Jurnal Kejuruteraan 36(4) 2024: 1759–1768
https://doi.org/10.17576/jkukm-2024-36(4)-38

Brain Stroke Prediction Using Stacked Ensemble Model

Hemalatha Gunasekaran[a], Angelin Gladys[a], Deepa kanmani[b], Rex Macedo[a] & Wilfred Blessing N R [a]

[a]*College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri, Sultanate of Oman*

[b]*Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, India*

**Corresponding author: hemalatha.david@utas.edu.om*

Received 5 January 2024, Received in revised form 27 March 2024
Accepted 27 April 2024, Available online 30 July 2024

ABSTRACT

Stroke is a potentially fatal illness that requires emergency care. There is a greater chance that the patient will recover and resume their regular life when they receive treatment and diagnosis as soon as feasible. Artificial Intelligence has the potential to significantly impact stroke diagnosis and facilitate prompt patient treatment for physicians. Machine learning can be utilized in stroke prediction by evaluating huge volumes of patient data and detecting patterns and risk variables that may contribute to the likelihood of a stroke. In this study, we explored a stacked ensemble model that uses four base models—Decision Tree, XGBoost, RandomForest, and ExtraTree classifiers to predict the stroke. We discovered that the accuracy of the stacked ensemble model was 96.35%, higher than that of the traditional machine-learning models, other ensemble models, and ANN model.

Keywords: Ensemble Model, Bagging, Voting, Stacked ensemble, Boosting

INTRODUCTION

Many risk threats, including health conditions like diabetes, heart disease, high blood pressure, high cholesterol, and atrial fibrillation, as well as unhealthy lifestyle choices like smoking, obesity, eating poorly, and not exercising, can lead to strokes (Dritsas, 2022). The primary symptoms of a stroke can be identified using the acronym "FAST". F represents Face Drooping which means that they are unable to close their mouth or laugh and not able to close their eyes also. A is Arm Weakness, may not be able to raise both arms. S is Speech Difficulty and T is for Time which refers to the time of attack and also the time for immediate action.

Identifying brain strokes early on and mitigating their risk factors is thought to be a saving grace. With machine learning and artificial intelligence (AI) making great progress in forecasting many diseases, it is possible to utilize these approaches to estimate the risk of stroke. Various classification methods have been employed for foreseeing strokes, yielding satisfactory outcomes. The accuracy of the ensemble approach in predicting different ailments has made it a popular choice for medical applications (Mienye et.al. 2020, Shilpa et. Al. 2022, Asghari et.al 2022) To improve overall performance, these strategies incorporate the prediction results of multiple classification models. The classification process is finished in two steps: first, the full dataset is trained using various base models, and then the final prediction is produced by training a meta-learner classifier prediction findings (Akinbo et.al 2021, Srinivas et al. 2023). These approaches guarantee more dependable and accurate outcomes by applying multiple models, rather than depend on exclusively on the output of a single model (Rosly et.al. 2018, Dzeroski et.al. 2009)

The core role of this paper is to develop a prediction model for stroke using stacking ensemble classifier and to analyze the performance of individual machine learning techniques with ensemble and ANN model. The primary objectives of this research are:

1. Classify the stroke dataset using different machine learning techniques such as K-Nearest Neighbor,

Support Vector Machines, Random Forest, Naive Bayes, Logistic Regression, Decision Trees, Extra Tree Classifier, and XGBoost.
2. Classify the stroke dataset using ensemble techniques such as boosting, bagging, voting, and stacking classifiers.
3. Classify the stroke dataset with the ANN model.
4. Compare and contrast the three techniques with respect to accuracy, precision, and recall.

