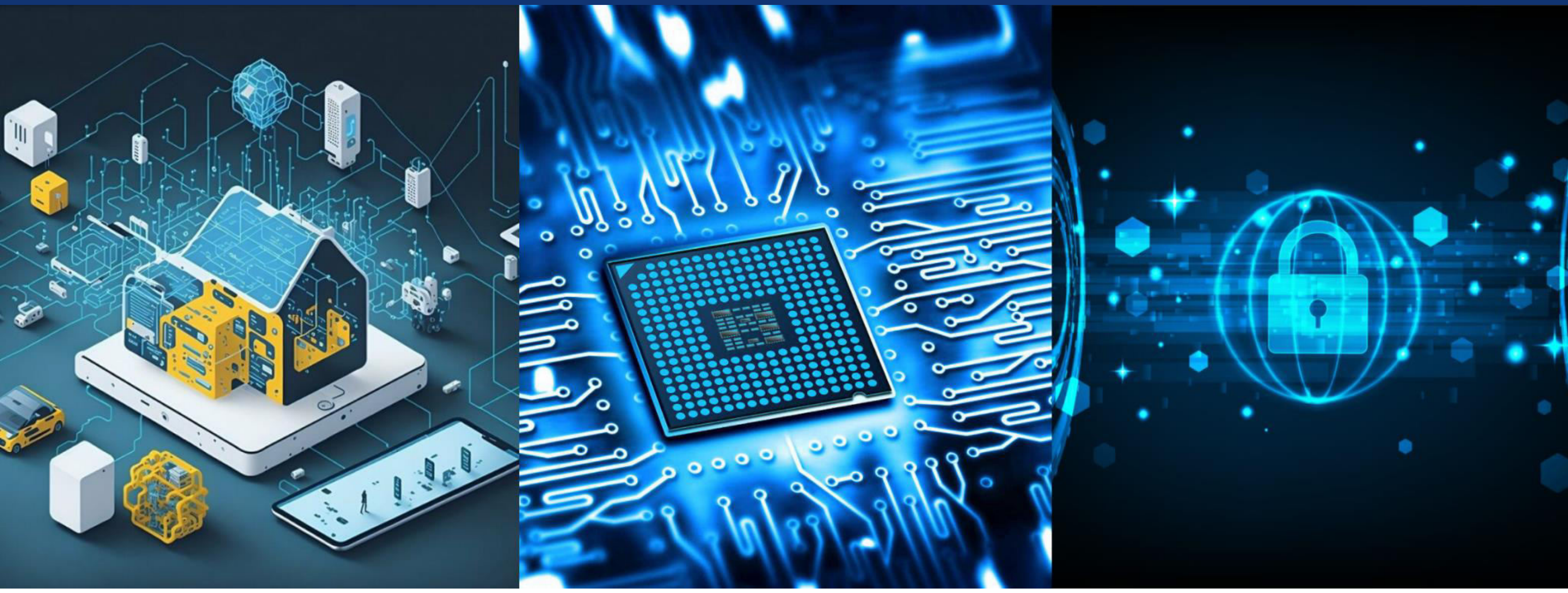
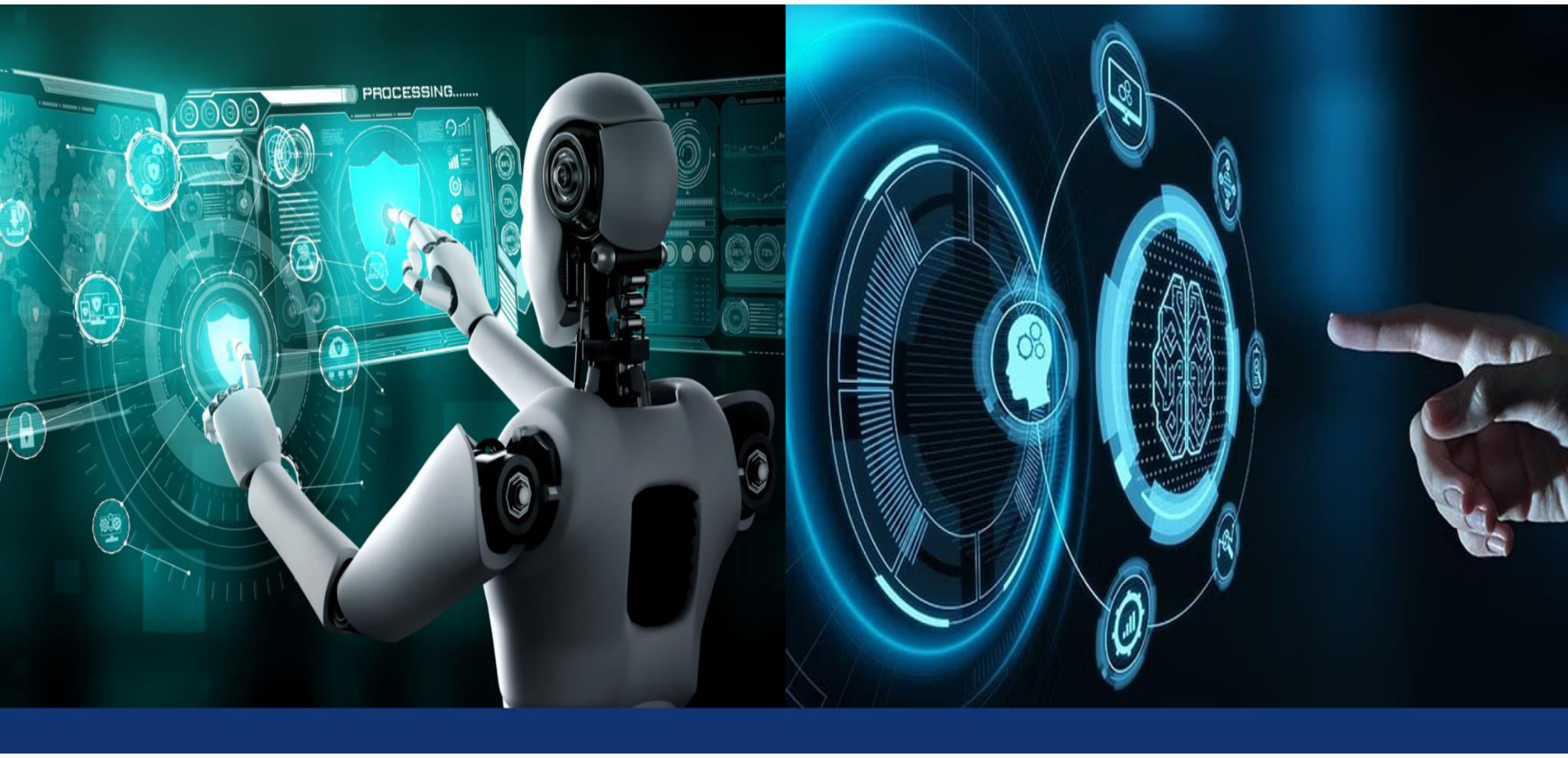
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

