Predicting and Analyzing Early Onset of Stroke Using Advanced Machine Learning Classification Technique

Puja Kumari
CHRIST (Deemed to be University)
India
puja.kumari@mca.christuniversity.in

Jossy George
CHRIST (Deemed to be University)
India
frjossy@christuniversity.in

Akhil M Nair
Luxsh Technologies Pvt Ltd,
United Kingdom
akhil.nair@christuniversity.in

Bosco Paul Alapatt
CHRIST (Deemed to be University)
India
bosco.paul@christuniversity.in

Riya Baby
CHRIST (Deemed to be University)
India
Riya.baby@christuniversity.in

Jiby Jose
CHRIST (Deemed to be University)
India
jiby.jose@christuniversity.in

2024 10th International Conference on Electrical Energy Systems (ICEES) | 979-8-3503-5377-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICEES61253.2024.10776853

Abstract—Around the world, stroke is the leading cause of death. When blood vessels in the brain rupture, they cause damage. Alternatively, blockage in a blood vessel that supplies oxygen and other nutrients may also lead to this disease. This study uses various machine learning models to predict whether someone will have a stroke or not. Different physiological features were taken into account by this study while using Logistic Regression; Decision Tree Classification; Random Forest Classification; K-Nearest Neighbors (KNN); Support Vector Machine (SVM); Naïve Bayes classifier algorithm; and XGBoost classification algorithm – these were used for six different models to ensure accurate predictions are made. We will accomplish the finest exactness with Bayes cv look which may be a hyper-tuning classifier with 92.87%. This consideration can be utilized for future work by doing the increase and include designing on the dataset. It is constrained to literary information, so it might not continuously be right for foreseeing stroke. so utilize the datasets that contain pictures and work on those datasets.

Keywords—Stroke, Machine learning, Decision Tree Classification, Random Forest Classification ,Logistic Regression , K_Nearest neighbors, Support Vector Machine, Naïve Bayes classification , Bayes cv search.

I. INTRODUCTION

Stroke, a weakening condition characterized by the sudden impedance of brain work remains a noteworthy open well-being concern all inclusive. Interests, and their effect rise above personal well-being, amplifying financial burdens and healthcare frameworks around the world. The capacity to anticipate and avoid strokes, especially among youthful grown-ups, holds gigantic significance in relieving its unfavorable results and making strides in the general populace's well-being incredibly. In this setting, the application of machine learning (ML) methods offers a promising road for upgrading stroke expectation models, encouraging early intercession too much, and eventually decreasing the burden of stroke-related horribleness and mortality insanely!

Machine Learning can predict the onset of a stroke given an advancement in medical technology. A correct analysis is provided by constructive algorithms in Machine Learning..The center of this consideration revolves around the improvement and assessment of Machine Learning(ML) -based prescient models for stroke in youthful grown-ups.

Whereas strokes customarily influence more seasoned populaces, later patterns show a rising rate among more youthful age bunches, underscoring the requirement for focused on preventive techniques custom fitted to this statistic. By leveraging comprehensive datasets enveloping assorted clinical and statistical factors, we point to tackle the prescient control of ML calculations to distinguish people at tall chance of stroke and illuminate personalized intercession techniques.

The assignment starts by selecting a dataset from Kaggle which has some physiological features as its attributes.These features are then studied and used for the final prediction. The dataset is cleaned first to prepare it for machine learning. This is known as Data Preprocessing. In this process, the dataset is checked for missing values and fill them in. Then Name encoding is performed to convert string values into integrability followed by one-hot encoding if required.

After data Preprocessing, the dataset is split into train and test data. A model is then created using this fresh data using different Classification Algorithms. Accuracy is calculated for all these algorithms and compared to get the best trained model for prediction. Finally we are applying different algorithms to find the best accuracy.

II. LITERATURE REVIEW

In G. a. G. L. A. K. Sailasya [1], stroke prediction was made on the same datasets that I chose for my research, this dataset using six machine learning techniques. Among these, Naïve Bayes achieved the highest accuracy (approximately 82%). This paper emphasizes the importance of early stroke detection and highlights Naïve Bayes as a promising algorithm for predicting brain strokes.

In research paper by Krishna Mridha ,Sandesh Ghimire , et.al [2], the researchers developed automated stroke prediction

979-8-3503-5377-8/24/$31.00 ©2024 IEEE

algorithms, which would allow for early intervention and perhaps save lives. He Used many algorithms like XGBoost, KNN, Logistic regression, SVM, and Random Forest.