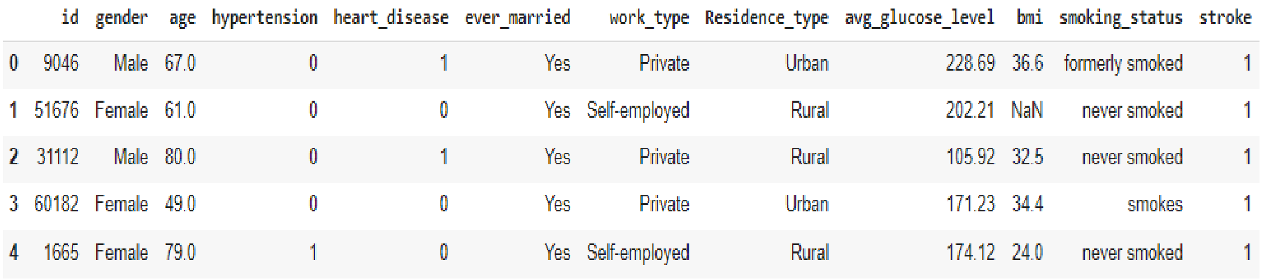
LITERATURE SURVEY

M. S. Singh et al. have used cardiovascular dataset for stroke prediction. The author has used decision tree for feature selection. The selected features are reduced using principal component analysis technique and finally the classification is done using backpropagation neural network. The author achieved an accuracy of around 97.7% accuracy.

D Ushasree et al. proposed two techniques one for selecting the best feature for stroke prediction known as "Hybrid Measures Approach for Feature Engineering (HMA-FE)" and to create ensemble ML model for classification known as Hybrid Ensemble and Feature Engineering for Stroke Prediction (HEFE-SP). The author achieved an accuracy of around 95.25%.

Abd Mizwar A. Rahim et al. used the Xtreme Gradient Boosting algorithm for classification of stroke dataset. The author achieved an accuracy of 96%.

Soumyabrata Dev et al. used the Principal Component Analysis to predict the most important features required for classification. The features such as average blood sugar level, age, heart disease and hypertension were found to be the important features. Finally, the author used perceptron neural network for classification and achieved an accuracy of 78% for this combination of features.



	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

FIGURE 1. Brain Stroke Dataset

Yuru Jing et al. used four unique methods to compare and enhance the less dominant class classification in the skewed dataset such as stroke dataset: The Synthetic Minority Over-Sampling Technique (SMOTE), Principal Component Analysis combined K-Means Clustering (PCA-K means), Focal Loss combined with the Deep Neural Network (DNN), and the ensemble weight based voting classifier. The analysis results showed that, PCA-K means combined with SMOTE and DNN Focal Loss achieved an accuracy of around 92% which is higher than the other methods.

Alruily et al. proposed a stacking ensemble model with three machine learning models such as Random Forest, XGBoost and LightGBM. The author used KNN Imputer technique to handle the missing values and SMOTE technique the handle the imbalanced dataset. The proposed stacking ensemble model achieved an accuracy of around 96.34%.

DATA SET

The stroke dataset is taken from Kaggle (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset) used in this research. It contains 5110 rows and 11 features including the target column as shown in Figure 1. The target column contains two values 1 indicates the patient has a stroke and 0 if not.

DATA PRE-PROCESSING

Data pre-processing is a series of steps as shown in Figure 2 applied on the data to make it suitable for analysis. Data can have missing values, outliers, categorical values, or data can be over-sampled or under-sampled. All the mentioned problems should be dealt with before applying a machine learning algorithm.

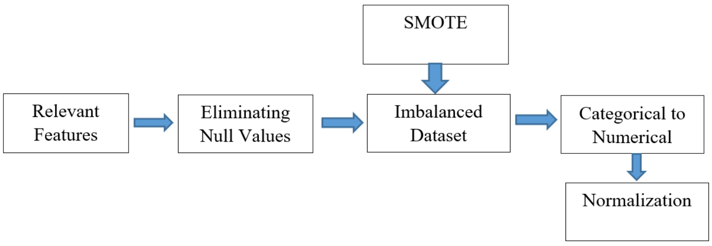


FIGURE 2. Steps in Pre-Processing

RELEVANT FEATURES

There are many dimensions or attributes in the prediction of stroke dataset, so it is necessary to identify the most important/relevant feature to predict the target variable. To understand the correlation between the different variables and the target variable heat map is generated as shown in Figure 3.

The correlation map helps to understand the relationship between different variables and helps to identify the most important feature for building the machine-learning model. Correlation varies between -1 and +1. The id column has a correlation of 0 as a result it is dropped from the dataset.

ELIMINATING NULL VALUES

The stroke prediction dataset contains 201 missing data in the BMI column. The missing data must be eliminated or replaced to create a robust machine-learning model. Hence, the 201 null values in the BMI column are replaced with the average value of that vertical data.

