# Fine-Tuning Hyperparameters for Enhanced Stroke Disease Prediction: A Comparative Analysis of Machine Learning Approaches


Dr.R.Rajadevi
Associate Professor
Department of Artificial Intelligence
Kongu Engineering College
Perundurai,India
rajdevi.it@kongu.edu

Dr. R. Roopadevi
Associate Professor (SRG),
Department of Information Technology,
Kongu Engineering College,
Erode, India.
roopadevi.it@kongu.edu

Dr. R.S. Latha
Associate Professor
Department of Artificial Intelligence,
Kongu Engineering College,
Erode, India.
latha.cse@kongu.edu

R.S.Pradhanya
Student
Department of Artificial Intelligence
Kongu Engineering College
Perundurai, India
pradhanyars.22aim@kongu.edu

A.Sesili
Student
Department of Artificial Intelligence
Kongu Engineering College
Perundurai, India
sesilia.22aim@kongu.edu

K.T.Sutheeb
Student
Department of Artificial Intelligence
Kongu Engineering College
Perundurai, India
sutheebkt.22aim@kongu.edu



***ABSTRACT:*** A stroke is a medical condition characterized by the bursting or blockage of blood vessels in the brain, leading to brain tissue damage. This disruption in blood flow and vital nutrients can trigger symptoms. According to the World Health Organization (WHO), stroke is the foremost cause of both mortality and disability worldwide. Identifying early warning signs is essential for minimizing the impact of a stroke. Several machine learning (ML) models have been developed to evaluate stroke risk.In this study, various ML techniques, including the voting classifier, random forest (RF), decision tree (DT), and logistic regression (LR), are utilized to build predictive models.

The research places significant emphasis on optimizing hyperparameters for the Random Forest algorithm, such as the number of trees (n_estimators), maximum depth (max_depth), and minimum samples split (min_samples_split), to enhance model performance. The Random Forest algorithm, with optimized hyperparameters, achieved an accuracy rate of nearly 97%, emerging as the top performer among all models evaluated. This notably higher accuracy compared to prior studies indicates increased reliability in the models employed in this research. The study conducts a comparative analysis of multiple machine learning algorithms and concludes that the Random Forest model excels in predicting strokes. Such a model shows promise for integration into clinical practice by healthcare professionals



***Keywords—Machine Learing, Random forest, KNN, Decision Tree, Hyperparameter Tuning***


## I. INTRODUCTION

The study focuses on achieving 100% accuracy in stroke prediction via machine learning. It targets vessel-related dysfunctions lasting over 24 hours, categorizing strokes into four types. Patient characteristics like age, gender, BMI, glucose level, and medical history are extracted from electronic health records. Machine learning and neural network techniques are employed, integrating dimensionality reduction and dataset balance methods. Various models including Random Forest, Decision Tree, KNN, Naive Bayes, and MLP are explored, emphasizing hyperparameter tuning. Parameters like learning rates and regularization strengths are meticulously adjusted to enhance stroke prediction accuracy. This ensures robust generalization to unseen cases while balancing model complexity and interpretability, crucial for clinical use. Hyperparameter optimization leads to the development of highly effective stroke prediction models, potentially saving lives through timely intervention.

## II LITERATURE SURVEY

In this section, we will look at current research that employ machine learning to predict the risk of strokeShoily et al. (2019) investigated four machine learning techniques applied to patient data, likely including symptoms, to train various algorithms. Their objective was to develop a machine learning-based system for rapid stroke identification, aiming to assist medical practitioners in making timely and accurate diagnoses. Pradeepa et al. (2020) proposed a technique utilizing spectral clustering to extract stroke symptoms and prevention strategies from social media. They employed support vector machines, naive Bayes, ten-fold cross-validation, and probabilistic neural networks (PNNs), with PNNs achieving the highest accuracy of 89.90% . Sailasya and Kumari (2021) applied various machine learning methods to the Kaggle dataset for stroke prediction. Among the techniques used, Naive Bayes outperformed the others, achieving an accuracy of 82%. Govindarajan et al. (2020) used a stroke dataset from Sugam Multispecialty Hospital in India to classify different types of strokes. They

