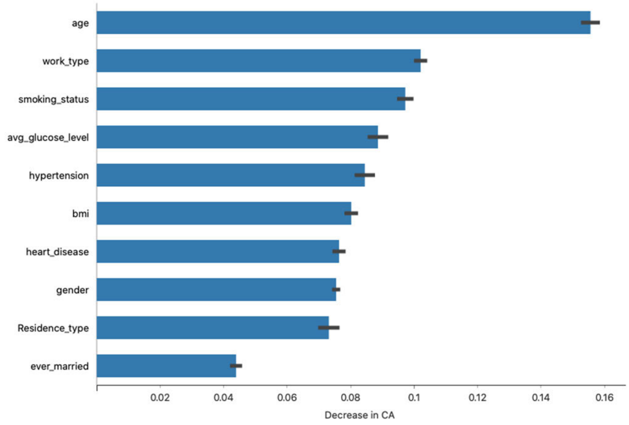


FIGURE 17. Feature importance matrix plots of the best prediction (CNN+LSTM) model.

- Interpretability is crucial in medical applications to gain the trust of healthcare professionals and ensure that predictions are clinically meaningful. Hybrid CNN+LSTM models may lack interpretability, making it challenging for healthcare professionals to understand the model's predictions and trust its decisions.
- Medical datasets often suffer from missing data and class imbalance, where instances of strokes may be relatively rare compared to non-stroke instances. Handling missing data and class imbalance is essential to prevent biases in the model's predictions and ensure robust performance.

TABLE 7. Comparison of the current study with previous studies that use the same dataset.

Ref.	Authors (Year)	Balancing Methods/ Data Preprocessing	Deployed Classifiers	Accuracy Performance
[19]	Sailasya et al. [2021]	Missing value handling, label encoding, and imbalanced data handling using ROS	LR, DT, RF, K-NN, SVM and NB	NB (82%)
[20]	Tazin et al. [2021]	Use SMOTE for data balancing	LR, DT, RF, and VC	RF (95.7%)
[22]	Akter et al. [2022]	Train-Test Split Method	RF, SVM, DT	RF (95.3%)
[23]	Dev et al. [2022]	Dimensionality reduction using PCA	DT, RF, ANN	ANN with PCA (75%)
	Current Study	• 10 – Fold Cross-validation • SMOTE • Delete sample that has "unknown" value	RF, XGBoost, KNN, LR, CNN, ANN, LSTM, CNN+LSTM	CNN+LSTM (95.9%)

B. POTENTIAL IMPLICATIONS

Despite the abovementioned limitations, the proposed model holds significant clinical implications and could potentially revolutionize stroke management. Here are some potential clinical implications:

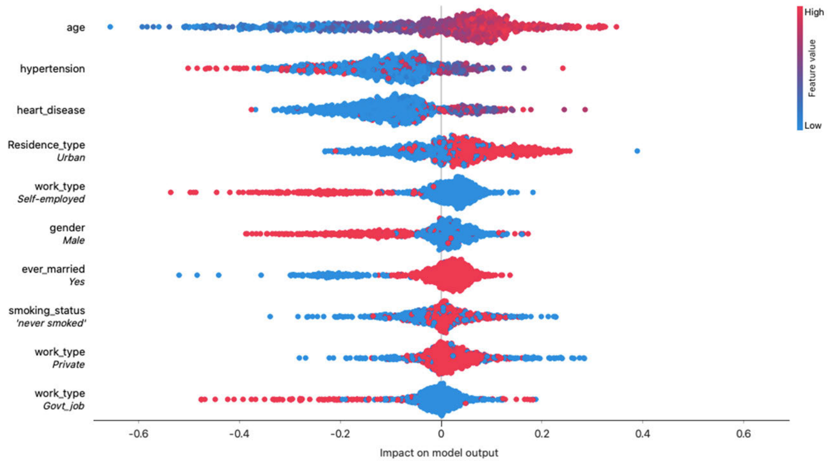


FIGURE 18. Parameters impact on CNN+LSTM model output for the stroke prediction.

