Machine Learning Based Stroke Detection: A Predictive Approach

Akinsola Adeniyi F.[1].; Sokunbi. M. A.[2]; Ogundele. I. O.[3]; Onadokun I. O.[4]

[1] Yaba College of Technology, Computer Technology Dept.,
Yaba, Lagos Nigeria.

[2] Yaba College of Technology, Computer Technology Dept.,
Yaba, Lagos Nigeria.

[3] Yaba College of Technology, Computer Technology Dept.,
Yaba, Lagos Nigeria.

[4] Yaba College of Technology, Computer Technology Dept.,
Yaba, Lagos Nigeria.

Publication Date: 2026/03/13

Abstract: Stroke is one of the biggest challenges facing the world's public health today and is ranked as the second most common cause of death and the third most common cause of long term disability globally [1]. Early diagnosis is important for lowering the risk of death as well as long-term disability from stroke. Traditional methods of diagnosing stroke, such as CT scans and MRI's, however, can be expensive, time-consuming and require specialists to interpret them. The challenges associated with these traditional methods of diagnosis can delay the making of decisions regarding treatment, especially in low-resource settings in which there is a lack of access to advanced imaging technologies or qualified personnel. Due to this, new approaches are now being studied in an effort to identify stroke rapidly and efficiently. The stroke detection model was developed based on the machine learning application to structured data from patients. The four supervised learning methods used were: LightGBM, CatBoost, XGBoost, and Random Forest. These methods were applied using a 70/30 split for the training and testing set. Accuracy, Precision, Recall, F1-Score, and the Area Under Curve Receiver Operating Characteristic Curve (AUC-ROC) were all used as measures of model performance. Of the models compared in the study, the Random Forest method demonstrated superior model performance with an accuracy of 90% and an F1-Score of 0.95. Additionally, the Random Forest model was able to achieve higher performance than gradient boosting methods for each of the most important performance metrics. These results indicate that machine learning algorithms particularly those based on ensemble techniques (such as Random Forest) can potentially be used to enhance current diagnostic pathways for predicting stroke by providing faster, more scalable and more accessible predictions of stroke risk than are currently available, which could provide the opportunity for earlier clinical interventions and better patient outcomes, specifically in low resource health care settings. Machine learning tools should not be used to supplant clinical expertise; however, if integrated into routine practice, they could mark an important step toward using data more effectively and equitably when caring for patients with stroke.

Keywords*: Stroke Detection, Machine Learning, Random Forest, Predictive Modeling, Healthcare AI.*

How to Cite: Akinsola Adeniyi F.; Sokunbi. M. A.; Ogundele. I. O.; Onadokun I. O. (2026) Machine Learning Based Stroke Detection: A Predictive Approach. *International Journal of Innovative Science and Research Technology*, 11(2), 3159-3165. https://doi.org/10.38124/ijisrt/26feb1439

I. INTRODUCTION

The World Health Organization (WHO) recognizes stroke as a significant global public health challenge, and stroke continues to be one of the most serious forms of non-communicable disease. According to recent data from the World Stroke Organization (WSO), Global Stroke Fact Sheet 2025, stroke was the second leading cause of death globally and the third leading cause of disability when considering disability-adjusted life-years (DALYs) [1]. Additionally, the global burden of stroke has dramatically increased since 1990. In fact, between 1990 and 2021, there was a 70% increase in the number of new incident stroke cases; stroke-related deaths increased by 44%; the number of people with stroke increased by 86%; and DALYs associated with stroke increased by 32% [1] [2]. These trends demonstrate a persistent increase in both the occurrence of stroke and its long term impact.

The total amount of money that can be attributed to stroke as an economic burden, is as large as the money lost due to loss of work productivity and medical costs of strokes. Annually, this loss exceeds $890 billion and represents about .66% of the world's GDP. This loss is concentrated in lower and middle income countries and accounts for almost 87% of all stroke related deaths and 89% of all DALYs lost around the globe. The reasons for these inequalities are a direct result of unequal access to timely diagnosis and treatment, and also, equal or greater inequality to preventives; therefore, there is a dire need for affordable and easily scalable methods for detecting strokes, especially in developing health care systems.