achieved over 95% accuracy with an artificial neural network using stochastic gradient descent, 91% accuracy with a bagged ensemble, and similar results with support vector machines Nwosu et al. (2019) identified stroke risk variables using electronic health records. They applied random forest, decision tree, and neural network algorithms, achieving accuracies of 74.53%, 74.31%, and 75.02%, respectively. Murray and Forouzanfar (2016), in their paper published in The Lancet Neurology, conducted a systematic study on the global impact of stroke. They highlighted the need for effective preventive measures and healthcare strategies to address this significant public health issue. He et al. (2008) address imbalanced datasets. Their approach, ADASYN, generates synthetic samples to balance class distributions, potentially improving machine learning algorithm performance. He and Garcia (2009) present valuable insights and methods for handling imbalanced datasets, crucial for improving machine learning model performance in real-world applications. Lamari et al. (2021) improve machine learning for imbalanced data by integrating SMOTE-ENN sampling with enhanced dynamic ensemble selection. This approach enhances predictive accuracy by effectively managing class imbalance, resulting in notable improvements, particularly in fields like medical diagnostics. F. Saeed, T. Al-Hadhrami, F. Mohammed, & E. Mohammed provides valuable insights into the classification of medical data. It is part of a comprehensive volume that explores advanced computational techniques for improving healthcare outcomes. Finally, Lee et al. (2020), a machine learning method is employed to identify strokes within a critical timeframe of 4.5 hours post-onset. This study underscores the significance of timely intervention in stroke management and illustrates the effectiveness of advanced computational techniques in healthcare. By employing machine learning algorithms, the research aids in refining early diagnosis and treatment strategies, ultimately enhancing patient outcomes It emphasizes the importance of timely intervention in stroke management and demonstrates the efficacy of advanced computational techniques in healthcare. Through the application of machine learning algorithms, the research contributes to refining early diagnosis and treatment approaches, ultimately leading to improved patient outcomes.

## III.METHODOLOGY

Machine learning is a revolutionary tool in system development, enabling systems to autonomously learn from past data across various industries like healthcare, banking, retail, and transportation, all managing extensive databases. By automatically identifying patterns, machine learning boosts accuracy and efficiency, notably in healthcare, where it aids in disease diagnosis, outcome prediction, and treatment planning. To maximize effectiveness, preprocessing steps like feature engineering and data cleaning are crucial. This prepares data for analysis, with subsequent division into testing and training sets. Testing evaluates algorithm performance on new data, while training educates the system on data patterns. Predicting stroke risk exemplifies machine learning's practical application, using various factors like medical history and lifestyle choices. This predictive ability can save lives and reduce healthcare costs through early intervention. Different methods like Naive Bayes, K-Nearest Neighbors, and Random Forest are employed for stroke prediction, each with its strengths and weaknesses. Parameter tuning, adjusting factors like learning rates and regularization strengths, is crucial to optimize model performance. Accurate predictive models necessitate an iterative process of parameter tuning, especially in healthcare applications like stroke prediction.To maximize machine learning models' performance, parameter adjustment is essential. Through the manipulation of variables including learning rates, regularization strengths, and tree depths, scientists can refine algorithms to get enhanced precision and resilience. Creating accurate predictive models for stroke prediction and other healthcare applications requires an iterative process of parameter adjustment.

## IV.PREPROCESSING

### *A. Dataset description*:

The study utilized a large dataset from Kaggle, including 3254 participants aged 18 or older. The dataset features nine input properties, mostly nominal variables, and provides a detailed target class description crucial for data analysis and interpretation, as detailed in Table1.

*TABLE 1. Description about dataset*

| Risk factor | Description | Details |
|---|---|---|
| ***Age (year)*** | The actual age of participants | All of the participants are older than 18 |
| ***Gender*** | Whether the participants is male or female | In the dataset, 1260 participants are males, and 1994 participants are female |
| ***Hypertension*** | The participant suffering from hypertension or not | 12.54% of the participants in the dataset are hypertensive |
| ***Heart disease*** | The participants suffering from heart diseases in general or not | 6.33% of the participants in the dataset suffer from heart diseases |
| ***Marital status*** | The participant is married or not | In the dataset, 79.84% of the participants are married |
| ***Work type*** | The work status of the participants | 65.02% of them work in the private sector, 19.21% are self-employed, 15.67% have a job, while 0.1% have never worked |
| ***Residence type*** | Whether the participant lives in an urban or rural place | 51.14% of the participants in the dataset live in urban place whereas the rest live in rural places |
| ***Avg glucose level (mg/dL)*** | The average level of a participant's blood glucose | Numerical values for each patient |
| ***BMI (Kg/m2)*** | Participant's body mass index of the participants | Numerical values for each patient |
| ***Smoking status*** | Whether a participant currently smokes or not | 22.37% of the participants smoke, 24.99% of them have smoked in the past, and 52.64% of them have never smoked |
| ***Stroke history*** | Whether the participant has had a stroke previously or not | 5.53% of the participants have previously had a stroke |

***B.Data pre-processing***: To address missing values, redundancy, and noisy data, actions like data discretization were applied. Resampling methods, such as SMOTE, balanced the dataset by evenly distributing participants between stroke and non-stroke groups, ensuring representativeness.