SKEWED DATASET

The number of instances for each class is not well balanced. The dataset has 4861 rows for class 0 and 249 rows for class 1. The figure shows the class distribution for the dataset. The machine learning algorithm performs poorly if the dataset is unbalanced. So, we have applied the SMOTE method to make the information to be balanced. In SMOTE, the class 1 instances will be oversampled to 4861 rows. As a result, both classes 0 and 1 will be perfectly balanced class. The number of instances in each class after the SMOTE technique is given in the Figures 4 and 5.

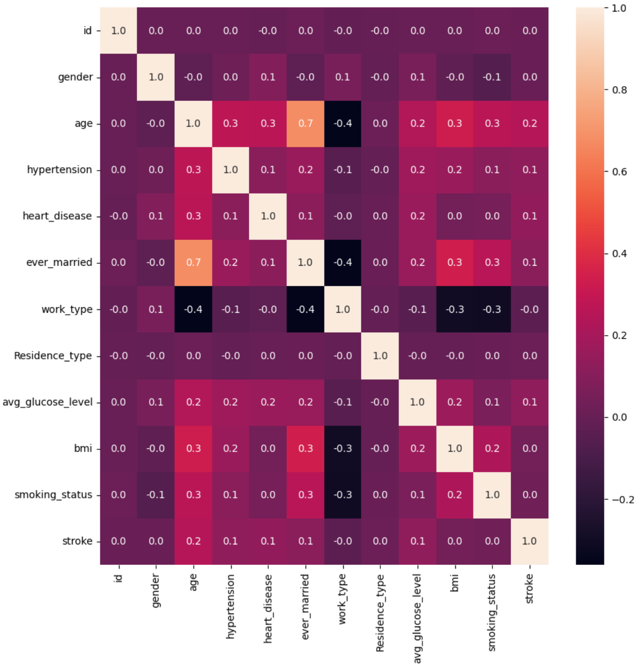


FIGURE 3. Correlation heat map

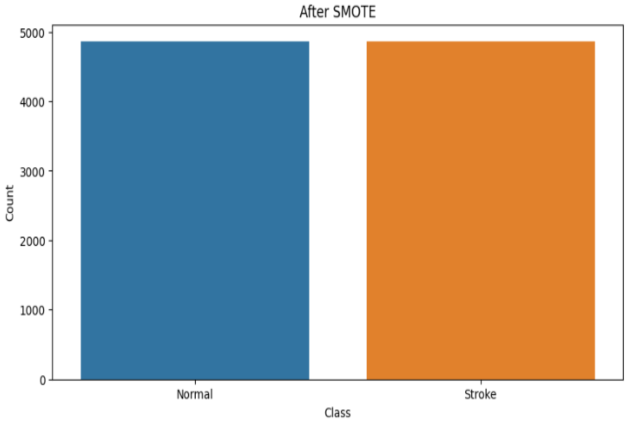


FIGURE 4. Before SMOTE

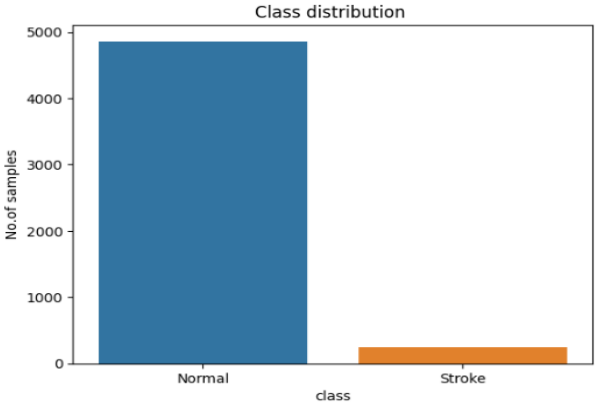


FIGURE 5. After SMOTE

CATEGORICAL TO NUMERICAL

The features like ‘work_type’, ‘ever_married’, ‘gender’, ‘smoking_status’ , ‘Residence_type’, are having categorical values. Most of the machine-leaning models can work only with numbers so the categorical values are converted into numerical values using a technique called LabelEncoder.