Although CT and MRI are the clinical "gold standard" for diagnosing a stroke, they can be expensive, require specialized equipment and experts to interpret, and provide slow diagnoses due to the time it takes to obtain them. These factors limit the ability of these tests to assist with early diagnosis and timely treatment of stroke victims [3]. For that reason, researchers and clinicians are exploring the use of ML-based models as a way to rapidly diagnose and predict stroke risk based upon standardized health data routinely collected from electronic health records.

We compare the performance in terms of predictive accuracy of four supervised machine learning algorithms (Random Forest, LightGBM, CatBoost and XGBoost), as a first step towards identifying which one may be most suitable for inclusion within an intelligent health care support system to enable early detection of stroke and ultimately lead to better patient outcomes.

II. REVIEW OF RELATED WORK

In response to the rising global burden of stroke there is a considerable amount of literature on the potential of Machine Learning (ML) and Deep Learning (DL) in the early detection of stroke, the assessment of stroke risk and the prediction of outcomes from stroke events. There are now numerous studies indicating that by developing predictive models based upon the data collected, clinicians can utilize such predictive models to aid in their decision making and to enhance the accuracy of diagnoses made in a variety of different healthcare environments.

Several previous studies have investigated the utility of Classical Machine Learning (ML) techniques for analyzing Structured Clinical Data. For example, Alageel et al.[5] investigated seven different types of Machine Learning (ML) models by utilizing Electronic Health Records (EHRs). The results were that the Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Decision Trees achieved the best predictive results for the identification of stroke. Additionally, Rahman et al.[6] compared Machine Learning (ML) and Deep Learning (DL) models for the prediction of stroke and concluded that classical Machine Learning models consistently performed better than deep Neural Networks.

In contrast to some previous studies, which utilized various deep learning architectures to identify temporal patterns in patient data, several researchers like Kaur et al. [7] have explored the use of Recurrent Neural Network (RNN) Models for Early Stroke Detection. Specifically, the authors employed Long Short-Term Memory (LSTM) Networks, Bidirectional LSTM (biLSTM) Networks, Gated Recurrent Units (GRUs), and Feed Forward Neural Networks (FFNNs). Results indicated that GRU-based models achieved the highest level of accuracy at 95.6%. Thus, it appears that Temporal Modeling can be particularly effective when a patient's longitudinal data is available.

In addition to utilizing structured Clinical Records, some studies have also included Biomedical Signals and Medical Imaging. For example, Pitchai et al. [8] developed an Artificial Intelligence (AI)-based stroke prediction model using Synthetic Electromyography (EMG) signals, along with an SVM Classifier. The authors reported a high degree of accuracy, or 92.6%. Unfortunately, the subsequent retraction of this study by the journal because of its lack of replicability highlights the problems associated with validating data and models in AI-Driven Healthcare Research. Dritsas and Trigka [4], however, found that classical Machine Learning (ML) Techniques, such as Logistic Regression, Naive Bayes, K-Nearest Neighbor (KNN), and Ensemble Methods such as Random Forest, could be used to achieve strong levels of performance.

Deep Learning Models are being used in Imaging-Based Stroke Analysis. Alotaibi et al.[9] utilized a Convolutional Neural Network (CNN)-Based Architecture, which had Recurrent Layers (CNN-LSTM and CNN-BiLSTM), to Analyze MRI Scans for Ischemic Stroke Tissue Fate Detection, they achieved an accuracy of 89.2%. While these imaging-based methods appear to be promising, they do present limitations of advanced imaging technologies in terms of potential scalability limitations for use in resource-constrained environments.

In addition to Classification Tasks, Predictive Analytics Studies have highlighted the need for Identifying Key Stroke Risk Factors. Dev et al. [10] Combined Statistical Analysis with Neural Networks to Confirm Age, Hypertension, Heart Disease and Glucose Levels as Significant Indicators of Stroke Occurrence. Building on this, Alanazi et al. [11] Utilized Laboratory Test Results to Classify Patients into Categories of Stroke Risk Based on Symptom Severity and Frequency. In the area of Outcome Prediction, Heo et al. [12] showed that Deep Neural Networks were capable of Predicting Long-Term Ischemic Stroke Outcomes with an AUC of 0.888; better than both Random Forest and Logistic Regression Models.

Collectively, these studies demonstrate that Random Forest — in particular — is effective at providing good results across a wide variety of tasks of predicting strokes from various types of data including both clinical and imaging.