***C. Data Exploration and Visualization:***

a) Gender Distribution Analysis:

To standardize categories and reduce dimensionality, 'Other' instances are relabelled as 'Female'. Gender distribution is visualized with a pie chart in Fig1.

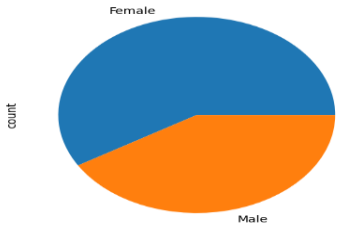


*Fig 1. Gender Distribution: Visual Representation*

b) *Target Attribute Distribution:*
The frequency of stroke occurrences is depicted in a bar chart (Fig 2), providing insight into the distribution of stroke incidences within the dataset.

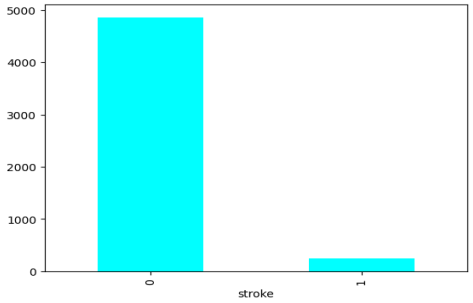


*Fig 2. Stroke Frequency Visualization*

*c)Hypertension Attribute Distribution:*
A bar chart (Fig 3) illustrates the distribution of hypertension incidences, showing the prevalence of hypertension in the dataset.

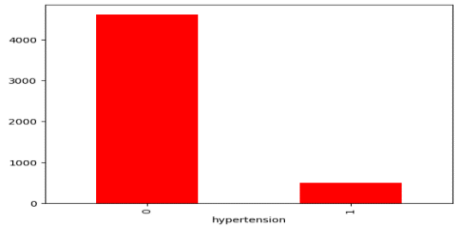


*Fig 3. Hypertension Incidence Visualization*

*d)Work Type Attribute Distribution:*
A pie chart (Fig 4) visualizes the distribution of work types in the dataset, highlighting the predominant categories.

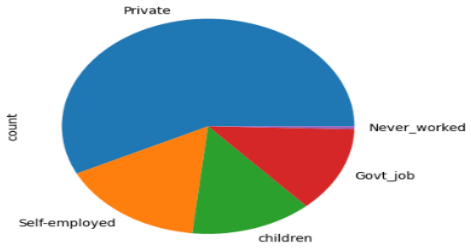


*Fig 4. Work Type Distribution Overview*

*e) Smoking Status Attribute Distribution:*
Through a pie chart, the distribution of smoking statuses within the dataset is displayed, providing insights into the prevalence of different smoking habits among individuals is represented in Fig 5.

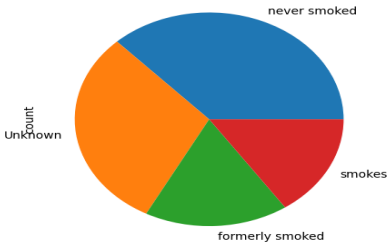


*Fig 5. Smoking Status Breakdown*

*f) Residence Type Attribute Distribution:*
A pie chart (Fig 6) shows the distribution of urban and rural residences in the dataset.

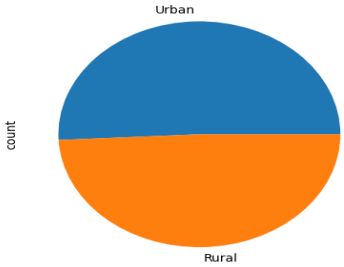


*Fig 6. Residence Type Distribution Overview*

*g) BMI Distribution:*
The distribution of BMI values is examined using a histogram, allowing for the visualization of the spread and density of BMI measurements within the dataset is represented in Fig 7.

