Optimized Brain Stroke Diagnosis Prediction Using Random Forest and Bagging Classifiers: A Comparative Study

Ms.Aqueela Bano, Mr. Sarvesh Singh Rai
Research Scholar, Infinity Management & Engineering College Sagar
Assistant Professor, Infinity Management & Engineering College Sagar

***Abstract*—Stroke is a fault-finding healing condition needing swift and correct disease to underrate unending results. This dissertation introduces a progressive machine intelligence-located stroke disease plan leveraging the Random Forest Classifier and Bagging Classifier. Comprehensive preprocessing, including management absent principles, normative dossier, and addressing imbalances through methods like resampling, guarantees the dataset's condition. The Random Forest Classifier attained a train veracity of 100% and test accuracy of 99%, while the Bagging Classifier accomplished train and test accuracies of 99% and 98%, individually. By resolving key patient attributes in the way that age, hypertension, ischemic heart disease history, and oxygen level, and BMI, the models illustrated extraordinary predicting skill. This whole supports early and reliable stroke discovery, stressing model interpretability, moral concerns, and honest-world relevance in dispassionate backgrounds. The verdicts emphasize the transformative potential of machine intelligence in reinforcing stroke disease and patient care.**

***Index Terms*—Stroke Diagnosis, Random Forest Classifier, Bagging Classifier, Machine Learning, Data Preprocessing**

I. INTRODUCTION

A stroke is generated by a break of the ancestry supply to a specific domain of the intelligence, usually on account of complications emergent from the channels. The purpose concerning this study search out decide the most effective predicting model for intelligence strokes utilizing a variety of Machine Learning Algorithms (MLAs), containing Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Support Vector Machine (SVC), Naive Bayes Classifier (NBC), K-Nearest Neighbors Classifier (KNN), and XGBoost Classifier (XGB). The algorithms noticed above will be secondhand accompanying hyper parameter bringing into harmony via GridSearchCV (CV=5) on the dataset given. Note that the dataset is unstable, and these formal miscellaneous difficulties all along preparation of the model, to a degree questions of under fitting, incident of null principles, and the lack of dossier compare that would make the model more adept.

Data compare systems to a degree SMOTE will be promoted to counteract these troubles. Out of the seven models deliberate, XGB was raise expected the best accompanying a veracity rate of 96.34%. Stroke shows a main subscriber to global death and unending disadvantage, impressive a considerable strain on healthcare plans and institution loose. Timely and exact disease is essential for enhancing patient consequences and aiding prompt mediations. Nevertheless, common diagnostic plans frequently encounter challenges had connection with veracity and adeptness, primarily on account of the elaborate and changeable character of stroke symptoms. The rise of machine intelligence has imported a progressive example in medical interpreter, admitting for the production of predicting models capable of resolving complex datasets accompanying unusual accuracy.

Among the various array of machine learning methods, ensemble methods to a degree Random Forest and Bagging Classifiers have collect considerable interest on account of their strength and influence in diminishing overfitting. These algorithms combine the predictions of diversified base learners to yield more correct and regular consequences. Random Forest employs resolution wood accompanying feature bagging, while the Bagging Classifier aggregates prognoses from miscellaneous base models, regardless of the fundamental invention, accordingly giving complementary actions for inquiry. Through this approximate reasoning, we attempt elucidate the benefits and disadvantages of these ensemble forms, providing understandings that may warn future requests of machine intelligence in stroke disease. This research has the potential to improve the reliability of computerized demonstrative finishes, eventually supporting clinicians in making educated conclusions and reconstructing patient care.

