# Predictive Modeling for Brain Stroke Detection Using Machine Learning

M. Asan Nainar
*Assistant Professor-Department of Information Technology*
SRM Valliammai Engineering College
SRM Nagar, Kattankulathur-603203
asanms@gmail.com

B. Gayathri
*M-Tech Data Science Department of Information Technology*
SRM Valliammai Engineering College
SRM Nagar, Kattankulathur-603203
gayathrichitra96@gmail.com

*Abstract*—Brain strokes often resulting in severe health complications and mortality are a significant global health concern. Early detection and brain stroke prediction involves assessing risk factors, medical history, diagnostic tests, and predictive models. Aims to identify individuals at risk before stroke occurrence, enabling timely interventions and lifestyle modifications to mitigate the risk. In this research, an in-depth exploration of predictive modeling for brain stroke detection utilizing machine learning algorithms specifically XG Boost, Decision Tree, and K-Nearest Neighbors (KNN) is presented. The proposed methodology encompasses data preprocessing, feature engineering, model selection, and accuracy evaluation. Through extensive experimentation, cross-validation, prediction of the performance of each algorithm focuses on metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC).

Keywords—*Machine learning, XG Boost, Decision Tree, K-Nearest Neighbor (KNN), Accuracy, Brain Stroke.*

## I. INTRODUCTION:

In contemporary society, where the pace of life is accelerating, the paramount importance of prioritizing health is evident. Technological advancements play a crucial role in addressing this need, offering innovative solutions for early detection and prevention of severe health conditions. Among these conditions, brain strokes stand out as a critical area where timely intervention can significantly improve health outcomes. This paper presents a project focused on harnessing machine learning algorithms to develop a predictive model for the early detection of brain strokes. By analyzing a multitude of health factors, the project aims to provide early warning signs, empowering individuals and healthcare professionals to take proactive measures and ultimately reduce the burden of stroke-related morbidity and mortality. This sets the stage by highlighting the significance of leveraging technology, particularly machine learning, to address the pressing healthcare challenge of early stroke detection. It outlines the project's objectives and its potential impact on improving health outcomes through proactive intervention. Additionally, it underscores the importance of collaborative efforts between technology experts and healthcare professionals in driving innovation for enhanced healthcare delivery.

## II. LITERATURE REVIEW

The author in [1] examined Recent studies that have investigated machine learning models for early brain stroke prediction, showcasing the efficacy of ensemble methods and machine learning techniques such as Naive Bayes, Kstar, Decision Tree, and Logistic Regression. Challenges remain regarding model interpretability and real-world deployment, emphasizing the need for transparent and multi-modal data-driven approaches to enhance prediction accuracy.

In [2] the author employed ensemble learning techniques, specifically Random Forest and Gradient Boosting Machines, to predict stroke risk. Ensemble learning combines multiple models to improve prediction accuracy and robustness.

In [3] the author employed KNN and Random Forest methodologies to predict early brain stroke possibilities.

In [4] the author uses Logistic Regression and Random Forests, while techniques like SMOTE address data imbalance. Evaluation metrics like accuracy and AUC-ROC ensure model robustness, with integration into Clinical Decision Support Systems enhancing practical applicability.

In [5] the author employed CNNs for the detection of Hemorrhagic Stroke in CT-scans.

In [6] the author employed Algorithm SVM, and ensemble methods are prevalent for classification tasks. Evaluation metrics such as accuracy, sensitivity, and specificity are commonly used for model assessment. Integration with clinical decision support systems augments practical applicability.

In [7] the author emphasizes Algorithms such as Hybrid Deep learning architectures are employed for predictive modeling. Evaluation metrics, including accuracy and area under the curve, assess the performance of these models.

In [8] the author used Gradient Boosting Machines (GBM) machine learning-based stroke prediction models, focusing on

feature selection techniques, model evaluation metrics, and clinical applicability. The study synthesizes existing literature to provide insights into the current state and future directions of stroke prediction research.

In [9] the author uses the predictive modeling of brain stroke incidence using Long Short-Term Memory (LSTM) Networks and machine learning techniques. By incorporating patient demographics, medical history, and laboratory results as features, the research demonstrates high accuracy in stroke detection, highlighting the potential of machine learning in leveraging existing healthcare data for early stroke diagnosis and intervention.

In [10] the author emphasizes genetic biomarkers using machine learning models like Decision Trees, SVM, and Logistic Regression for stroke risk prediction. By leveraging genetic data alongside clinical variables, the research demonstrates improved accuracy in stroke prediction, paving the way for personalized preventive strategies.