www.ijircce.com ijircce@gmail.com +91-9940572462 +91 63819 07438

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304091

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Machine Learning Based Stroke Prediction

Dr. G. Bala Krishna, N. Sai Sri Sowmya, K. Sandeepthi, M. Sai Teja, K. Shashank

Asst. Professor, Department of CSE, School of Engineering, Malla Reddy University, Hyderabad, Telangana, India

Department of CSE, School of Engineering, Malla Reddy University, Hyderabad, Telangana, India

ABSTRACT: Stroke is a major reason for disability and mortality across the globe, making initial prediction and intervention critical to reducing its impact. This project leverages machine learning techniques to develop a reliable and efficient system for predicting stroke risk. The model analyzes various clinical and lifestyle factors, including age, gender, hypertension, heart disease, cholesterol levels, and smoking habits to identify individual at high risk, The dataset used for training and evaluation is sourced from kaggle consisting of 5110 entries. The data preprocessing phase involves managing missing values, normalization, and feature selection to enhance model accuracy. Key features are identified using techniques like correlation analysis and Recursive Feature Elimination (RFE) to improve predictive performance the system implements advanced ML techniques implemented through python which is embedded with many diverse libraries which cater the user requirements to evaluate and redefine the performance of the applied ML technique parameters including accuracy precision recall and F1-score are considered. This project provides an affordable, scalable, and accurate solution for stroke prediction equipping healthcare professionals with actionable insights for timely interventions, reducing stroke-related fatalities, and improving patient outcomes.