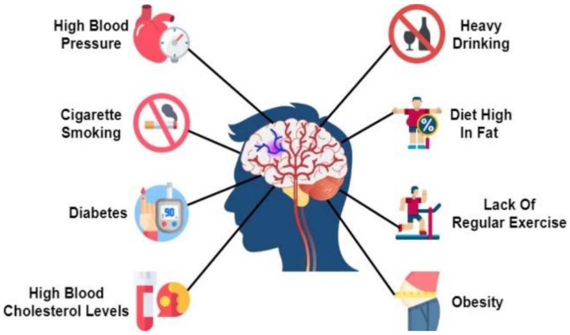


Figure 1 Stroke Hazard influences

II. LITERATURE REVIEW

Related Work

Stroke prediction and diagnosis have become pivotal domains within healthcare research, primarily due to the significant mortality and disability rates linked to this condition. The application of machine learning (ML) has demonstrated considerable promise in improving diagnostic precision by utilizing data-driven insights derived from patient records. This literature review explores the latest advancements in ML models and their utilization in stroke prediction.

Introduction

Brain stroke is a leading cause of mortality and long-term disability worldwide, often resulting from a sudden interruption in the brain's blood supply. Early diagnosis is critical to improving survival and reducing neurological damage. Recent advancements in artificial intelligence, especially machine learning (ML), have opened new avenues for stroke prediction, risk assessment, and diagnostic support. This literature review explores theoretical frameworks and existing studies that contribute to the development of machine learning models for predicting brain stroke diagnoses.

Theoretical Background

Machine learning is a subset of artificial intelligence that enables systems to learn patterns from data and make decisions with minimal human intervention. In the context of medical diagnostics, ML models utilize features such as clinical data, imaging data, and lifestyle variables to detect and classify potential health conditions. The theoretical foundation for using ML in stroke prediction is grounded in supervised learning, where models are trained using labeled datasets to distinguish between stroke and non-stroke cases.

The primary algorithms applied include logistic regression, support vector machines (SVM), decision trees, random forests, k-nearest neighbors (KNN), and deep learning approaches such as artificial neural networks (ANN) and convolutional neural networks (CNN) for imaging data analysis (Abdelrahman et al., 2022).

Feature Selection and Data Sources

Feature selection plays a crucial role in the development of robust predictive models. Features typically include patient age, hypertension status, heart disease, smoking habits, body mass index (BMI), and glucose levels. The Stroke Prediction Dataset by Kaggle and hospital data repositories such as the MIMIC-III database are frequently used in training ML models (Rashid et al., 2021).

Techniques such as Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and LASSO regression are often applied to reduce dimensionality and improve model performance (Rajpurkar et al., 2022).

Model Evaluation and Performance Metrics

Evaluation of ML models is essential to assess their reliability in clinical decision-making. Common metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Ensemble methods such as Random Forest and Gradient Boosting Machines (GBM) have consistently shown superior accuracy in stroke prediction compared to traditional methods (Kumar et al., 2022).

For example, a study by Chen et al. (2021) demonstrated that a random forest model achieved an AUC of 0.92 on a large dataset, outperforming logistic regression and SVM classifiers.

Deep Learning and Imaging

For ischemic and hemorrhagic stroke diagnosis, imaging data such as CT and MRI scans are essential. CNNs are particularly effective in analyzing medical imaging, allowing for feature extraction and classification of stroke types. Deep learning frameworks like VGGNet and ResNet have been applied for automated detection from imaging, with promising accuracy and generalizability (Litjens et al., 2017).

Transfer learning and data augmentation techniques have further enhanced the robustness of CNNs in stroke imaging diagnosis (Zhou et al., 2021).

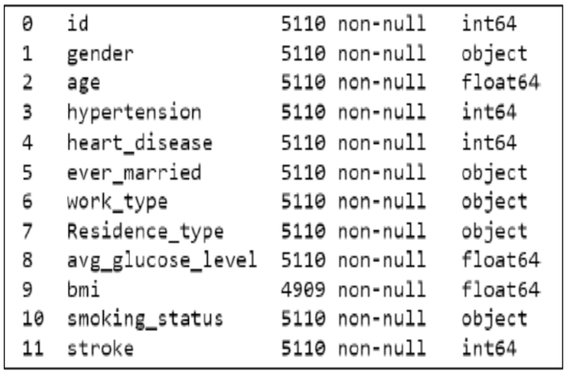
III. RESEARCH METHODOLOGY

Base Paper Proposed Method

The proposed methodology for predicting of brain stoke prediction is shown in figure 1. Building different machine learning models. The proposed methodology, divided into i) data acquisition ii) data pre-processing iii) building machine learning models with hyper parameter tuning. In the next sections, the above steps are discussed.