NORMALIZATION

Normalization is the process of changing the numerical value in a dataset to a common scale and to make it closer to normal distribution. There are different normalization methods available such as *MinMaxScaler*, *RobustScaler*, *StandardScaler*, and *Normalizer.* The features ‘age’, ‘avg_glucose_level’, and ‘bmi’ have numerical values with different scales. In the stroke prediction dataset, the features are normalized using *MinMaxScaler.* A Min-Max scaling is typically done via the Eq. (1):

$$X_{sc} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

where X is the attribute data, X_{min}, X_{max} are the minimum and maximum absolute value of X respectively.

METHODOLOGY

We compare the effectiveness of different algorithms using Machine learning with ensemble techniques and deep learning techniques.

LOGISTIC REGRESSION

Logistic regression constitutes a statistical framework to build a binary or linear classification model. The logistic regression model calculates the Likelihood of an event taking place, such as encountered stroke or not derived from a provided dataset of variables that are not dependent on each other. Since the output is a probability, the dependent variable range is from 0 to 1. It uses the sigmoid function to estimate the probability as given Eq. (2).

$$g(z) = \frac{1}{1+\exp(-z)} \qquad (2)$$

$where\ z = (a + bX)$, a and b are the parameters of the model.

SUPPORT VECTOR MACHINES

SVM is used for both classification and regression. The main principle of SVM is to find the best optimal hyperplane in N-dimensional space that can separate the data points in different classes in feature space. But if the dataset contains more noise, SVM does not perform well.

DECISION TREES

The decision tree is a hierarchical model used in decision support. Nodes are the attribute test condition and leave nodes are classes. The instances are classified by moving down the tree from root to leave. The most popular criteria used for selecting attributes as the splitting condition is Gini and entropy. ID3, c4.5, and CART are widely used decision tree algorithms.

RANDOM FOREST

Random forest is an ensemble classification technique that combines the predictions of multiple decision trees using voting or averages of all the predictions. This model will perform better than the single decision tree.

NAIVE BAYES

It is the most popular machine learning classification model. It uses Bayes' theorem of probability to predict the class labels as given in Eq. (3).

$$P\left(\frac{c}{x}\right) = \frac{P\left(\frac{x}{c}\right)p(c)}{p(x)} \qquad (3)$$

where, p(x/c) is the likelihood, p(c) is the class prior probability and p(x) is the predictor prior probability.

K-NEAREST NEIGHBOR

K-Nearest Neighbor plots all the instances in a n-dimensional space. It classifies the new data point by calculating the similarity measure eg. Euclidean distance as given in Eq. 4. Classification is done by taking a majority vote of the *k* nearest neighbor.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (4)$$

where, p and q are the points in 2D space.

EXTRA TREE CLASSIFIER

Extra Tree Classifier is also an ensemble model very similar to a random forest.

XGBOOST

Similar to Random Forest, XGBoost is also an ensemble model that uses more than one decision tree. The gradient is used to minimize the loss function during the weight calculation. This model is best to reduce overfitting and improve generalization and performance.

ENSEMBLE MACHINE LEARNING MODELS

Ensemble is the process of combining the prediction of multiple models/classifiers to improve the accuracy of the model. The ensemble can be built from a sample type of classifier or from different type of classifier. In the case of Random Forest, XGBoost, Extra Tree Classifier, the base basic classifier used is the decision tree. There are three types of ensemble models as shown in Figure. 6.

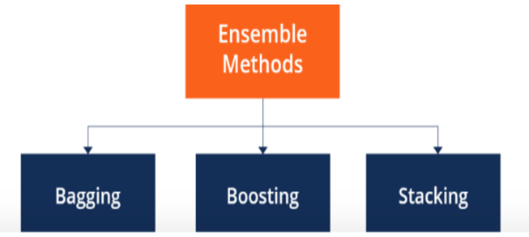


FIGURE 6. Types of Ensemble Learning

BAGGING

Bagging method used bootstrapping sampling techniques to create training samples called as bags for the models in ensemble learning as shown in Figure 7. Any machine-learning model can be selected to create the ensemble model. Bagging is mainly used to reduce the variance of the model due to the missing values in the dataset.