- Early prediction: Early prediction of ischemic stroke risk factors can enable healthcare providers to intervene proactively, thereby preventing or minimizing the occurrence of strokes. By leveraging predictive modeling techniques, the proposed model can identify individuals at high risk of developing ischemic stroke based on their demographic information, medical history, lifestyle factors, and other relevant data.
- Timely Intervention: Early identification of individuals at risk of ischemic stroke allows for timely interventions, such as lifestyle modifications, medication management, and targeted preventive therapies. Healthcare providers can use the model's predictions to tailor preventive strategies to each patient's specific risk profile, potentially reducing the incidence and severity of ischemic strokes.
- Personalized Medicine: The model can facilitate personalized medicine approaches by identifying individualized risk factors and tailoring interventions accordingly. Healthcare providers can prioritize resources and interventions for patients at the highest risk of ischemic stroke, optimizing healthcare delivery and resource allocation.
- Integration into Healthcare Systems: The model could be integrated into existing EHR systems or clinical decision support systems used by healthcare providers. Integration could involve developing user-friendly interfaces or dashboards that display patients' stroke risk scores and recommendations based on the model's predictions. Automated alerts or notifications could prompt healthcare providers to review and address patients' stroke risk factors during routine clinical encounters.

VI. CONCLUSION

In conclusion, our study highlights the transformative potential of ML-based predictive models, particularly the CNN+LSTM architecture, in revolutionizing ischemic stroke prediction. With promising accuracy rates of 95.9% and achieving the highest AUCROC value of 98.9%, our findings underscore the effectiveness of this approach. However, addressing limitations such as data quality and interpretability is crucial, which may hinder healthcare professionals' understanding and trust in the model's predictions. Despite these challenges, the proposed model is promising to optimize healthcare resource allocation and enhance patient outcomes. Yet, successful integration into real-world settings demands collaborative efforts and addressing multifaceted challenges. Future research should prioritize validating the model's generalizability and comparing it with other methodologies. We can propel stroke prediction and management to new heights through ongoing innovation and collaboration, ultimately improving healthcare delivery and patient outcomes.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia, for funding this research work through project number RI-44-0213.

REFERENCES

[1] L. Hwangbo, Y. J. Kang, H. Kwon, J. I. Lee, H.-J. Cho, J.-K. Ko, S. M. Sung, and T. H. Lee, "Stacking ensemble learning model to predict 6-month mortality in ischemic stroke patients," *Sci. Rep.*, vol. 12, no. 1, p. 17389, Oct. 2022.

[2] M. Mahmud, M. S. Kaiser, M. M. Rahman, M. A. Rahman, A. Shabut, S. Al-Mamun, and A. Hussain, "A brain-inspired trust management model to assure security in a cloud based IoT framework for neuroscience applications," *Cognit. Comput.*, vol. 10, no. 5, pp. 864–873, Apr. 2018.

[3] M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. A. Mamun, and M. Mahmud, "Application of deep learning in detecting neurological disorders from magnetic resonance images: A survey on the detection of Alzheimer's disease, Parkinson's disease and schizophrenia," *Brain Informat.*, vol. 7, no. 1, pp. 1–21, Oct. 2020.

[4] M. Mahmud, M. S. Kaiser, T. M. McGinnity, and A. Hussain, "Deep learning in mining biological data," *Cognit. Comput.*, vol. 13, no. 1, pp. 1–33, Jan. 2021.

[5] G. Fang, Z. Huang, and Z. Wang, "Predicting ischemic stroke outcome using deep learning approaches," *Frontiers Genet.*, vol. 12, Jan. 2022, Art. no. 827522.

[6] M. Chun, R. Clarke, B. J. Cairns, D. Clifton, D. Bennett, Y. Chen, Y. Guo, P. Pei, J. Lv, C. Yu, L. Yang, L. Li, Z. Chen, and T. Zhu, "Stroke risk prediction using machine learning: A prospective cohort study of 0.5 million Chinese adults," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 8, pp. 1719–1727, May 2021.

[7] M. S. Jabal, O. Joly, D. Kallmes, G. Harston, A. Rabinstein, T. Huynh, and W. Brinjikji, "Interpretable machine learning modeling for ischemic stroke outcome prediction," *Frontiers Neurol.*, vol. 13, May 2022, Art. no. 884693.

[8] A. Adiguzel, E. M. Arsava, and M. A. Topcuoglu, "Temporal course of peripheral inflammation markers and indexes following acute ischemic stroke: Prediction of mortality, functional outcome, and stroke-associated pneumonia," *Neurological Res.*, vol. 44, no. 3, pp. 224–231, Sep. 2021.

[9] D. Chaudhary, V. Abedi, J. Li, C. M. Schirmer, C. J. Griessenauer, and R. Zand, "Clinical risk score for predicting recurrence following a cerebral ischemic event," *Frontiers Neurol.*, vol. 10, p. 1106, Nov. 2019.