KEYWORDS: Stroke Prediction, Machine Learning, Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks.

I. INTRODUCTION

Stroke is a primary health concern throughout the world and a major, chronic disorder. It happens when the brain's blood supply is either blocked or reduced, cutting off the oxygen supply to brain cells, which can lead to severe damage. Timely identification and preventive measures play a crucial role in minimizing severe complications associated with stroke. With advancements in Machine Learning (ML) algorithms, healthcare professionals can analyze vast amounts of patient data to predict the likelihood of stroke. By leveraging supervised learning techniques, ML models can identify patterns in medical history, lifestyle habits, and demographic factors to classify individuals as high-risk or low-risk for stroke.

This study aims to develop a stroke prediction system using machine learning (ML) techniques, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks. The WHO stares that almost 15 million individuals experience a stroke each year, with fatalities occurring every four-five minutes. The model will be trained on a dataset containing key patient attributes including factors like age, hypertension, heart disease, glucose levels, BMI, and smoking status to enhance prediction accuracy.

This study utilizes a dataset sourced from Kaggle, incorporating diverse physiological attributes for analysis and prediction. The extracted features are analyzed and utilized for the final prediction. Prior training the machine learning model, the dataset undergoes Data Preprocessing, which includes cleaning and preparation to enhance model interpretability. The process includes handling missing values, encoding categorical values through Label Encoding and One-Hot Encoding, and ensuring data consistency. Once the preprocessing is done, the dataset will be divided into training and testing subsets. Different algorithms are then implemented to classify and develop predictive models for stroke detection and their performance is assessed using accuracy metrics to determine the most effective model.

Integrating ML-driven methodologies enhances early diagnosis, facilitates risk assessment, and supports timely medical needs, thereby improving patient care and easing healthcare system challenges. Strokes are categorised majorly into two main types: ischemic and hemorrhagic. Ischemic strokes occur because of narrowed arteries that supply oxygenated blood to the brain, while hemorrhagic strokes occurs because of rupture of a weakened blood vessel, causing internal bleeding in the brain. Preventing strokes requires adopting a healthy lifestyle, minimizing risk factors such as smoking and

excessive alcohol intake, maintaining a balanced BMI and stable blood glucose levels, and ensuring proper heart and kidney function.
Early stroke prediction is crucial, as timely detection can help prevent severe complications, including long-term disability or fatal outcomes.

The final-year students made significant contributions to the research, with one handling data collection and preprocessing, including cleaning and encoding dataset attributes, another focusing on model selection and implementing various classification algorithms, a third evaluating performance and comparing model accuracies, and the last member working on report writing, visualization, and interpretation of results. The guide provided essential expertise on machine learning methodologies, structured the project, and validated the final model to ensure accuracy and reliability. This collaborative effort contributed significantly to the evolution of an effective stroke prediction system.

II. LITERATURE SURVEY

In this paper [2], Five machine learning models were applied to predict strokes using data from the Cardiovascular Health Study (CHS) dataset. The most effective approach combined Decision Trees with the C4.5 algorithm, Principal Component Analysis (PCA), Artificial Neural Networks (ANNs), and Support Vector Machines (SVMs) to enhance prediction accuracy. However, the limited number of input parameters in the CHS dataset impacted the model's overall efficiency.

In this article [3], they explored stroke prediction by analyzing user-generated content from social media. The authors utilized the DRFS method to detect potential symptoms linked to stroke. Additionally, Natural Language Processing (NLP) was applied to extract textual data; however, this significantly heightened the model's computational complexity and execution time.