BOOSTING

In the Boosting technique, the same learning samples are used to train more than one classifier. However, the incorrectly classified instances are given as input to the next classifier to correct the weight assigned by the classifier to make the right prediction. Thus, the classifier learns the mistakes in each level and finally, the prediction are improved as shown in Figure 8.

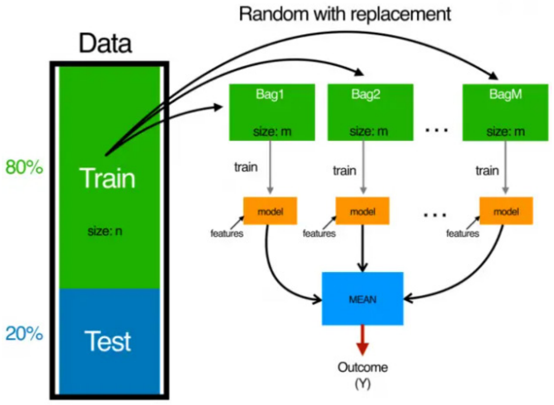


FIGURE 7. Bagging

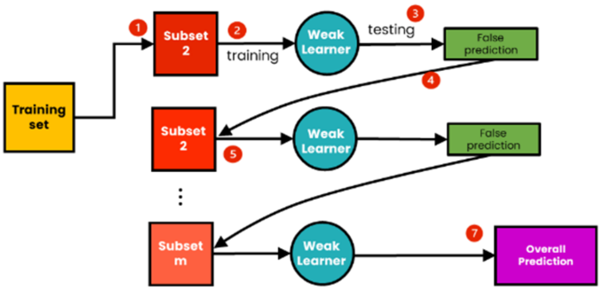


FIGURE 8. Boosting Ensemble

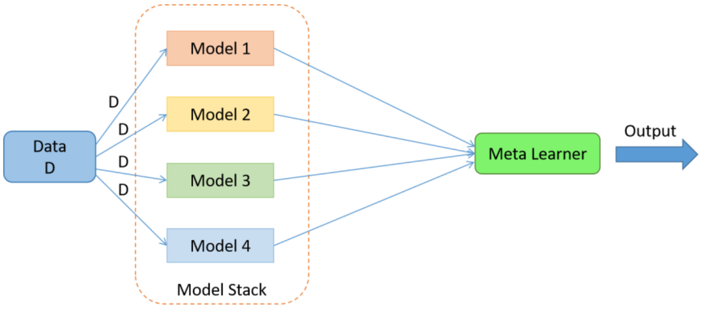


FIGURE 9. Stacking Ensemble

STACKING ENSEMBLE

In stacking ensemble, classification happens in two levels. Level 0 will have some machine learning models like Model 1, Model 2, Model 3, Model 4 as shown in figure 9. The prediction made by the model in level 0 is combined by the meta-learner in level 1. How best to combine the predictions of the model is decided by the meta-learner that is available in level 1.

NEURAL NETWORK MODEL

Finally, artificial neural network model is built to perform classification. The network has three layers:

1. Input layer
2. Hidden layer
3. Output Layer

The model architecture is given in the Figure 10

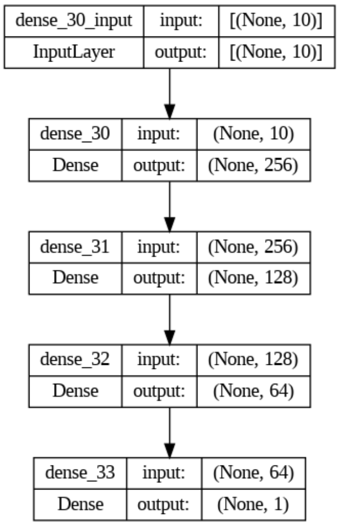

dense_30_input | input: | [(None, 10)]
InputLayer | output: | [(None, 10)]
dense_30 | input: | (None, 10)
Dense | output: | (None, 256)
dense_31 | input: | (None, 256)
Dense | output: | (None, 128)
dense_32 | input: | (None, 128)
Dense | output: | (None, 64)
dense_33 | input: | (None, 64)
Dense | output: | (None, 1)