[10] Z. Yuanyuan, W. Jiaman, Q. Yimin, Y. Haibo, Y. Weiqu, and Y. Zhuoxin, "Comparison of prediction models based on risk factors and retinal characteristics associated with recurrence one year after ischemic stroke," *J. Stroke Cerebrovascular Diseases*, vol. 29, no. 4, Apr. 2020, Art. no. 104581.

[11] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, "Artificial intelligence transforms the future of health care," *Amer. J. Med.*, vol. 132, no. 7, pp. 795–801, Jul. 2019.

[12] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in *Proc. 8th Annu. Ind. Autom. Electromechanical Eng. Conf. (IEMECON)*, Aug. 2017, pp. 158–161.

[13] Y. Zhang, W. Song, S. Li, L. Fu, and S. Li, "Risk detection of stroke using a feature selection and classification method," *IEEE Access*, vol. 6, pp. 31899–31907, 2018.

[14] L. García-Temza, J. L. Risco-Martín, J. L. Ayala, G. R. Roselló, and J. M. Camarasaltas, "Comparison of different machine learning approaches to model stroke subtype classification and risk prediction," in *Proc. Spring Simul. Conf. (SpringSim)*, Apr. 2019, pp. 1–10.

[15] M. S. Singh, P. Choudhary, and K. Thongam, "A comparative analysis for various stroke prediction techniques," in *Proc. Int. Conf. Comput. Vis. Image Process.*, Jaipur, India, Sep. 2019, pp. 27–29.

[16] J. Heo, J. G. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Heo, "Machine learning-based model for prediction of outcomes in acute stroke," *Stroke*, vol. 50, no. 5, pp. 1263–1265, May 2019.

[17] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of stroke disease using machine learning algorithms," *Neural Comput. Appl.*, vol. 32, no. 3, pp. 817–828, Jan. 2019.

[18] Y. Wu and Y. Fang, "Stroke prediction with machine learning methods among older Chinese," *Int. J. Environ. Res. Public Health*, vol. 17, no. 6, p. 1828, Mar. 2020.

[19] G. Sailasya and G. L. A. Kumari, "Analyzing the performance of stroke prediction using ML classification algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 539–545, 2021.

[20] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. M. Khan, "Stroke disease detection and prediction using robust learning approaches," *J. Healthcare Eng.*, vol. 2021, pp. 1–12, Nov. 2021.

[21] N. Darabi, N. Hosseinichimeh, A. Noto, R. Zand, and V. Abedi, "Machine learning-enabled 30-day readmission model for stroke patients," *Frontiers Neurol.*, vol. 12, Mar. 2021, Art. no. 638267.

[22] B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas, and U. Sara, "A machine learning approach to detect the brain stroke disease," in *Proc. 4th Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Tirunelveli, India, Jan. 2022, pp. 897–901.

[23] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," *Healthcare Analytics*, vol. 2, Nov. 2022, Art. no. 100032.

[24] *Stroke Prediction Dataset*. Accessed: May 25, 2023. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[25] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset," *Sensors*, vol. 22, no. 9, p. 3246, Apr. 2022.

[26] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[27] J. Wang, X. Gong, H. Chen, W. Zhong, Y. Chen, Y. Zhou, W. Zhang, Y. He, and M. Lou, "Causative classification of ischemic stroke by the machine learning algorithm random forests," *Frontiers Aging Neurosci.*, vol. 14, Apr. 2022, Art. no. 788637.

[28] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, Jun. 2022, Art. no. 100071.

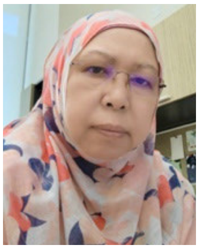
[29] C.-C. Chung, E. C.-Y. Su, J.-H. Chen, Y.-T. Chen, and C.-Y. Kuo, "XGBoost-based simple three-item model accurately predicts outcomes of acute ischemic stroke," *Diagnostics*, vol. 13, no. 5, p. 842, Feb. 2023.

[30] C. Ranathunge, S. S. Patel, L. Pinky, V. L. Correll, S. Chen, O. J. Semmes, R. K. Armstrong, C. D. Combs, and J. O. Nyalwidhe, "Promor: A comprehensive R package for label-free proteomics data analysis and predictive modeling," *Bioinf. Adv.*, vol. 3, no. 1, Mar. 2023, Art. no. vbad025.