However, most of these studies are focused on using highly specialized data sets (such as MRI or EMG) and/or only one specific risk factor for stroke prediction. In addition to these limitations, few of these studies have demonstrated how their predictive performance can be optimized across clinically meaningful metrics (such as recall, F1-score, and AUC-ROC).

The current study addresses these issues through the development of a generalizable machine learning based framework for detecting stroke that uses four robust algorithms: Random Forest, LightGBM, CatBoost, and XGBoost.

The results show that Random Forest is significantly better than all other models tested in this study, with an accuracy of 90%, an F1-score of 0.95, and an AUC-ROC of 0.95.

Therefore, the results indicate that Random Forest offers a reliable, easily accessible, and scalable method for stroke detection that balances high predictive performance with the practicality of being able to use it in a clinical setting.

III. METHODOLOGY

A. Dataset

This research will use two main data sources to provide the foundation for developing and evaluating the proposed stroke prediction models. First, the Kaggle Stroke Prediction Data Set, which is publicly available and contains thousands of structured records regarding patients; the Kaggle Data Set includes demographic, clinical, and lifestyle characteristics of patients such as age, gender, hypertension status, heart disease history, marital status, work type, blood sugar levels, smoking status, BMI and type of home where they reside. Due to its size and diversity, it can be viewed as a good representation of global stroke risks and has been used by many researchers using machine learning.

To complement this global dataset, a second dataset was acquired from selected federal hospitals in Lagos State Nigeria. These records capture region-specific demographic characteristics, comorbidity profiles and health patterns that may not be fully reflected in international datasets. To improve the external validity of the models and ensure that predictive performance more accurately reflects real-world conditions within the Nigerian healthcare context, locally sourced clinical data were included in the model development process.

The data sets were combined to create a greater representation of stroke risk factors that is contextually appropriate. The combined dataset contains approximately 5000 patients' health profiles; every row is one patient's health record. The variables that are recorded as part of this health profile contain all of the well-established clinical risk factors for stroke; they include age, gender, whether or not there is a history of hypertension, whether or not there is a history of heart disease, marital status, type of work, BMI, smoking habits, average blood sugar levels, and whether the person lives in an urban or rural area. These attributes were selected based on clinical evidence to support their use in supervised machine-learning-based stroke prediction models.

Before developing the model, all of the pre-processing that is typically done was completed. Missing data were appropriately imputed. Categorical features were converted from their categorical format to a numerical format so they could be processed by the models. All feature values were then normalized to scale them equally. Finally, a 70% training subset and a 30% test subset of the dataset were created to provide an unbiased measure of how well the model performed.

B. Machine Learning Algorithms Used

To provide a basis for comparison among predictive capabilities of stroke prediction, we employed four state-of-the-art ensemble and gradient-boosting machine learning methods with established applications in structural medical data (Random Forest, LightGBM, CatBoost, XGBoost) as they have demonstrated to effectively address nonlinearities and interaction between features, as well as class imbalances that are common characteristics of most clinical datasets.

➢ *Random Forest*

Random Forest is a technique for creating an ensemble of multiple decision tree models to generate a single classification result by using a voting mechanism where each sample from the data set is classified by all of the constituent models and then the outcome with the most votes is selected as the final classification. The random selection of subsets of both the samples used for training (i.e., the "rows") and the attributes or variables used to train the model (i.e., the "columns") results in a reduction in the variability of the model and minimizes overfitting allowing the model to be generalized to data sets it has never seen before. The model also tends to be robust to noisy data and can perform reasonably well in large dimensionality attribute spaces and can provide the user with intrinsic measures of the importance of the different attributes, a very useful measure when trying to predict disease occurrence in patients such as in the case of stroke detection. In this study, the Random Forest model was able to produce consistent and excellent performance on a variety of evaluation metrics demonstrating the potential for inclusion into clinical decision support systems [13].

- *Light Gradient Boosting Machine (LightGBM)*

LightGBM is an open source gradient boosting framework utilizing tree based machine learning that was created to be as fast as possible with minimal computational overhead. It differs from other traditional gradient boosted methods by using a leaf-wise approach to build the decision tree instead of a level-wise approach. This can result in faster convergence and lower training loss when comparing models. LightGBM was also developed to enable it to operate on very large data sets (big data) and reduce its overall memory footprint at no cost to the model's predictive accuracy. As such, LightGBM can process both numeric and categorical feature types common to many health care related data sets and thus, may be suitable for use within a stroke prediction task [14].