The study focused on evaluating stroke severity using an improved version of the Random Forest algorithm. This technique effectively assessed stroke risk levels and demonstrated superior performance compared to existing models. However, its scope was restricted to specific stroke types, limiting its applicability to newly emerging stroke variations [4].

They developed a stroke prediction model utilizing Decision Trees, Random Forest, and Multi-layer Perceptrons. The study observed that all three models delivered comparable accuracies, with Decision Trees achieving 74.31%, Random Forest 74.53%, and Multi-layer Perceptrons performing slightly better at 75.02% [5].
Amini and others [6] conducted a study on stroke prediction, collecting data from 807 individuals, both healthy and at risk. They identified 50 major risk factors, including diabetes, cardiovascular conditions, smoking, and hyperlipidemia. The most accurate techniques used in their study were the C4.5 Decision Tree algorithm, which achieved 95% accuracy, and the K-Nearest Neighbor (KNN) model, which reached 94%.

In this research [7] utilized the Cardiovascular Health Study (CHS) dataset for stroke prediction, implementing A novel automated feature selection method is introduced, leveraging the conservative mean principle to enhance model performance. This approach systematically evaluates feature importance by considering statistical stability and relevance, ensuring that only the most significant attributes are retained for optimal predictive accuracy. The selected features were then integrated with a Support Vector Machine (SVM) algorithm to enhance predictive accuracy. However, the increased number of vectors generated in this process led to performance inefficiencies.

[8] The researchers used Artificial Neural Networks (ANNs) to predict thrombo-embolic strokes, leveraging their ability to identify complex patterns in medical data and enhance diagnostic accuracy. The model employed the Backpropagation algorithm to optimize stroke predictions, achieving approximately 89% accuracy. However, the computational demands of neural networks, particularly in terms of training time and processing complexity, posed significant challenges.

III. EXISTING SYSTEM

The existing system for stroke prediction primarily relies on traditional statistical models and risk assessment tools used in clinical settings. A widely used approach for stroke risk assessment is the Framingham Stroke Risk Score, which estimates the likelihood of stroke by considering factors such as age, blood pressure, diabetes, smoking status, and cardiovascular conditions. Similarly, clinical scoring models like the CHA_2DS_2-VASc score help assess stroke risk in patients with atrial fibrillation. These traditional methods, while widely used, have limitations in accuracy, personalization, and adaptability to diverse patient populations. In addition to statistical models, hospitals and healthcare institutions rely on electronic health records (EHRs) to track patient history and risk factors. Physicians analyze this data manually or using basic rule- based algorithms to estimate stroke risk. However, these systems often struggle with missing data, subjective clinical judgment, and lack of real-time predictive capability. With advancements in technology, imaging-based stroke detection has become more prevalent. Computed Tomography (CT) scans and Magnetic Resonance Imaging (MRI) are used to detect strokes after they occur, but they do not help in predicting stroke risk beforehand. Some AI-powered diagnostic tools, such as those developed using deep learning, assist in identifying strokes in medical images, but their implementation in real-world healthcare settings is still in progress. The primary drawbacks of the existing system include its reliance on historical data and static risk factors, limited predictive accuracy, and the inability to adapt dynamically to new patient information. These limitations have led to the increasing adoption of machine learning models, which can analyze large datasets, incorporate real-time data, and improve prediction accuracy. By leveraging artificial intelligence, future Stroke Risk Assessment System aim to address the shortcomings of traditional techniques, enhance early diagnosis and prevention.

IV. PROPOSED SYSTEM

The proposed stroke prediction model leverages AI-driven learning to enhance accuracy, efficiency and real-time decision-making. Unlike traditional statistical methods, machine learning models can analyze vast patient datasets, uncover complex relationships and continuously refine their predictive capabilities. This system integrates structured data from electronic health records, medical imaging and wearable devices to deliver a holistic stroke risk assessment. As its core, the model employs an automated ML pipeline encompassing data preprocessing, feature selection, model training and performance evaluation. The preposing phase includes handling missing data, standardizing medical records and mitigating class imbalances using techniques such as the Synthetic Minority Over-Sampling Technique(SMOTE).