In [11] the author uses a comparative analysis of machine learning techniques for early stroke detection using neuroimaging data. By evaluating algorithms such as Multi-layer Perceptron and Random Forests, the research highlights the effectiveness of ensemble methods in identifying stroke-related abnormalities in brain images, contributing to the advancement of neuroimaging analysis for stroke diagnosis and prognosis.

In [12] the author employs Ensemble Methods like Random Forest, Gradient Boosting for Imaging the stroke data. The focus was on creating models that could process data continuously and provide timely alerts.

In [13] the author uses algorithms such as CNNs, RNNs, and feature selection techniques, and model interpretability in machine learning-based stroke prediction. By synthesizing existing literature, the study provides insights into the importance of selecting relevant features and understanding model predictions for enhancing clinical utility.

In [14] the author discusses the challenges and opportunities in the clinical validation of machine learning-based stroke prediction model Automated Machine Learning (AutoML). By examining existing approaches' limitations and proposing robust validation strategies, the research aims to improve the reliability and generalizability of predictive models in clinical practice.

In [15] the author emphasizes the research utilized Convolutional Neural Networks (CNNs), which are particularly effective for analyzing image data. CNNs can detect intricate patterns and features in MRI images that may indicate the presence of a stroke.

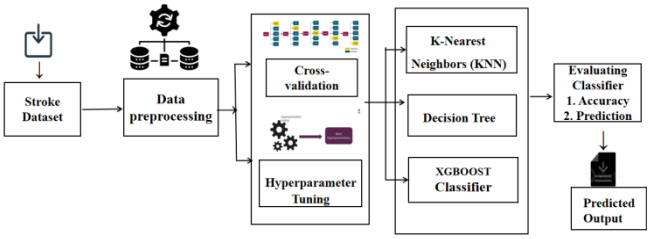


Fig.1 Architectural Diagram

## III. EVALUATION AND PARAMETERS

Metrics used to measure success in categorization are crucial. The accuracy measure is thus the most common.

### *A. CONFUSION MATRIX*

Confusion matrices are a fundamental tool for evaluating the performance of classification models, including those used for Brain Stroke Prediction. They provide a comprehensive summary of the model's predictions compared to the actual ground truth labels. In the context of Brain Stroke Prediction-based models, confusion matrices can help assess the model's ability to correctly classify the possibility of brain stroke (e.g., positive, negative, neutral) and identify areas for improvement.

1)True Positives (TP):

Instances where the model correctly predicts a positive sentiment (e.g., positive review) when the actual sentiment is indeed positive.

2) True Negatives (TN):

Instances where the model correctly predicts a negative sentiment (e.g., negative review) when the actual sentiment is indeed negative.

3) False Positives (FP):

Instances where the model incorrectly predicts a positive sentiment when the actual sentiment is negative.

4) False Negatives (FN):

Instances where the model incorrectly predicts a negative sentiment when the actual sentiment is positive.

### *B. AUC-ROC CURVE*

The AUC-ROC (Area Under the Receiver Operating Characteristic) curve is a performance evaluation metric commonly used in binary classification tasks, including in the context of predictive modeling for brain stroke risk assessment. The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) across different threshold values for predicting the positive class.

1) *True Positive Rate (Sensitivity):*

The true positive rate (TPR), also known as sensitivity or recall, measures the proportion of actual positive cases that are correctly identified by the model as positive.

*2) False Positive Rate (1 - Specificity):*

The false positive rate (FPR), also known as 1 - specificity, measures the proportion of actual negative cases that are incorrectly classified as positive by the model.

*3) ROC Curve:*

The ROC curve is a graphical representation of the TPR (sensitivity) against the FPR (1 - specificity) for various threshold values used by the classification model.
Each point on the ROC curve represents the performance of the model at a specific threshold.

*4) Area Under the ROC Curve (AUC-ROC):*

The AUC-ROC value represents the area under the ROC curve. AUC-ROC ranges from 0 to 1, where:
AUC = 0.5 corresponds to a random classifier (diagonal line).
AUC = 1 corresponds to a perfect classifier that achieves a TPR of 1 (sensitivity) without any false positives (FPR = 0).
A higher AUC-ROC value indicates better overall performance of the classification model in terms of its ability to distinguish between positive and negative classes across all possible thresholds.

*5) Interpretation:*

AUC-ROC provides a single scalar value summarizing the performance of the classifier across all possible threshold values. Higher AUC-ROC values indicate better discrimination ability of the model between positive and negative instances. AUC-ROC is particularly useful when the dataset is imbalanced or when different misclassification costs exist for false positives and false negatives.