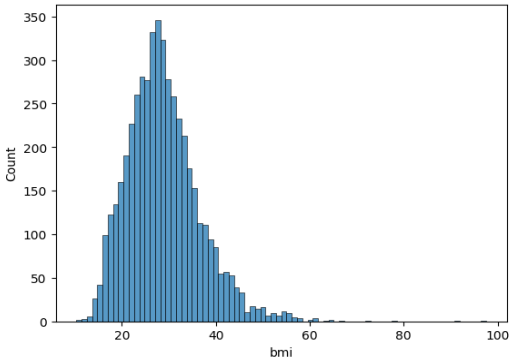


*Fig 7. BMI Distribution Analysis*

*h) BMI Outlier Analysis:*
A boxplot is generated to identify outliers and assess the variability and distribution of BMI values within the dataset. Outliers, if present, can indicate potential data anomalies or extreme values that warrant further investigation is represented in Fig 8.

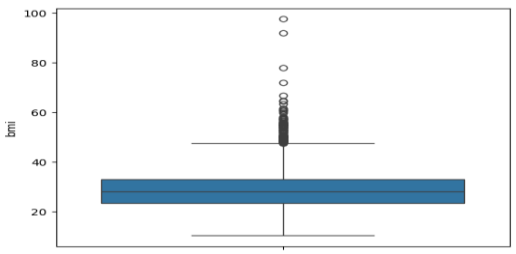


*Fig 7. BMI Distribution Analysis*

### *D. Data Encoding and Correlation Analysis*

a) Encoding Categorical Variables: Categorical variables, such as 'Male', are transformed into numeric representations.
b) Computing Correlation Matrix: A correlation matrix is calculated to assess the relationships between variables.
c) Heatmap Visualization: The correlation matrix is visually depicted using a heatmap to identify patterns and correlations between variables.

### *F. Evaluation:*

Various evaluation metrics assess machine learning models. Key metrics include accuracy, precision, recall, and F1-Score. Accuracy measures correct classifications. Precision gauges accurate positive predictions. Recall assesses positive instance identification. F1-Score balances precision and recall. These metrics offer insights into model performance, aiding decision-making.

## V. Hyper Parameter Tuning

The study utilizes diverse algorithms for stroke prediction, including Random Forest, Decision Tree, MLP, KNN, and Naive Bayes. Random Forest achieves exceptional performance with 99.33% accuracy, highlighting its robust predictive capabilities. Its ensemble approach and adaptability make it reliable for early stroke detection, crucial for patient outcomes

*Random Forest*: Random Forest (RF) classification employs multiple decision trees trained on random data subsets, aggregated through voting for predictions, effective for tasks like relapse detection. Key hyperparameters like Estimators, Features, Depth, Sample Split, Sample Leaf, and Bootstrap are optimized via grid and random search. The optimized configuration includes 100 estimators, 'sqrt' features, depth of 30, sample split of 2, sample leaf of 1, and bootstrap set to False.

*Decision Tree:* Decision Tree (DT) classification segments data based on specified parameters. Key hyperparameters, such as criterion, splitter, max depth, min samples split, min samples leaf, and max features, are optimized via grid and random search. Optimal values include the entropy criterion, best splitter, no max depth, min samples split of 2, min samples leaf of 1, and 'sqrt' max features.

*Multilayer Perceptron:* Multilayer Perceptrons (MLPs) are crucial for stroke prediction, capturing complex patterns in patient data. They offer high predictive accuracy and mitigate overfitting through regularization. Optimal parameters from grid search include a (100, 50) hidden layer size, tanh activation, Adam solver, alpha of 0.001, and constant learning rate. Random search suggests a (50,) hidden layer size. *K*-Nearest Neighbour: K-Nearest Neighbours (KNN) defers training until classification is needed, operating on stored datasets. With 80% accuracy, KNN shows recall and precision rates of 83.7% and 77.4%, respectively, and an F1 Score of 80.4%. Optimal parameters from grid search include 3 neighbors, distance weighting, auto algorithm selection, leaf size of 20, and P value of 1. Random search suggests similar values with kd_tree algorithm and leaf size of 30.

Naive Bayes: Naive Bayes assumes feature independence within classes based on Bayes Theorem. Achieving 82% accuracy, 79.2% precision, and 85.7% recall, it has an F1 Score of 82.3%. Key hyperparameters from grid search include smoothing (1e-05) and class priors (None), with random search suggesting smoothing (1e-05) and class priors ([0.4, 0.6]).