To maximize predictive accuracy, the system utilizes a range of machine learning models , including logistic regression, decision trees, random forests, support vector machines(SVM) and XGBoost, aw well as deep learning approaches like artificial neural networks(ANNs) and convolutional neural networks(CNNs). Furthermore, it incorporates explainable AI techniques such as SHAP(Shapely additive Explanations) and LIME(Local Interpretable Model-Agnostic Explanations) to improve transparency and highlight key risk factors affecting stroke predictions.

Additionally, IoT-enabled wearable devices, such as smartwatches that monitor blood pressure, heart rate, and physical activity, facilitate continuous health tracking. These devices provide real-time alerts to healthcare professionals when abnormal readings are detected. This advanced system aims to surpass traditional methods by delivering greater accuracy, real-time stroke risk assessment, and seamless connectivity with hospital management systems. By leveraging cloud-based deployment and mobile applications, it ensures accessibility for both patients and healthcare providers, fostering proactive healthcare management.

electronic health records, medical imaging and wearable devices to deliver a holistic stroke risk assessment. As its core, the model employs an automated ML pipeline encompassing data preprocessing, feature selection, model training and performance evaluation. The preposing phase includes handling missing data, standardizing medical records and mitigating class imbalances using techniques such as the Synthetic Minority Over-Sampling Technique(SMOTE).

To maximize predictive accuracy, the system utilizes a range of machine learning models , including logistic regression, decision trees, random forests, support vector machines(SVM) and XGBoost, aw well as deep learning approaches like artificial neural networks(ANNs) and convolutional neural networks(CNNs). Furthermore, it incorporates explainable AI techniques such as SHAP(Shapely additive Explanations) and LIME(Local Interpretable Model-Agnostic Explanations) to improve transparency and highlight key risk factors affecting stroke predictions.

Additionally, IoT-enabled wearable devices, such as smartwatches that monitor blood pressure, heart rate, and physical activity, facilitate continuous health tracking. These devices provide real-time alerts to healthcare professionals when abnormal readings are detected. This advanced system aims to surpass traditional methods by delivering greater accuracy,

real-time stroke risk assessment, and seamless connectivity with hospital management systems. By leveraging cloud-based deployment and mobile applications, it ensures accessibility for both patients and healthcare providers, fostering proactive healthcare management.

V. IMPLEMENTATION

A dataset is a systematically structured collection of data, where each **row** represents an individual record, and each **column** corresponds to a specific attribute or characteristic. In stroke prediction, datasets commonly include medical history, lifestyle choices, and clinical parameters that influence stroke risk. This project utilizes a dataset obtained from sources like Kaggle, containing thousands of patient records. The dataset comprises essential health indicators, including age, hypertension, heart disease, glucose levels, BMI, and smoking status, all of which are critical for evaluating an individual's stroke risk.

Machine Learning Algorithms for Stroke Prediction

Logistic Regression

Logistic Regression is a widely used statistical technique for binary classification, such as determining whether a patient is at high or low risk of stroke (0 = No, 1 = Yes). It utilizes a sigmoid function to estimate the probability of the target variable, making it effective in analysing the impact of factors like age and glucose levels on stroke risk.

Random Forest

Random Forest is an ensemble learning technique that constructs multiple decision trees using a method known as Bootstrap Aggregation (Bagging). This approach enhances model stability and accuracy by enhancing model robustness and adaptability. In stroke prediction, it generates several decision trees by selecting different subsets of the dataset. The final classification is derived by combining predictions from all decision trees, typically through majority voting and enhances accuracy while minimizing the risk of **overfitting**, ensuring more reliable results.

Decision Trees

Decision Trees are interpretable classification models that partition datasets into hierarchical branches based on feature values such as glucose levels and age. Each decision node represents a test on an attribute, leading to different outcomes. Although highly interpretable, Decision Trees tend to overfit the data, but this issue can be mitigated through pruning or by integrating them into ensemble techniques like Random Forest

Support Vector Machines (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm used for both classification as well as regression tasks. In stroke prediction, it determines the optimal hyperplane that separates high-risk and low-risk patients within a multidimensional space. By utilizing kernel functions such as the Radial Basis Function (RBF) and polynomial kernels, SVM effectively captures complex patterns and relationships in the data.