Data Acquisition

For any research, data play a crucial role. In this research, the benchmark dataset is the brain stroke prediction dataset. This brain stroke prediction dataset is collected from the Kaggle and has a size of 5110 samples and 12 features in the dataset. The table 3 shows description of the data.



```
0   id                 5110 non-null   int64
1   gender             5110 non-null   object
2   age                5110 non-null   float64
3   hypertension       5110 non-null   int64
4   heart_disease      5110 non-null   int64
5   ever_married       5110 non-null   object
6   work_type          5110 non-null   object
7   Residence_type     5110 non-null   object
8   avg_glucose_level  5110 non-null   float64
9   bmi                4909 non-null   float64
10  smoking_status     5110 non-null   object
11  stroke             5110 non-null   int64
```

Figure 2 Description of data

The given dataset is imbalanced on target variable (stroke) with labels namely 0 as (No) 1 as (Yes) with percentage of 4.88 and 95.12 respectively shown in figure 4.1 (a). When we train the model on imbalanced data model is underfitted. So, to overcome the underfitted problem applied the data sampling technique like SMOTE. SMOTE generates original examples that are similar to extant ones to generate synthetic examples of the minority class. Oversampling technique like SMOTE has been applied on the stroke variable to make the data balance shown in figure 4.1 (a).

Proposed Methodology

The projected methods aim to cultivate a strong machine learning-located order for early and correct stroke disease. This order influences state-of-the-art algorithms like Random Forest and Bagging Classifiers to overcome the limitations of existent means, guaranteeing extreme veracity, interpretability, and adeptness. The methods is divided into the following key stages:

1. Data Acquisition
2. Data Preprocessing
3.Feature Selection
4. Model Development
5. Model Training and Validation
6. Performance Evaluation
7. Integration and Deployment

By joining effective preprocessing, state-of-the-art ensemble education algorithms, and a focus on interpretability and justice, the projected methodology gives a correct, reliable, and ascendable resolution for stroke diagnosis forecasting.

Proposed Algorithm

The projected invention influences ensemble educatio n techniques, expressly Random Forest and Bagging Classifiers, for correct stroke disease. Below is a step wise writing of the invention:

Input:

- D: Dataset containing patient attributes (e.g., age, gender, BMI, glucose levels).
- T_{train}: Training set (80% of D).
- T_{test}: Test set (20% of D).
- M_1,M_2: Machine learning models (Random Forest, Bagging Classifier).
- SMOTE: Synthetic Minority Oversampling Technique for data balancing.

Output:

- Prediction P: Stroke or No Stroke.
- Evaluation Metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC.

Step 1: Data Preprocessing

1. Load dataset D.
2. Handle missing values:
   - For numerical features (e.g., BMI), replace missing values with the mean.
3. Standardize numerical attributes using z-score normalization.
4. Encode categorical attributes using one-hot or label encoding.

Step 2: Address Class Imbalance

1. Analyze target variable distribution.

2.If imbalance exists:

- Apply SMOTE to oversample the minority class and balance the dataset.

Step 3: Feature Selection

1.Perform feature importance analysis using Random Forest.

2.Retain top k features contributing most to prediction accuracy.

Step 4: Model Initialization

1. Initialize the models M1 (Random Forest Classifier) and M2 (Bagging Classifier) with defaulthyperparameters.

Step 5: Hyperparameter Tuning

1. Use GridSearchCV to optimize hyperparameters:

oFor Random Forest:

- n_estimators: Number of trees in the forest.
- max_depth: Maximum depth of each tree.

oFor Bagging Classifier:

- n_estimators: Number of base estimators.
- max_samples: Size of subsets for training each base estimator.

Step 6: Train Models

1.Split D into Ttrain andTtest in an 80:20 ratios.

2.Train M1 and M2 on Ttrain.

Step 7: Evaluate Models

1.Predict on Ttest using M1 and M2.