- *CatBoost*

CatBoost is a machine learning (ML) algorithm that was designed to improve upon gradient boosting algorithms to better handle categorical data by using an optimal method of encoding that will help to minimize prediction bias and prevent target leakage. CatBoost also uses an ordered boosting process to provide additional opportunities for improving model generalizability on smaller sample size datasets. Due to these features, CatBoost is ideal for datasets containing many categorical demographic and lifestyle variables (e.g., gender, marital status, smoking habits). The ability of CatBoost to perform consistently well across multiple datasets provides it with sufficient reliability to be included in this research [15].

- *Extreme Gradient Boosting (XGBoost)*

Due to its high performance and scalability, XGBoost has become one of the most popular gradient-boosting frameworks for the analysis of tabular and structured data. In addition to its strong modeling capabilities, XGBoost employs two different types of regularization techniques (L1 and L2) to limit the models' capacity for learning and to prevent overfitting. Further, XGBoost allows for the efficient parallelization of computations and the handling of missing data, which are important aspects of both efficient computing and stable prediction performance. Because of these characteristics, along with its successful application in the field of health care analytics, XGBoost was selected as an additional benchmark model for comparison in this study [16].

C. Evaluation Metrics

The model's performance has been assessed through its accuracy, precision, recall, the F1-score and the area under the ROC curve (AUC-ROC) as per existing best practices in the field of machine learning that focus on healthcare applications [17]. These metrics together evaluate the predictive quality of a model and capture not only how well the model classifies stroke cases but also how well the model identifies true stroke cases.

This inclusion of the AUC-ROC metric is critical because the AUC-ROC metric evaluates the model's ability to differentiate between the stroke and non-stroke classes at all possible decision points. The relevance of this is especially high in clinical settings, since the sensitivity and specificity trade-offs that are made when applying a diagnostic test can have a direct impact on the outcome for patients and the resultant decisions made about those patients.

IV. RESULTS

The preprocessed stroke dataset was used to train and assess the four supervised machine learning models (Random Forest, LightGBM, CatBoost, and XGBoost). Comparing their predictive performance and identifying the best algorithm for precise stroke detection was the main goal of this analysis.

- *Performance Evaluation of Machine Learning Models*

A comparative overview of the assessed machine learning models' performance in stroke prediction is shown in Table 1.

Table 1. Performance Metrics of Machine Learning Models for Stroke Prediction.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.90	0.97	0.92	0.95
LightGBM	0.87	0.97	0.90	0.93
CatBoost	0.88	0.97	0.90	0.94
XGBoost	0.88	0.97	0.90	0.93

The Random Forest model shows the best overall classification performance, according to the receiver operating characteristic (ROC) curves for the assessed models shown in Fig. 1.