In research paper by K Lipska, P N Sylaja, et.al "[3] the study underscores the importance of targeting adolescents and young adults for screening and prevention to reduce the burden of ischemic stroke in this age group. It can Evaluate risk factors .

In research paper by Anirudha S. Chandrabhatla ,et.al [4], this paper reviews FDA-approved technologies that utilize artificial intelligence/machine learning (AI/ML) for the diagnosis and management of stroke and in this some challenges like anatomical variation which affect the Performance.

In research paper by Xueyang Wang, Jinhao Lyu ,et.al [5], it investigates the role of cerebral SVD in predicting outcomes. It uses predictive models like Random forest Logistic regression, Gaussian process regression, and XGboost.

In research paper by Enzhao Zhu, Zhihao Chen, et.al [6], the author extracted a database from Medical Information Mart from the Intensive Care (MIMIC)-IV. He contrasted the demographic variables of control group with those of death group. Moreover, he built up predictive models for stroke deaths using machine learning (ML). Here, six ML algorithms were used: Neural Oblivious Decision Ensemble (NODE) like Catboost; XGBoost; LightGBM; fully connected neural network (FCNN); and logistic regression (LR).Among these Catboost has the highest accuracy of 89.95%.

The research paper by Soumyabrata Dev, Hewei Wang, et.al "[7]": A perceptron neural network using these four attributes age, heart disease, avg. glucose level and hypertension achieved the highest accuracy.

In a Research paper by Lea Fast , Uchralt Temuulen , et.al [8]. ML model was trained to predict the outcomes. And it's future applications include personalized rehabilitation programs and decision support for clinicians.

In research paper by Olusola Olabanjo, Ashiribo Wusu, et.al [9], They uses Ml and Deep learning for predicting stroke risk. The future work of this paper is to explore model robustness and scalability to handle large-scale data and real-time predictions. Consider deployment in telemedicine or remote monitoring scenarios.

In research paper Mylapalli Kanthi Rekha, Phani Kumar [10], proposed a predictive model using the Random Forest and AdaBoost algorithms to predict the likelihood of a brain stroke based on various risk factors. Both achieved an accuracy of 90%. Explore hyperparameter tuning and Model optimization to enhance the accuracy.

In research paper Giuseppe Miceli , Maria Grazia Basso , et.al and [11], They used the TOAST classification system, and in this, AI plays a vital role in IS diagnosis, subtype classification, and risk prediction. It enhances accuracy, aids in early detection, and improves patient outcomes.

In research paper Mr. Dilesh Yuvraj Bagul , Dr. P. B. Bharate ,et.al. [[12] used four Ml models random forest, Decision tree, Naïve Bayes, and logistic regression. Among these random forest has the highest accuracy of 92.09%. It suggests using a larger dataset for prediction and enhancing the accuracy by using various models.

III. METHODOLOGY

To implement this, we have employed multiple Kaggle datasets.Cleaning, transforming and preparing data for analysis is called data preprocessing after it has been collected as a dataset.

When the data has been preprocessed and ready for model building, processed datasets are used in combination with machine learning algorithms. We have tried different models such as Random Forest, KNN, Decision tree, Logistic Regression, Naïve Bayes, SVM and Adaboost to predict stroke. Also we balanced the data using undersampling. To improve prediction accuracy hyperparameter tuning was done which included Grid search and Bayes CV search. Flow chart of the methodology is in Fig. 1.