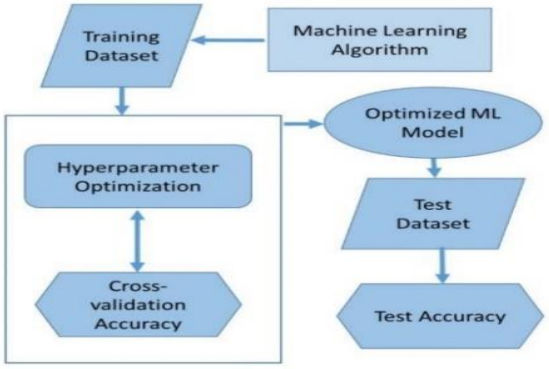


Table 2:

| Classification Model | Parameter | Parameter Value | Best parameter Grid Search | Best parameter Random Search |
|---|---|---|---|---|
| Random search | **Estimators** | [100, 200, 300, 400, 500] | 100 | 100 |
| | **Features** | ['auto', 'sqrt'] | auto | Sqrt |
| | **Depth** | [10, 20, 30, 40, 50,60,70,80,90,100] | 30 | 30 |
| | **Sample Split** | [2, 5, 10] | 2 | 2 |
| | **Sample leaf** | [1, 2, 4] | 1 | 1 |
| | **Bootstrap** | [True, False] | False | False |
| Decision Tree | **criterion** | ['Gini', 'entropy'] | Entropy | Entropy |
| | **splitter** | ['best', 'random'] | Best | Best |
| | **Max depth** | [None, 10, 20, 30, 40, 50,80,100,120] | None | None |
| | **Min samples split** | [2, 5, 10,15 | 2 | 2 |
| | **Min samples leaf** | [1, 2, 4,10] | 1 | 1 |
| | **Max features** | [None, 'auto', 'sqrt', 'log2'] | sqrt | Sqrt |
| KNN | **N neighbours** | [3, 5, 7, 9, 11] | 3 | 3 |
| | **Weights** | ['uniform', 'distance'] | Distance | Disrtance |
| | **Algorithm** | ['auto', 'ball tree', 'kd tree', 'brute' ,'sparse'] | Auto | Kd_tree |
| | **Leaf size** | [20, 30, 40, 50] | 20 | 30 |
| | **P value** | [1, 2] | 1 | 1 |

| | | | | |
|---|---|---|---|---|
| **Naive Bayes** | **Smoothing** | [1e-9, 1e-8, 1e-7, 1e-6, 1e-5] | 1e-05 | 1e-05 |
| | **Class Priors** | [None, [0.1, 0.9], [0.2, 0.8], [0.3, 0.7], [0.4, 0.6], [0.5, 0.5]] | None | [0.4, 0.6] |
| **MLP** | **Hidden Layer Size** | [(50,), (100,), (50, 50), (100, 50)] | (100, 50) | (50,) |
| | **Activation** | ['logistic', 'tanh', 'relu'] | Tanh | Tanh |
| | **Solver** | ['sgd', 'adam'], | Adam | Adam |
| | **Alpha** | [0.0001, 0.001, 0.01] | 0.001 | 0.001 |
| | **Learning Rate** | ['constant', 'adaptive'] | Constant | Constant |

Finally, the comparative analysis of bar chart with and without using hyperparameter is given in Fig 10.

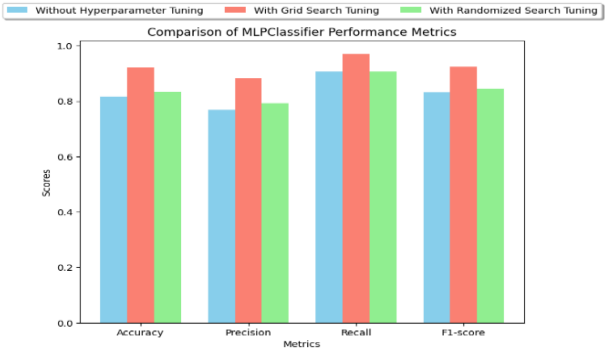


*Fig 10. Comparison of MLP classifier Performance Metrics*

## VI.RESULTS

Evaluation measures for different stroke prediction algorithms were compared through thorough data collection, preprocessing, and visualization combined with painstaking hyperparameter tuning. The performance of each algorithm was optimized by hyperparameter optimization, which guaranteed the best possible parameter values for increased predicted accuracy. Models were optimized for best performance across various patient datasets through methodical hyperparameter research. The resultant thorough summary in the table offers insightful information on how hyperparameter tuning affects algorithm performance, helping healthcare decision-makers choose the best algorithm for better patient care and results.The comparative analysis of algorithm is given in Table 3.