FIGURE 10. ANN Model Summary

MODEL EVALUATION

The machine learning, ensemble, and deep learning models are evaluated using various metrics such as accuracy, precision, recall, and f1-score as given the Eq. 5 – Eq. 8.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

$$F1 - Score = 2\,\frac{recall*precision}{recall+precision} \tag{8}$$

RESULTS

The experiment was conducted in Google Colab Pro with Python 3 Google Compute Engine backend (GPU –A100) with 40GB GPU RAM. The dataset is divided into training and testing in a ratio of 80% and 20% respectively. The performance of individual machine learning models are given in Table1 and Figure 11. Bagging classifier is created with XGBoost classifier as the base model with no_of estimator as 40. Voting classifier is created with four base classifier such as ExtraTree, RandomForest, XGBoost and Decision Tree classifiers with hard and as well as soft voting. In stacked classifier first level includes base estimator such as ExtraTree, RandomForest, XGBoost and Decision Tree classifiers. The performance metric for all the ensemble models are given in Table 2 and Figure. 13. Finally, ANN model is created with 3 hidden layers/dense layer with 256, 128 and 64 neurons and an output layer with one neuron and the performance of the model is measure by training the model to 500 epochs. The training and validation accuracy and loss are given in the Figure 14.

TABLE 1. Performance Metrics of Traditional Machine – Learning Model

	Accuracy	Precision	Recall
Logistic Regression	0.8010	0.8289	0.7844
Support Vector Machine	0.8031	0.8340	0.7847
Decision Trees	0.9075	0.9258	0.8926
Naïve Bayes	0.7959	0.8773	0.7538
K-Nearest Neighbor	0.8725	0.9423	0.8264

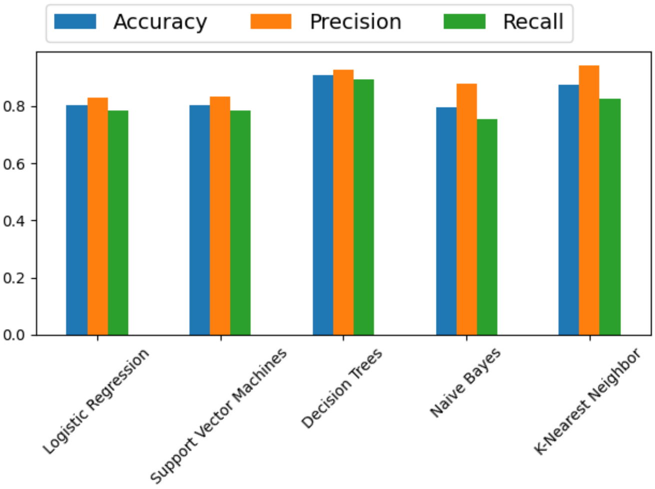


FIGURE 11. Performance Metrics of Traditional Machine-learning Model

TABLE 2. Performance Metrics of Ensemble Models

	Accuracy	Precision	Recall
Random Forest	0.9517	0.9660	0.9389
Extra Tree Classifier	0.9522	0.9691	0.9372
XGBoost	0.9542	0.9629	0.9463
Bagging	0.9506	0.9649	0.9379
Soft Voting	0.9558	0.9711	0.9420
Hard Voting	0.9568	0.9557	0.9576
Stacked	0.9635	0.9629	0.9639