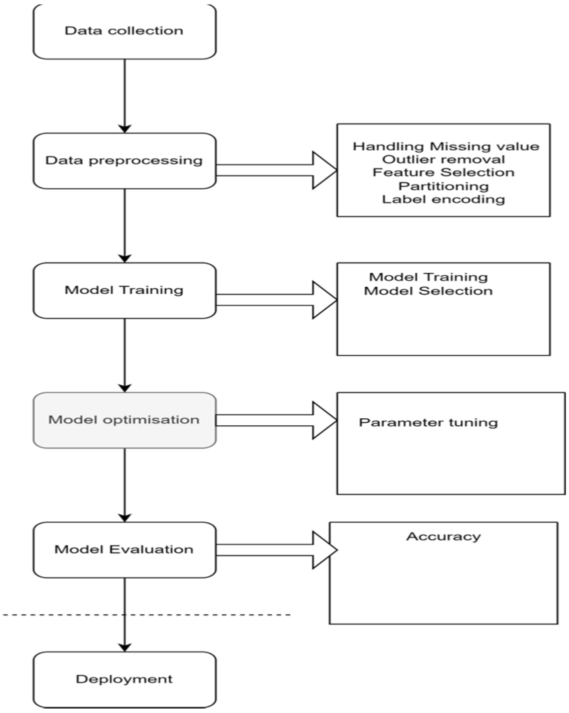

Fig 1: Flowchart of the model

IV. IMPLEMENTATION

Dataset:
The given stroke prediction dataset comes from Kaggle. There are 5110 records and 12 columns in this dataset. 'id',' gender', 'age', 'hypertension', heart_disease', 'ever_married', 'work_type', 'Residence_type', 'average _glucose_level', 'Bmi', 'smoking_status' and 'stroke' which is the target

attribute are the columns. The column named ‘stroke’ contains either ‘1’ or ‘0’. When it is equal to zero, no risk of stroke has been detected while one indicates possible risk for getting a stroke. This dataset is imbalanced with respect to class distribution since frequency count for occurrence of ‘0’ in output variable (‘stroke’) is significantly higher than that of ‘1’. Accordingly, only 249 records have value as one while remaining 4861 entries are labelled as zero under stroke column. So as to provide more accurate results during analysis, it becomes necessary undertaking data preprocessing procedures aimed at balancing out numbers between different categories within our dataset . The table below shows summary statistics relating to the aforementioned dataset (Table 1).

Table 1: Stroke Dataset

Attributes	**Type**	**Explanation**
1.id	Integer value	It is unique integer value for Patients.
2.gender	String *(Male, Female, Other)*	It defines sex of the patients .
3.age	Integer value	It helps in defining age of the patients.
4.hypertension	Integer value *(1,0)*	It defines whether the patient has high blood pressure or not.
5.heart_disease	Integer value *(1,0)*	It tells whether patient has heart related disease or not.
6.ever_married	String *(Yes, No)*	It defines whether the patient was married or not.
7.work_type	String(self employed, govt.job , private , never worked)	It defines the different categories..
8.Residence_type	String *(Urban, Rural)*	It defines the residence type of the patient.
9. avg_glucose_level	Floating point number	Defines how much the glucose present in the blood.
10.bmi	Floating point number.	Defines Body Mass Index of the patients.
11.smoking_status	String *(never smoked ,formerly smoked, smokes,unknown)*	It defines the smoking status of the patient.
12.stroke	Integer *(1,0)*	Target column that provides the stroke status of the patient.

Data Preprocessing:

After collecting the datasets, data preprocessing is done where we can handle the missing values, balance the dataset, and convert string value to Numeric value.

Handling Imbalanced Data

We have datasets of 5110 rows and 12 columns from that dataset 4861 have no occurrence of stroke and 249 have an occurrence of stroke. So we need to balance the data for that we are using Random Under-Sampling.

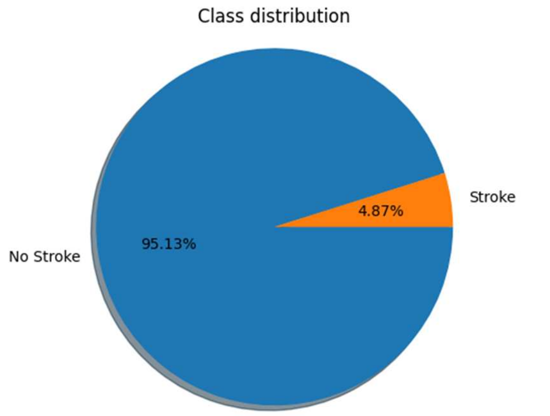

Fig 2: Before Undersampling.

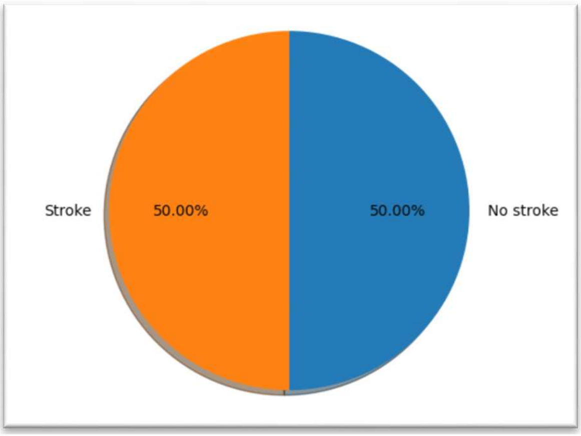

Fig 3: After Undersampling.