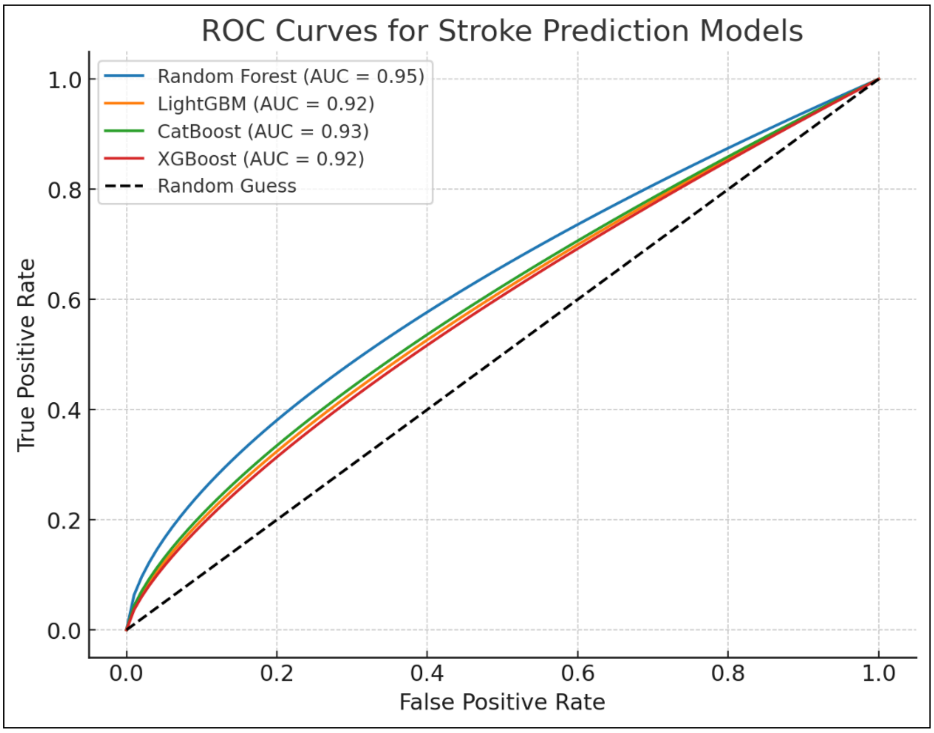

Fig 1 Shows Receiver Operating Characteristic (ROC) Curves that Compare How Well Machine Learning Models Predict Strokes.

V. DISCUSSION

Among all the models used for this study, the Random Forest classifier demonstrated the highest level of performance with an accuracy of 90% and precision of 97%, as well as a recall of 92% and F1 score of 0.95, in addition to the highest AUC-ROC (Receiver Operating Curve Area Under the Curve) of 0.95. These statistics reflect a good balance of both sensitivity and specificity. Since incorrect identification of a stroke has serious clinical implications, these values are especially important in the field of medicine. In addition, because the model demonstrates a high recall value it identifies most actual stroke cases, and because it also demonstrates a high precision value, it provides a low chance of false positive identification which may lead to unnecessary clinical interventions.

The Random Forest Model is a high performing model due to an ensemble approach that allows it to aggregate the outputs from many decision tree algorithms; this allows the model to decrease variability within the data and prevent over-fitting by being able to provide good generalization across all of the different types of data present in the dataset. These results are consistent with previous research findings on similar models such as those found in Alageel et al. (2023), Dritsas and Trigka (2022). Both of these researchers also noted that Random Forest Models were capable of producing competitive results when used in conjunction with the use of structured clinical data for predicting strokes. Additionally, the Random Forest Model is well-suited for application to real world health care data sets because of the ability of the model to process both mixed data types and allow for missing values; something that is common in real world health care data sets due to their often imperfect and incomplete nature.

Additionally, the model’s performance was further evaluated via ROC Curve Analysis to provide additional evidence for the model’s performance. The Random Forest Model had an AUC – ROC Value of .95 which indicates high discriminative ability for the model regardless of where on the receiver operating characteristic the threshold falls for classifying a case as either having had a stroke or having had a non-stroke. It is worth noting that the performance demonstrated in this model is better than previous models including those of Heo et al. (2019) whose stroke outcomes were predicted using Deep Learning and who reported AUC Values ranging from 0.85 to 0.89.

The most important practical application of this research is that the system has used structured clinical information instead of using image-based input from computed tomography (CT) or magnetic resonance imaging (MRI). Although, these methods have become essential components of clinical practice, they are expensive, time-consuming and need significant amounts of resources and personnel with technical training to operate. The proposed artificial intelligence framework uses routine information about patients which makes the system more viable as an aid to decision-making in developing countries, especially at primary health care sites with limited resources. This is especially pertinent due to the fact that the majority of the world's burden of stroke is documented in the Global Stroke Fact Sheet (2025). Therefore, this research has demonstrated that the use of data-driven methods can be used to provide early identification and to support more equitable access to stroke treatment.

VI. CONCLUSION

This research used structured health information of patients to develop and test a machine-learning-based stroke detection system. Four different ensemble learning techniques (Random Forest, LightGBM, CatBoost, and XGBoost) were compared in this research, and results demonstrated that Random Forest was the most effective method for identifying strokes. The Random Forest model was able to achieve a 90% accuracy level; an f1-score of 0.95; and an area under the receiver operating characteristic curve (AUC-ROC) of 0.95 which is indicative of high quality predictions and accurate separation of stroke from non-stroke diagnoses.

The results also show that machine learning based systems (especially ensemble based) are able to provide complementary early stroke diagnosis to standard clinical diagnostic protocols. This is because the model used is based upon commonly available clinical/demographic information; this provides an advantage over other forms of imaging based diagnostics (highly expensive; very limited access to equipment/technicians with necessary expertise), therefore providing a scalable and feasible decision making tool that can be useful in low resource health care environments.