[31] S. Thammaboosadee and T. Kansadub, "Data mining model and application for stroke prediction: A combination of demographic and medical screening data approach," *Interdiscipl. Res. Rev.*, vol. 14, no. 4, pp. 61–69, 2019.

[32] N. Hatami, T.-H. Cho, L. Mechtouff, O. F. Eker, D. Rousseau, and C. Frindel, "CNN-LSTM based multimodal MRI and clinical data fusion for predicting functional outcome in stroke patients," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 3430–3434.

[33] Y. Zheng, Z. Guo, Y. Zhang, J. Shang, L. Yu, P. Fu, Y. Liu, X. Li, H. Wang, L. Ren, W. Zhang, H. Hou, X. Tan, and W. Wang, "Rapid triage for ischemic stroke: A machine learning-driven approach in the context of predictive, preventive and personalised medicine," *EPMA J.*, vol. 13, no. 2, pp. 285–298, May 2022.



SAPIAH SAKRI (Member, IEEE) received the B.Sc. degree in computer science from Universiti Kebangsaan Malaysia, Malaysia, in 1985, the M.Sc. degree in information science from the University of Malaya, Malaysia, in 1997, and the Ph.D. degree in information security from Universiti Kebangsaan Malaysia, in 2007. She served Malaysian public sector for 30 years. Her latest position (before retiring) was the Director of the Information Communication Technology Policy and Strategic Division, Prime Minister's Department. She is currently an Assistant Professor with the Department of Information Systems, Princess Nourah bint Abdulrahman University, Saudi Arabia. She has published articles mainly in the domain of data analytics. Her research interests include information security management, cloud computing security, data analytics, and data science. She received the highest award from the Prime Minister of Malaysia for her achievement in initiating the SMART city known as eKL, in 2012.

SHAKILA BASHEER is currently an Assistant Professor with the College of Computer and Information Systems, Princess Nourah bint Abdulrahman University. She is working on data mining, machine learning for vehicular networks, blockchain, and the Internet of Things. She has more than ten years of teaching experience and has published more technical articles in international journals/international conference proceedings/ edited chapters of famous publications. She has worked and contributed to data mining, image processing, and fuzzy logic. Her research also focuses on data mining algorithms using fuzzy logic.

ZUHAIRA MUHAMMAD ZAIN received the bachelor's degree in information science and the master's degree in computer science from the National University of Malaysia, in 2000 and 2007, respectively, and the Ph.D. degree in software engineering from Universiti Putra Malaysia, in 2013. She was a Programmer with Fujitsu Systems Business Malaysia Bhd., from 2000 to 2007. She is currently an Associate Professor with the College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include software engineering, software quality, software evaluation, and data analytics.

NURUL HALIMATUL ASMAK ISMAIL received the degree in computer science from Universiti Sains Malaysia, in 2000, the master's degree from Universiti Putra Malaysia, in 2009, and the Ph.D. degree in computer science, specialization in networking from Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia, in 2015. She was with Majlis Amanah Rakyat, Malaysia, as a Lecturer for the Higher National Diploma in Computing and Edexcel program, U.K. She was involved with research groups, program accreditation, and syllabus construction, and years of experience with Edexcel program. Recently, she joined Princess Nourah bint Abdulrahman University, Saudi Arabia, as an Assistant Professor with the Department of Computer Science and Information Technology, College of Community. Her research interests include 6LoWPAN routing protocol, the Internet of Things (IoT), and machine learning.

DUA' ABDELLATEF NASSAR received the Doctor of Philosophy degree in information systems from UNITEN, Malaysia. She is an Assistant Professor with the Information Systems Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University. With more than ten years of expertise in the classroom, she has authored numerous pieces that have been published in ISI and SCOPUS index publications. Living in Jordan, she is accessible for speaking engagements and freelance work. She is also available for freelance research work.

GHADAH NASSER ALDEHIM received the Ph.D. degree specializing in data mining from the University of East Anglia, U.K. She is currently an Assistant Professor with the Information Systems Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University. She is the Director of the Data Science and Analysis Program and a member of the Information Systems Program Committee, the Data Science Program Development Committee, and the Research Center Committee.

MAIS AYMAN ALHARAKI received the Bachelor of Science degree from Arab Open University, Saudi Arabia. She is currently pursuing the Master of Science degree in information systems with the College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Saudi Arabia. She is an Independent Researcher. She is active in Saudi Arabia and is open to taking on freelancing work.