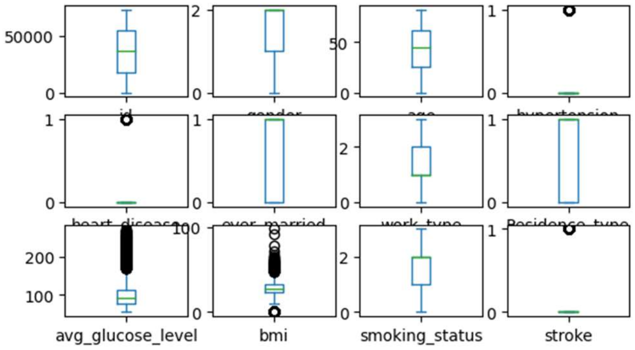

Fig 4: Boxplot of all the parameters

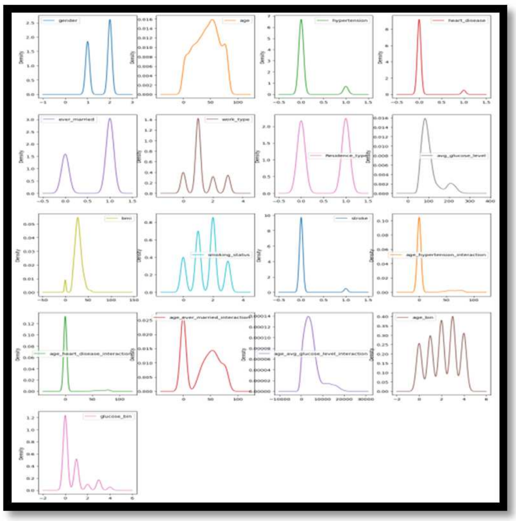

Fig 5: Density plot

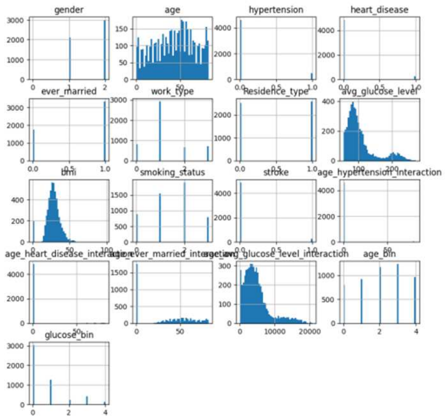

Fig 6: Histogram of all the parameters.

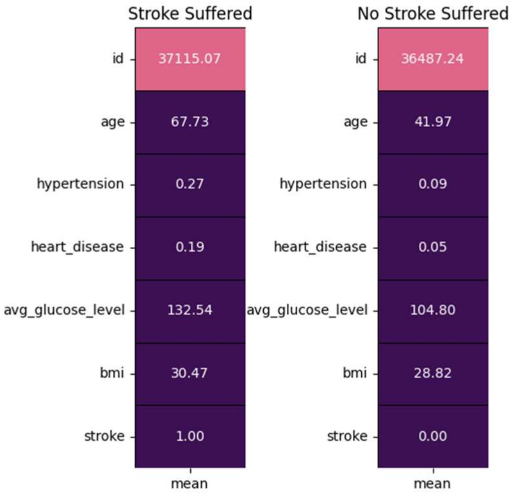

Fig 7: Heatmap of Stroke and No Stroke

MODEL BUILDING

Splitting the Data:

Upon the completion of data preprocessing and dealing with the imbalanced dataset, comes the step of creating a model. The information is divided into training and testing data by splitting it in 80% to train and 20% to test. Now we will train our model using different classification algorithms such as KNN, SVM, Random forest, logistic regression, Naïve Bayes, Decision tree and Adaboost.

Classification Algorithms:

Logistic regression: Logistic Regression is used for predicting the binary outcome. It works on 0 or 1, True or False , Yes or No etc.

Decision tree classification: Decision Tree is a tree like structure which is used for classification and regression tasks. It helps in making Decisions. It has Decision node ,and Leaf nodes .

Random forest classification: Random forest is used for increasing the accuracy by building multiple decision trees during training. It is a powerful model.

K-nearest Neighbours classification: KNN is also a ML model which is used for classification and regression tasks. Basically it decides the class of a new data point by looking at the classes of its closest neighbours.