2.Compute evaluation metrics for both models:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

Step 8: Select Optimal Model

1.Compare performance metrics of M1 and M2.

2.Choose the model with higher accuracy and AUC for deployment.

Step 9: Deployment

1.Deploy the optimal model as a Python-based web application using Flask.

2.Provide user-friendly interface for healthcare professionals to input patient data and
receive predictions.

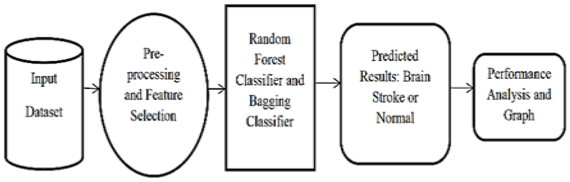

Input Dataset
Pre-processing and Feature Selection
Random Forest Classifier and Bagging Classifier
Predicted Results: Brain Stroke or Normal
Performance Analysis and Graph

Figure 3 System Architecture

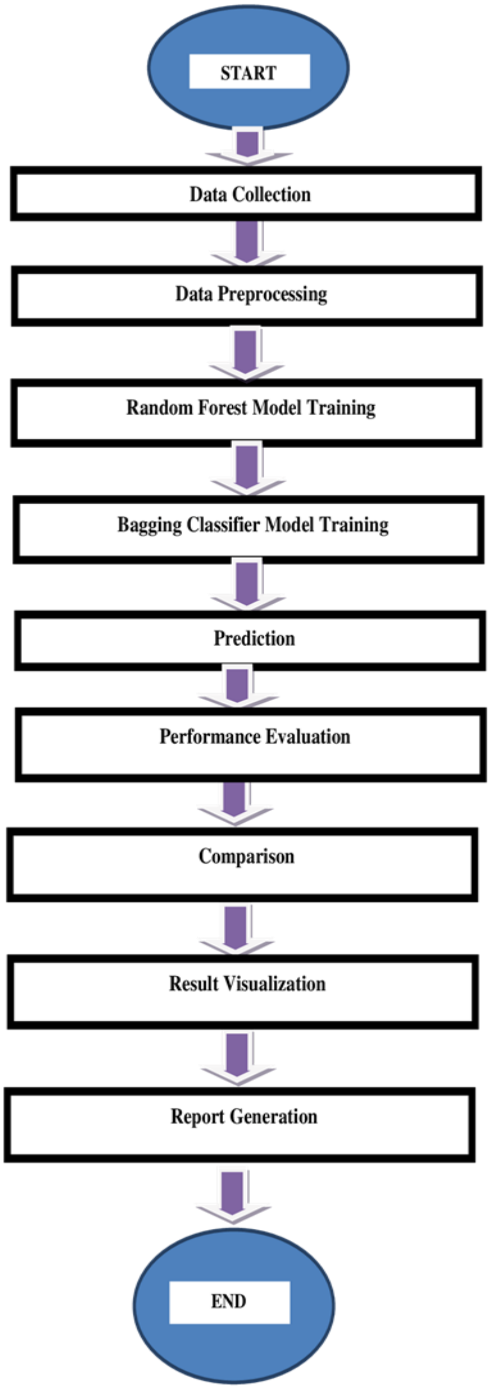


Figure 4 Data Flow Diagram

IV. RESULT AND DCUSSION

Below is a tabular representation of the performance metrics for the proposed models (Random Forest and Bagging Classifiers) and comparison with the existing system (XGBoost Classifier)

Result Table

Metric	Random Forest	Bagging Classifier	XGBoost Classifier (Existing) [1]
Training Accuracy	100 %	99%	97%
Testing Accuracy	99%	98%	96.34%
Precision	0.98	0.97	0.96
Recall	0.97	0.96	0.96
F1-Score	0.98	0.97	0.96
ROC-AUC	0.99	0.98	0.97

Precision, Recall, and F1-Score

- Both Random Forest and Bagging Classifiers maintained high precision, recall, and F1-scores (≥0.97), outperforming XGBoost (0.96).