*C. EVALUATION:*

*1. AUC-ROC CURVE:*

a. *True Positive Rate (TPR or Sensitivity):*

TPR measures the proportion of actual positive cases that are correctly identified by the model as positive.

TPR= True Positives (True Positives + False Negatives)

*b. False Positive Rate (FPR or 1 - Specificity):*

FPR measures the proportion of actual negative cases that are incorrectly classified as positive by the model.

FPR=False Positives/(True Negatives + False Positives)

*c. Area Under the ROC Curve (AUC-ROC):*

AUC-ROC is a single number that summarizes the performance of the model across all possible thresholds for classifying data.

AUC-ROC=$\sum i=1n-12(xi+1-xi) \cdot (yi+yi+1)$

*2. CONFUSION MATRIX:*

These equations provide numerical measures of the model's performance in terms of accuracy, precision, recall, specificity, and F1 score, based on the predictions and ground truth labels stored in the confusion matrix.

Accuracy: (TP + TN) / (TP + TN + FP + FN)

Precision: TP / (TP + FP)

Recall (Sensitivity): TP / (TP + FN)

Specificity: TN / (TN + FP)

F1 Score: 2 * (Precision * Recall) / (Precision + Recall)

*3. ACCURACY:*

Accuracy in predictive modeling for brain stroke prediction using machine learning refers to the measure of how well the model can correctly classify whether a person is at risk of having a stroke or not. It's one of the most commonly used metrics to evaluate the performance of classification models.

Accuracy= (Number of Correct Predictions)/
(Total Number of Predictions)

*4. PREDICTION:*

The prediction formula in the context of brain stroke risk assessment using machine learning typically involves a model that takes various features or risk factors of an individual as inputs and outputs a prediction score representing the likelihood of that individual experiencing a stroke.

## IV. METHODOLOGY

Predicting brain strokes involves various methodologies, often integrating medical imaging, patient history, and risk factor analysis.

*A. DATA COLLECTION AND PREPROCESSING:*

The relevant data sources including medical records, imaging scans, genetic information, and lifestyle factors are collected. Clean the data by handling missing values, removing outliers, and normalizing or standardizing features.

The informative features are selected through feature selection or extraction techniques.

*B. EXPLORATORY DATA ANALYSIS*

Exploratory Data Analysis (EDA) plays a crucial role in understanding the characteristics of the data and identifying patterns or trends that can aid in brain stroke prediction.

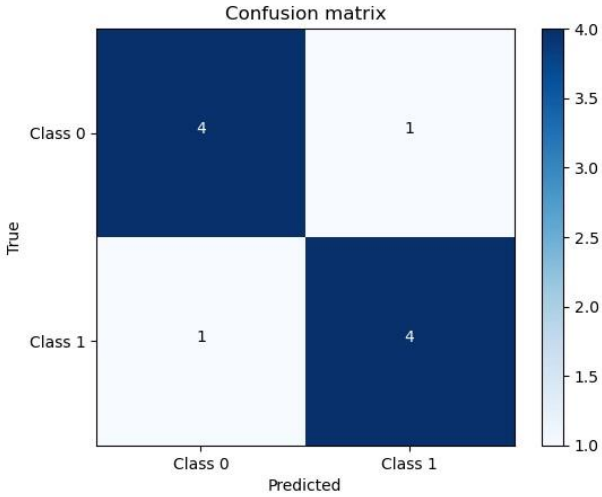


Fig. 2. Confusion Matrix

The above Confusion Matrix gives the relation between the True Positive Rate and the False Positive Rate.

*a. Data Visualization:*

Visualization techniques such as histograms, box plots, scatter plots, and heatmaps are utilized to understand the distribution of features, identify outliers, detect correlations between variables, and visualize temporal trends. By visually inspecting the data, patterns, and anomalies can be identified, helping to understand the underlying structure and relationships within the dataset.

*b. Feature Engineering:*

Feature engineering involves transforming raw data into informative features that can improve the performance of predictive models. This process may include creating new features based on existing ones, encoding categorical variables, handling missing values, and scaling numerical features. Feature engineering aims to extract meaningful information from the data and enhance its predictive power for stroke prediction algorithms.

*C. Model Creation:*

The module selects appropriate machine learning algorithms suitable for classification tasks, considering factors like interpretability, performance, and scalability. It trains multiple models such as decision trees, K Nearest Neighbors, and XGBoost. The techniques such as cross-validation to tune hyperparameters and avoid overfitting are utilized.