Support vector machine: SVM is a powerful machine learning model which is used for classification. It is done by drawing a boundary between different groups of data to separate.

Naïve Bayes classification: It is used for classification and regression tasks. It is basically works on the bayes theorem.

XGBoost: It is a machine learning model and it allows for the optimization of arbitrary differentiable loss functions.

C) Hyperparameter tuning:

Hyperparameter tuning is just like settings or options that we chose to build a model. It is specified before training the model and can impact the model’s performance and behavior. So we can talk about the two parameter tuning which is Grid CV search and Bayes CV search.

Bayes Search CV: It is a technique to find the best hyperparameters for Ml. It learns from past experiments to work efficiently.

Grid Search: Grid Search works by creating a grid of all possible combinations hyperparameters that we want to tune and evaluate its performance using a cross –validation.

V. RESULTS AND DISCUSSION

From the study, it can be concluded that Bayes search CV has performed better when compared to other algorithms which achieved 92.87%. Study shows that the accuracy we have achieved in the field of classification modeling. High accuracy highlights the robustness of our approach. We Used cross-validation Techniques to strengthen the reliability of the results. By systematically dividing the datasets and by training and testing the model with different combinations, we found this much accuracy.

Achieving consistently high accuracy across multiple cross-validation folds not only validates a model's generalizability but also provides confidence in its stability and reliability.

This not only highlights the suitability of the Bayesian framework to the datasets but also effective in capturing its patterns and relationships. It highlights accuracy, Validation techniques, and Comparative analysis, etc. The high accuracy of our stroke prediction model is very important in the medical sector.

Accurately identifying individuals at risk for stroke allows healthcare to implement targeted intervention strategies and preventive measures to reduce the occurrence of stroke, and used for improving patient outcomes and improving health. By doing comparative analysis we got to know that this is not accurately predicting the stroke.

***Accuracy**: It provides that how much the model is able to predict correct.*

$$Accuracy = \ No.of\ correct\ prediction\ / Total\ no.of\ \ prediction.$$

Precision: It focuses on proportion of true positive prediction out of all positive prediction.

$$Precision = True\ positive\ / True\ positive + \ False\ positive$$

Recall: It is also called sensitivity. And it works same as the precision.

$$Recall = TP \div TP + FP$$

F1 Score: It depends on the mean of precision and recall.

$$F1\ score = 2 \times \ Precision\ * Recall\ /\ \ Precision + \ Recall$$

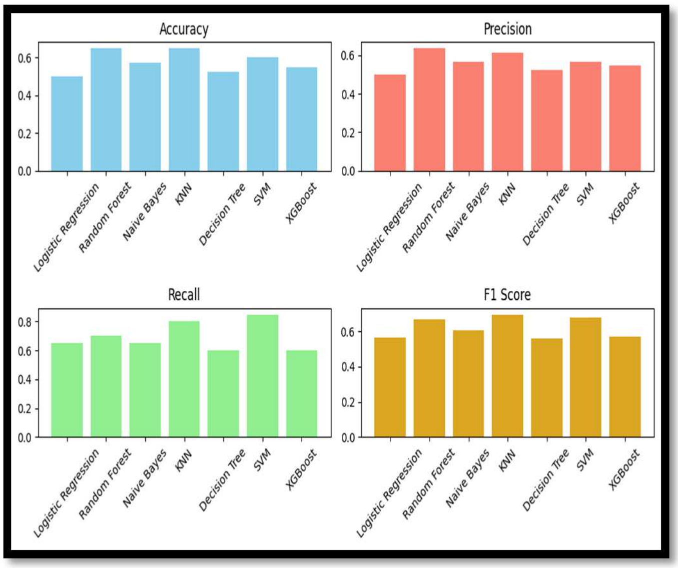

Fig.8: Performance of the chosen ML models based   on Evaluation Metrics.

Table 2: Comparison of all advance Ml models.

Model	**Accuracy**	**Recall**	**F1 Score**	**Precision**
Logistic regression	50	65	56.52	50
KNN	**65**	80	69.56	61.53
SVM	60	85	68	56.66
Decision Tree	52.5	60	55.81	52.17
Random Forest	65	70	66.66	63.63
Bayes cv search	**92.87**	_	_	_
XGBoost	55	60	57.14	54
Naïve Bayes	57.5	65	60.46	56.52

VI. CONCLUSION

Stroke happens when the blood supply to part of our brain is reduced either due to blockage or a burst blood vessel. This can cause brain cells to die. It is the second most common cause of death that should be treated before it worsens. By the report of WHO, it is clear that 13% -15% are affected by this. By building Machine learning (Ml) models it can help in predicting the stroke before it occurs. It can be predicted by looking at its parameters like age, heart disease, hypertension, etc. It can help in reducing the severe impact. we can work on different models to find that which works better in prediction of Stroke .