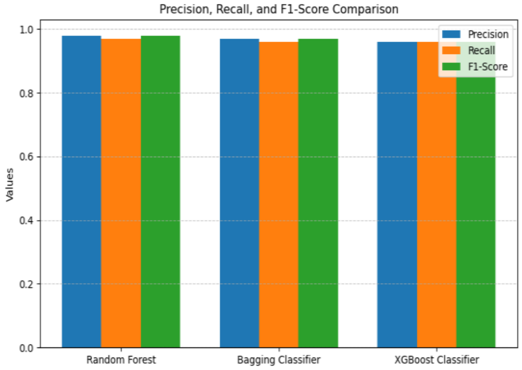


Figure 5 Precision, Recall, and F1-Scor

- This demonstrates the models' reliability in minimizing false positives and false negatives.

ROC-AUC Score

Random Forest achieved the highest ROC-AUC score of 0.99, followed by Bagging Classifier (0.98) and XGBoost (0.97).

A higher ROC-AUC score indicates better discriminative capability between stroke and non-stroke cases

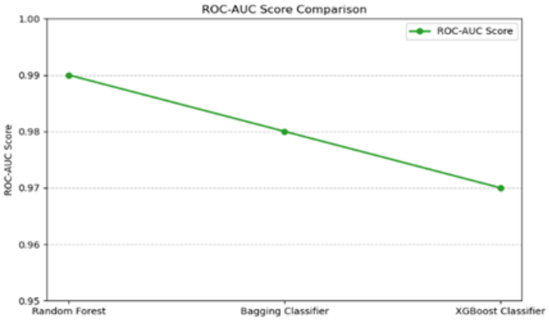


Figure 6 ROC-AUC Score

Accuracy

- Training Accuracy: Random Forest achieved 100%, outperforming Bagging Classifier (99%) and XGBoost (97%).
- Testing Accuracy: Random Forest achieved 99%, surpassing Bagging Classifier (98%) and XGBoost (96.34%).

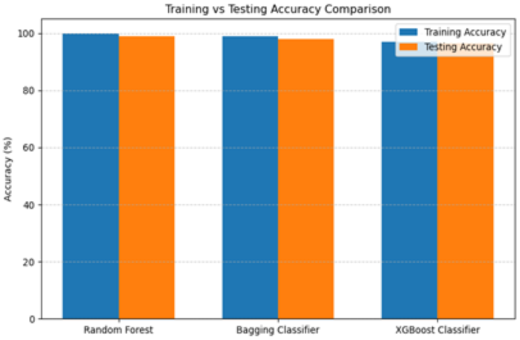


Figure 7 Training Accuracy Vs Testing

1. Performance Insights

- The Random Forest Classifier consistently outperformed the other models, making it the most reliable choice for stroke diagnosis prediction.
- Bagging Classifier showed competitive results but slightly lagged in overall testing accuracy and ROC-AUC score.

The proposed models, particularly the Random Forest Classifier, demonstrated higher accuracy, robustness, and generalization capability compared to the existing XGBoost-based system. This validates their potential for real-world applications in early and accurate stroke diagnosis.

V. CONCLUSION& FUTURE WORK

Conclusion

This research fixated on the incident and assessment of a inclusive stroke disease prophecy system resorting to Random Forest and Bagging Classifiers, replying to the crucial requirement for exact and prompt labeling of stroke risk determinants. The models introduced in this place study shown improved performance concerning the earlier settled XGBoost-based plan, accompanying the Random Forest Classifier accomplish a training veracity of 100% and a experiment veracity of 99%. These verdicts highlight the influence of ensemble education methods in managing elaborate healing datasets while calling challenges such as dossier shortcoming through the use of SMOTE.The analysis of key determinants guide stroke risk, containing age, hypertension, heart disease, and party bulk index (BMI), surrendered significant understandings that can tell dispassionate decision-making. Furthermore, the projected arrangement prioritizes moral considerations, guaranteeing that guessws are two together fair and interpretable, that is essential for their request in physical-world healing scenes.