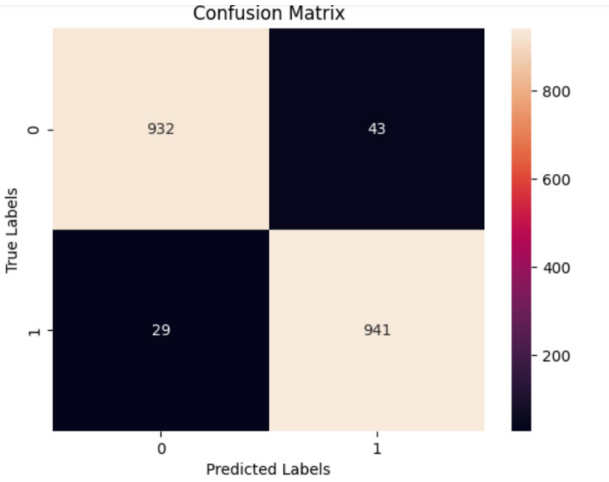


FIGURE 12. Confusion Metrics of Stacked Ensemble Model

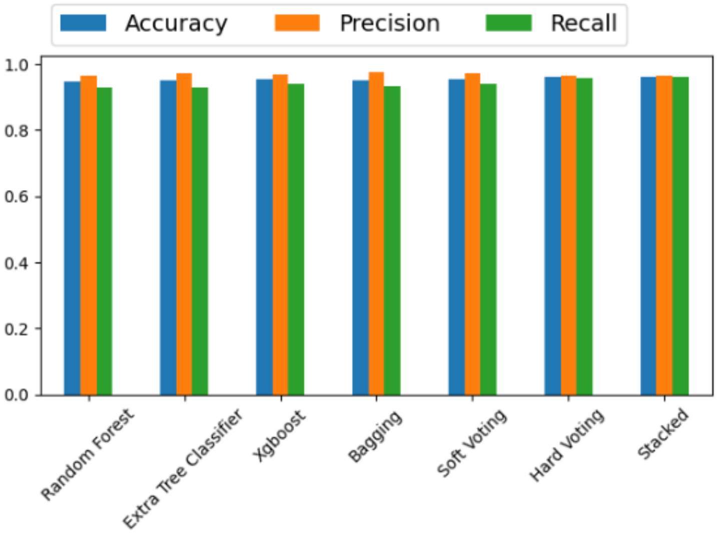


FIGURE 13. Performance Metric of Ensemble Models

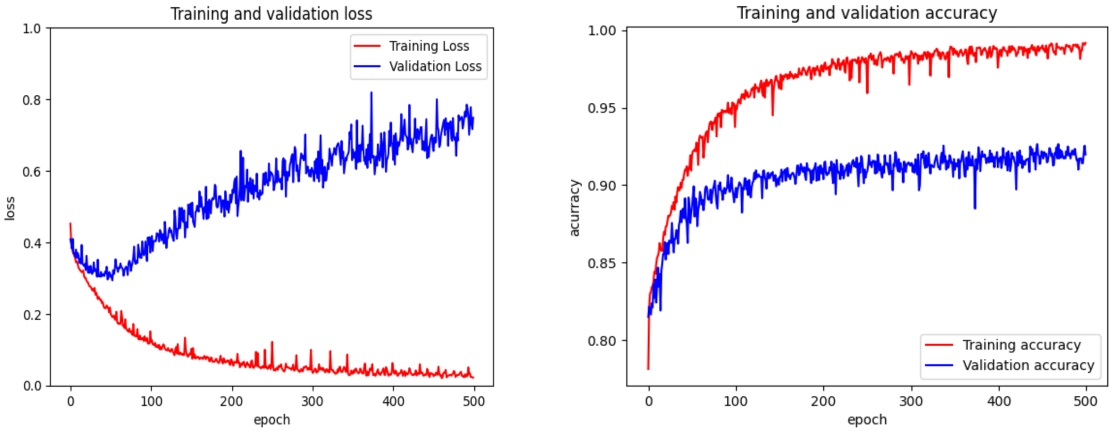


FIGURE 14. Training and Validation Accuracy and Loss

TABLE 3. Comparison of Proposed Stacked Ensemble and Recent Ensemble Models

Methodology and Reference	Accuracy
RF, DT and NB [Rehman, 2023]	94.781%
RXLM stacked Ensemble [Alruily, 2023]	96.34
Xtreme Gradient Boost [Alanazi, 2021]	96%
PCA-Kmeans + DNN + Focal Loss[Yuru, 2022]	92%
Proposed Stacked Ensemble	96.35%

The proposed stacked ensemble model created with XGBoost, Extra Tree Classifier, Random Forest, and Decision Tree is compared with other existing models as shown in Table 3. The stacked ensemble model created with RF, DT, and NB models reported in reference [14] achieved an accuracy of 94.781. The ensemble model in [12] achieved an accuracy of 96.34%. The Xtreme Gradient boost classifier in [13] resulted in 96% accuracy. Whether as, [12] reported 92% accuracy on the same dataset. The proposed ensemble model achieves higher performance of around 96.35% which is higher than all the existing models.