The Bayesian classifier is the most accurate of all the algorithms, with an accuracy of 92.87%. It's limited to textual data, so it may not always be a good indicator of stroke. To improve the reliability of the model, we can use datasets containing images and working on those datasets.

REFERENCES

[1] G. a. G. L. A. K. Sailasya, "Analyzing the performance of stroke prediction using ML classification algorithms.," *International Journal of Advanced Computer Science and Applications* , vol. 12, (2021).

[2] KRISHNA MRIDHA , (Member, IEEE), SANDESH GHIMIRE , JUNGPIL SHIN , (Senior Member, IEEE), ANMOL ARAN , MD. MEZBAH UDDIN1 , AND M. F. MRIDHA "Automated stroke prediction using machine learning: An explainable and exploratory study with a web application for early intervention.," *IEEE Access,* vol. 11, (2023).

[3] K Lipska, P N Sylaja, P S Sarma, K R Thankappan, V R Kutty, R S Vasan, K Radhakrishnan, "Risk factors for acute ischaemic stroke in young adults in South India.," *Journal of Neurology, Neurosurgery & Psychiatry,* 2007.

[4] Anirudha S. Chandrabhatla , Elyse A. Kuo , Jennifer D. Sokolowski , Ryan T. Kellogg , Min Park and Panagiotis Mastorakos, "Artificial intelligence and machine learning in the diagnosis and management of stroke: a narrative review of United States food and drug administration-approved technologies," *Journal of Clinical Medicine* , 2023.

[5] Xueyang Wang Jinhao Lyu Zhihua Meng Xiaoyan Wu Wen Chen Guohua Wang Qingliang Niu Xin Li Yitong Bian Dan Han Weiting Guo Shuai Yang Xiangbing Bian Yina Lan Liuxian Wang Qi Duan Tingyang Zhang Caohui Duan Chenglin Tian Ling Chen, | Xin Lou1, "Small vessel disease burden predicts functional outcomes in patients with acute ischemic stroke using machine learning," *CNS neuroscience & therapeutics,* 2023.

[6] Enzhao Zhu, Zhihao Chen, Pu Ai , Jiayi Wang1 , Min Zhu , Ziqin Xu , Jun Liu1 and Zisheng Ai "Analyzing and predicting the risk of death in stroke patients using machine learning," *Frontiers in Neurology,* 2023.

[7] Soumyabrata Dev , Hewei Wang , Chidozie Shamrock Nwosu , Nishtha Jain , Bharadwaj Veeravalli , Deepu John "A predictive analytics approach for stroke prediction using machine learning and neural networks.," *Healthcare Analytics 2,* 2022.

[8] Lea Fast1 , Uchralt Temuulen2 , Kersten Villringer2 , Anna Kufner, Huma Fatima Ali , Eberhard Siebert, Shufan Huo , Sophie K. Piper, Pia Sophie Sperber , Thomas Liman, Matthias Endres and Kerstin Ritter "Machine learning-based prediction of clinical outcomes after first-ever ischemic stroke.," *Frontiers in neurology* , 2023.

[9] Olusola Olabanjo, Ashiribo Wusu, Oseni Afisi and Boluwaji Akinnuwesi "Stroke Risk Factor Prediction Using Machine Learning Techniques: A Systematic Review.," *ournal of Applied Sciences,* 2024.

[10] Mylapalli Kanthi Rekha, I. Phani Kumar "Brain Stroke Prediction Using Random Forest And Adaboost Algorithm," 2023.

[11] Giuseppe Miceli , Maria Grazia Basso , Giuliana Rizzo , Chiara Pintus , Elena Cocciola , Andrea Roberta Pennacchio and Antonino Tuttolomondo. "Artificial intelligence in acute ischemic stroke subtypes according to Toast classification: a comprehensive narrative review," *Biomedicines* , 2023.

[12] Mr. Dilesh Yuvraj Bagul , Dr. P. B. Bharate , Dr. Aarti Sahasrabuddhe , "Use of Robust Machine Learning Approach in Prediction of Stroke.".