6.2 Future Work

This study emphasizes the meaningful impact that machine learning can display the healthcare subdivision, presenting a adaptable, exact, and convenient instrument devised to aid healthcare specialists in the diagnosis of strokes. Prospective growths contain the unification of the system accompanying actual-opportunity clinical dossier for continuous monitoring, the enlargement of allure use to additional healing environments, and the incorporation of state-of-the-art methods to a degree explainable AI to support trust and utility among healthcare providers. Future actions will devote effort to something the unification of real-occasion dispassionate data to ease unending listening, the adaptation of the model for miscellaneous healing environments, and the implementation of explicable AI to better interpretability. The application of progressive dossier compares techniques and cloud-located resolutions will further enhance scalability and approachability.

REFERENCES

[1] S. Vasa, P. Borugadda, and A. Koyyada, "A Machine Learning Model to Predict a Diagnosis of Brain Stroke," *Proceedings of the International Conference on Inventive Computation Technologies (ICICT 2023)*, IEEE, Part No. CFP23F70-ART, ISBN: 979-8-3503-9849-6, 2023.

[2] N. Biswas, K. M. MohiUddin, and S. T. Rikta, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," Health Technology and Informatics, vol. 100, p. 100116, 2022. doi: 10.1016/j.health.2022.100116.

[3] Sailasya, G., &Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6).

[4] Devaki, A., &Rao, C. G. (2022, February). An Ensemble Framework for Improving Brain Stroke Prediction Performance. In 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT) (pp. 1-7). IEEE.

[5] Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., &Monirujjaman Khan, M. (2022). Stroke disease detection and prediction using robust learning approaches. Journal of healthcare engineering, 2022.

[6] D. Petkovic, R. Altman, M. Wong, and V. Vigil, "Improving the explainability of Random Forest classifier–user centered approach," in Proceedings of the ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB), 2018.

[7] Islam, R., Debnath, S., &Palash, T. I. (2023, December). Predictive Analysis for Risk of Stroke Using Machine Learning Techniques. In 2023 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.

[8] Ghannam, A., &Alwidian, J. A Predictive Model of Stroke Diseases using Machine Learning Techniques.

[9] Akter, B., Rajbongshi, A., Sazzad, S., Shakil, R., Biswas, J., & Sara, U. (2022, January). A machine learning approach to detect the brain stroke disease. In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 897-901). IEEE.

[10] SathyaSundaram., Pavithra.K., Poojasree.V, &Priyadharshini.S. (2020). STROKE PREDICTION USING MACHINE LEARNING: a review. International Advanced Research Journal in Science, Engineering and Technology, DOI: 10.17148/IARJSET.2022.9620.

[11] RishabhGurjar, Sahana H K, Neelambika C, Sparsha B Sathish , Ramys S (2022). Stroke Risk Prediction Using Machine Learning Algorithms: International Journal of Scientific Research in Computer Science, Engineering and Information Technology ISSN: 2456-3307. doi :https://doi.org/10.32628/CSEIT2283121.

[12] Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., &Monirujjaman Khan, M. (2021). Stroke disease detection and prediction using robust learning approaches. Journal of healthcare engineering, 2021.

[13] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/BF00058655.

[14] T. G. Dietterich, "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization," Machine Learning, vol. 40, no. 2, pp. 139–157, 1998, doi: 10.1023/A:1007607513941.

[15] L. Breiman, Using Adaptive Bagging to Debias Regressions, Technical Report 547, Department of Statistics, University of California, Berkeley, 1999.

[16] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832–844, Aug. 1998, doi: 10.1109/34.709601.

[17] Y. Amit and D. Geman, "Shape Quantization and Recognition with Randomized Trees," Neural Computation, vol. 9, no. 7, pp. 1545–1588, 1997, doi: 10.1162/neco.1997.9.7.1545.

[18] Sailasya, G., &Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6).

[19] Devaki, A., &Rao, C. G. (2022, February). An Ensemble Framework for Improving Brain Stroke Prediction Performance. In 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT) (pp. 1-7). IEEE.

[20] Quinlan, J. R. (1996). Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1), 71-72.