*TABLE 3. Comparative analysis of algorithm*

| **Algorithm** | **Accuracy** | **F-1 Score** | **Recall** | **Precision** |
|---|---|---|---|---|
| **Random Forest** | 99.38 | 99.38 | 1.0 | 98.77 |
| **Decision Tree** | 97.73 | 97.78 | 1.0 | 95.66 |
| **Multilayer Perceptron** | 81.49 | 82.80 | 89.38 | 77.13 |
| **KNN** | 97.22 | 97.29 | 1.0 | 94.72 |
| **Naïve Bayes** | 67.66 | 75.34 | 99.07 | 60.78 |

## VII.CONCLUSION

In conclusion, the comparison of various machine learning algorithms for stroke prediction underscores the pivotal role of hyperparameter optimization in determining model performance. Random Forest (RF) classification, Decision Tree (DT) classification, Multilayer Perceptron (MLP), K-Nearest Neighbours (KNN) classification, and Naïve Bayes classification each exhibited unique strengths and weaknesses, influenced by hyperparameters like estimators, features, depth, sample split, and smoothing. Through techniques such as grid search and random search, optimal parameter values were identified to maximize predictive accuracy and generalization across diverse patient datasets. While RF and DT demonstrated robust performance with finely tuned hyperparameters, MLP showcased adaptability and high predictive accuracy, underscoring the critical role of hyperparameter optimization in optimizing machine learning models for effective stroke prediction in clinical practice.

## VIII.REFERENCES


[1] Shoily T.I., Islam T., Jannat S., Tanna S.A., Alif T.M., Ema R.R. Detection of stroke disease using machine learning algorithms; Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT); Kanpur, India. 6–8 July 2019; Piscataway, NJ, USA: IEEE; 2019. pp. 1–6.

[2] Pradeepa S., Manjula K., Vimal S., Khan M.S., Chilamkurti N., Luhach A.K. DRFS: Detecting risk factor of stroke disease from social media using machine learning techniques. Neural Process. Lett. 2020;2020:1–19. doi: 10.1007/s11063-020-10279-8.

[3] Sailasya G., Kumari G.L.A. Analyzing the performance of stroke prediction using ML classification algorithms. Int. J. Adv. Comput. Sci. Appl. 2021;12:539–545. doi: 10.14569/IJACSA.2021.0120662.

[4] Govindarajan P., Soundarapandian R.K., Gandomi A.H., Patan R., Jayaraman P., Manikandan R. Classification of stroke disease using machine learning algorithms. Neural Comput. Appl. 2020;32:817–828. doi: 10.1007/s00521-019-04041-y.

[5] Nwosu C.S., Dev S., Bhardwaj P., Veeravalli B., John D. Predicting stroke from electronic health records; Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); Berlin, Germany. 23–27 July 2019; Piscataway, NJ, USA: IEEE; 2019. pp. 5704–5707.

[6] Murray, C. J. L., & Forouzanfar, M. H. (2016). Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet Neurology, 15(9), 913–924.

[7] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of the International Joint Conference on Neural Networks, 3, 1322–1328.

[8] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284

[9] Jason, B. (2020). Tour of Evaluation Metrics for Imbalanced Classification. Machine Learning Mastery.

[10] Lamari, M., Azizi, N., Hammami, N. E., Boukhamla, A., Cheriguene, S., Dendani, N., & Benzebouchi, N. E. (2021). SMOTE--ENN-Based Data Sampling and Improved Dynamic Ensemble Selection for Imbalanced.

[11] In F. Saeed, T. Al-Hadhrami, F. Mohammed, & E. Mohammed (Eds.), “Medical Data Classification”, Advances on Smart and Soft Computing (pp. 37–49). Springer Singapore.

[12] Lee H., Lee E.J., Ham S., Lee H.B., Lee J.S., Kwon S.U., Kim J.S., Kim N., Kang D.W. Machine learning approach to identify stroke within 4.5 hours. Stroke. 2020;51:860–866. doi: 10.1161/STROKEAHA.119.027611.