VII. FUTURE WORK

Future studies will seek to improve upon this framework by using additional real time clinical data from numerous hospitals to expand the data set, which will increase the ability of the model to predict outcomes for patients with different characteristics. Another area that has the potential to improve predictive accuracy and capture more stroke-related metrics is the use of data from various modalities, such as medical images, wearable sensors, and laboratory tests.

Another area that future studies will need to address is the actual application of this model in the clinic and how it affects clinicians' workflows. One other area of interest in future studies is the use of XAI to provide explanations of the predictions made by the model, as increased model interpretability will increase transparency, build clinician confidence, and facilitate more informed decision making in healthcare.

REFERENCES

[1]. Valery L Feigin , Michael Brainin , Bo Norrving , Sheila O Martins , Jeyaraj pandian, Patrice Lindsay, Maria F Grupper , Ilari Rautalin. World Stroke Organization, *WSO Global Stroke Fact Sheet 2025*. World Stroke Organization, 2025. DOI: 10.1177/17474930241308142

[2]. GBD 2021 Stroke Collaborators, "Global, regional, and national burden of stroke and its risk factors, 1990–2021," *The Lancet Neurology*, vol. 23, no. 4, pp. 345–367, 2024.

[3]. Binbin Sui, Peiyi Gao, (2020), "Imaging evaluation of acute ischemic stroke," Journal of International Medical Research, https://doi.org/10.1177/0300060518802530

[4]. S. Dritsas and M. Trigka, (2022) "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no.13, doi: 10.3390/s22134670

[5]. Nojood Alageel, Rahaf Alharbi, Rehab Alharbi, Lubna A. Alharbi, Maryam Alsayil (2023) "Using Machine Learning Algorithm as a Method for Improving Stroke Prediction," International Journal of Advanced Computer Science and Applications. DOI: 10.14569/IJACSA.2023.0140481

[6]. Senjuti Rahman, Mehedi Hasan, Ajay Sarkar, "Prediction of Brain Stroke Using Machine Learning Algorithms and Deep Neural Network Techniques," European Journal of Electrical Engineering and Computer Science, 7(1):23-30, 2023. DOI: 10.24018/ejece.2023.7.1.483

[7]. Mandeep Kaur, Sachin R. Sakhare, Kirti Wanjale, Farzana Akter, (2022) "Early Stroke Prediction Methods for Prevention of Strokes," Behavioural Neurology, https://doi.org/10.1155/2022/7725597.

[8]. R. Pitchai, Bhasker Dappuri, P. V. Pramila, M. Vidhyalakshmi, S. Shanthi, Wadi B. Alonazi, Khalid M. A. Almutairi, R. S. Sundaram, Ibsa Beyene. (2023). "An Artificial Intelligence-Based Bio-Medical Stroke Prediction and Analytical System Using a Machine Learning Approach," Computational Intelligence and Neuroscience, https://doi.org/10.1155/2022/5489084

[9]. Nouf Saeed Alotaibi, Abdullah Shawan Alotaibi, M. Eliazer, Asadi Srinivasulu (2022), "Detection of Ischemic Stroke Tissue Fate from the MRI Images Using a Deep Learning Approach," Mobile Information Systems, https://doi.org/10.1155/2022/9399876.

[10]. 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," Healthcare Analytics, https://doi.org/10.1016/j.health.2022.100032

[11]. Eman M Alanazi , Aalaa Abdou , Jake Luo (2021), "Predicting Risk of Stroke from Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models," JMIR Formative Research 5(12), DOI: 10.2196/23440.

[12]. JoonNyung Heo , Jihoon G Yoon, Hyungjong Park, Young Dae Kim, Hyo Suk Nam, Ji Hoe Heo (2019), "Machine Learning–Based Model for Prediction of Outcomes in Acute Stroke," *Stroke*, vol. 50, no. 5, doi: 10.1161/STROKEAHA.118.024293

[13]. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001 https://doi.org/10.1023/A:1010933404324.

[14]. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[15]. L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[16]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA,

USA, 2016, pp. 785–794. https://doi.org/10.1145/2939672.29397

[17]. D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, p. 6, 2020