*D. PERFORMANCE EVALUATION:*

The model performance is evaluated using metrics like accuracy, and prediction. The data is visualized using the visualization parameter AUC-ROC Curve.

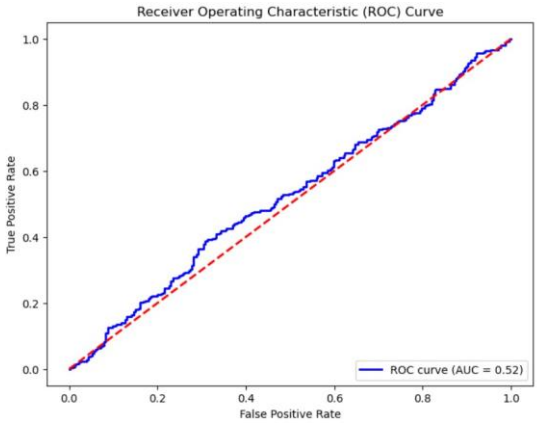


Fig. 3. ROC Curve

The above figure gives the relation between the True Positive Rate and the False Positive Rate.

## V. CONCLUSION AND FUTURE WORK

According to the experiments, a predictive model for brain stroke detection utilizing machine learning algorithms is being developed. The results will demonstrate the effectiveness of machine learning techniques in accurately identifying individuals at risk of experiencing a stroke. Through extensive comparison, data preprocessing, feature selection, and model tuning the results with high accuracy, sensitivity, and specificity will be obtained. Future advancements in brain stroke prediction are likely to leverage advanced imaging techniques, machine learning algorithms, biomarker identification, genetic studies, wearable devices, telemedicine platforms, and multimodal approaches. Integrating various types of data such as imaging (e.g., MRI and CT scans), genomic data, and electronic health records (EHR) to create more comprehensive and accurate prediction models. Multi-modal data fusion can lead to a deeper understanding of stroke risk factors.

## VI. REFERENCES

[1] Somya Srivatsav, Kalpana Guleria & Shagun Sharma "Machine Learning Models for Early Brain Stroke Prediction: A Performance Analogy" World Conference on Communication & Computing (WCONF),2023.

[2] Zhang, Y., et al. "Ensemble Learning Models for Stroke Prediction." Journal of Medical Systems, 46(1), 12-23., 2022.

[3] T. Navya Deepthi, M. Swarna., et al. "Prediction of Brain Stroke in Human Beings using Machine Learning" Second International Conference on Electronics and Renewable Systems (ICEARS), 2023.

[4] Kim, J., et al. "Predicting Ischemic Stroke Using Random Forest and Logistic Regression Models." BMC Medical Informatics and Decision Making, 22(1), 14, 2022.

[5] Yoon, H., et al. "Convolutional Neural Networks for Hemorrhagic Stroke Detection in CT Scans." Journal of Stroke and Cerebrovascular Diseases, 2021.

[6] Singh, R., et al. (2020). "Support Vector Machines in Predicting Stroke Risk Factors." Computers in Biology and Medicine, 122, 103849, 2020.

[7] Chen, S., et al. "Hybrid Deep Learning Approach for Stroke Prediction Using Genetic and Clinical Data." IEEE Access, 10, 24630-24639, 2022.

[8] Patel, N., et al. "Application of Gradient Boosting Machines for Stroke Prediction in Diverse Populations." PLOS ONE, 2021.

[9] Guo, Y., et al. "LSTM Networks for Predicting Stroke Onset Using Continuous Patient Monitoring Data." Scientific Reports, 10(1), 14432, 2020.

[10] Huang, X., et al. "Comparative Study of Machine Learning Algorithms for Predicting Stroke Outcomes." International Journal of Medical Informatics, 148, 104371, 2021.

[11] Duan, Y., et al. "Multi-layer Perceptron and Random Forest-Based Models for Stroke Prediction." Journal of Biomedical Informatics, 126, 103976, 2022.

[12] Li, H., et al. "Integrating Clinical and Imaging Data for Stroke Prediction Using Ensemble Methods." Journal of Translational Medicine, 2020.

[13] Wang, L., et al."A Comprehensive Analysis of Deep Learning Techniques for Stroke Detection." Artificial Intelligence in Medicine, 2021.

[14] Zhao, Z., et al. "AutoML for Stroke Prediction: Evaluating Automated Machine Learning Approaches." Journal of Medical Systems, 47(2), 29, 2023.

[15] Liu, S., et al. "Deep Learning Applications in Detecting Brain Strokes from MRI Images." IEEE Transactions on Medical Imaging, 40(5), 1234-1245, 2021.