CONCLUSION

In this research, we propose a stacked ensemble model using 4 ML models (ExtraTree, RandomForest, XGBoost and Decision Tree) in level 0 and a meta learner (Linear SVC) in level 1 to combine the prediction of the base learners. We found out that stacked ensemble model performs much better than the other traditional machine learning models, ensemble models and ANN models with an accuracy of 96.35%.

ACKNOWLEDGMENT

The authors would like to thank the Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, India for their support.

DECLARATION OF COMPETING INTEREST

None.

REFERENCES

Abd Mizwar A. Rahim , Andi Sunyoto , Muhammad Rudyanto Arief. 2022. Stroke prediction using machine learning method with extreme gradient boosting algorithm. Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer 21(3): 595-606.

Akinbo R. S and O. A. Daramola. 2021. Ensemble Machine Learning Algorithms for Prediction and Classification of Medical Images. Artificial Intelligence. IntechOpen. doi: 10.5772/intechopen.100602.

Alanazi EM, Abdou A, Luo J. 2021. Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models. *JMIR Form Res.* 2;5(12): e23440. doi: 10.2196/23440. PMID: 34860663; PMCID: PMC8686476.

Alruily, M.; El-Ghany, S.A.; Mostafa, A.M.; Ezz, M.; El-Aziz, A.A.A. 2023. A-tuning ensemble machine learning technique for cerebral stroke prediction. *Appl. Sci.* 13: 5047. https://doi.org/10.3390/app13085047

Asghari Varzaneh Z, M. Shanbehzadeh, and H. Kazemi-Arpanahi. 2022. Prediction of successful aging using ensemble machine learning algorithms. BMC Med. Inform. Decis. Mak., vol. 22, no. 1, p. 258.

Dritsas E, Trigka M. 2022. Stroke risk prediction with machine learning techniques. *Sensors* 21;22(13):4670. doi: 10.3390/s22134670. PMID: 35808172; PMCID: PMC9268898.

Džeroski, S, P. Panov, and B. Ženko. 2009. Machine Learning, Ensemble Methods. Encyclopedia of Complexity and Systems Science, New York, NY: Springer New York, pp. 5317–5325.

Mienye, I. D, Sun, and Z. Wang, 2020. An improved ensemble learning approach for the prediction of heart disease risk. *Inform. Med. Unlocked* 20: 100402.

Rehman A, Alam T, Mujahid M, Alamri FS, Ghofaily BA, Saba T. 2023. RDET stacking classifier: a novel machine learning based approach for stroke prediction using imbalance data. *PeerJ Comput Sci.* 21(9):e1684. doi: 10.7717/peerj-cs.1684. PMID: 38077612; PMCID: PMC10703010.

Rosly. R, M. Makhtar, M. Khalid Awang, M. Isa Awang, M. Nordin Abdul Rahman, and H. Mahdin. 2018. Comprehensive study on ensemble classification for medical applications. *Int. J. Eng. Technol*. 7(2.14): 186.

Shilpa k and T. Adilakshmi. 2022. Applying ensemble techniques of machine learning to predict heart disease. In *Proceedings of the International Conference on Cognitive and Intelligent Computing*. Singapore: Springer Nature Singapore.

Singh, M. Sheetal, and Prakash Choudhary. 2017. Stroke prediction using artificial intelligence. 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON). IEEE.

Soumyabrata Dev, Hewei Wang, Chidozie Shamrock Nwosu, Nishtha Jain, Bharadwaj Veeravalli, Deepu John. 2022. A predictive analytics approach for stroke prediction using machine learning and neural networks”, 2772-4425/© 2022 The Author(s). Published by Elsevier Inc.

Srinivas, Joseph Prakash Mosiganti. 2023. A brain stroke detection model using soft voting based ensemble machine learning classifier. *Measurement: Sensors* 29: 100871.

Ushasree D, AV Praveen Krishna, Ch Mallikarjuna Rao, D V Lalita Parameswari. 2023. SPE: Ensemble hybrid machine learning model for efficient diagnosis of brain stroke towards clinical decision support system. *International Journal of Intelligent Systems and Applications in Engineering* 11(1): 339–347.

Yuru Jing, 2022. Machine Learning Performance Analysis to Predict Stroke Based on Imbalanced Medical Dataset. University College London, Gower Street, London, UK, WC1E 6BT.