Gradient Boosting Classifier

Gradient Boosting is a powerful ensemble learning technique that enhances model performance by iteratively training weak learners, with each new model refining the errors of its predecessor. This method is particularly effective in stroke prediction, as it improves accuracy while identifying key risk factors through feature importance analysis.

Naïve Bayes

Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem, assuming that all input features are independent. This assumption simplifies computations while maintaining strong predictive performance in various applications. In stroke prediction, the model estimates the probability of a stroke by analyzing factors such as glucose levels, BMI, and smoking habits. Despite its simplified approach, Naïve Bayes is highly efficient and often delivers reliable results for binary classification tasks.

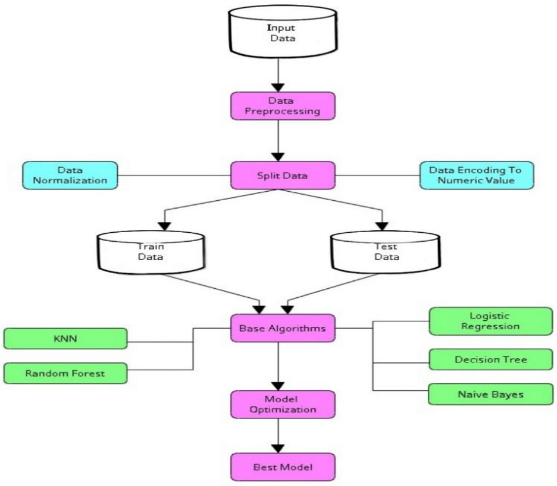


Fig. 1. Stroke Prediction Process

Fig. 1 outlines the step-by-step workflow for stroke prediction, highlighting the usage of various machine learning algorithms for classification and analysis.

VI. RESULTS

The stroke prediction results obtained from different self-learning methods, like Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Random Forest, showcase varying levels of accuracy and performance. Logistic Regression provides a baseline model with moderate accuracy, effectively handling linear relationships. The Decision Tree model, while interpretable, tends to overfit the data, leading to reduced generalization. SVM, known for its ability to find optimal hyperplanes, performs well with non-linear data but can be expensive. Among computationally these, the Random Forest model shows the highest performance and robustness by aggregating multiple decision trees, reducing overfitting, also improving prediction reliability. Overall, ensemble methods like Random Forest prove to be the most effective for stroke prediction, offering a balanced trade-off between accuracy and generalization.

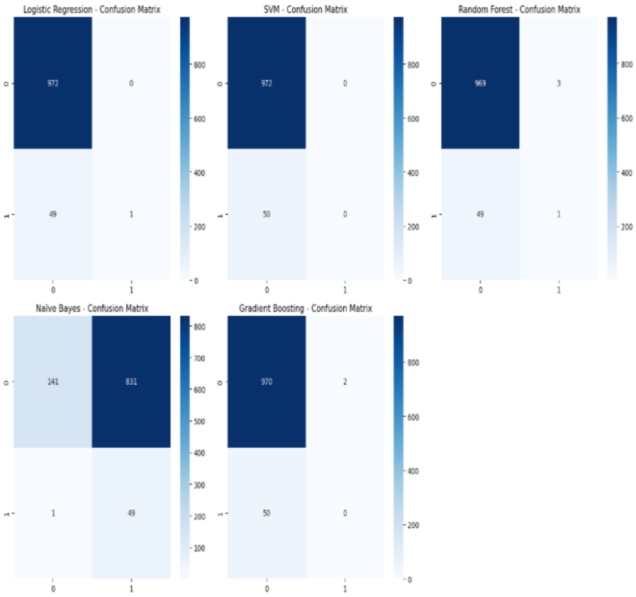


Fig. 2. Confusion Matrix for Stroke Prediction Models.

Fig. 2 illustrates confusion matrix, depicting the evaluation of machine learning models in classifying stroke predictions based on actual and predicted outcomes.

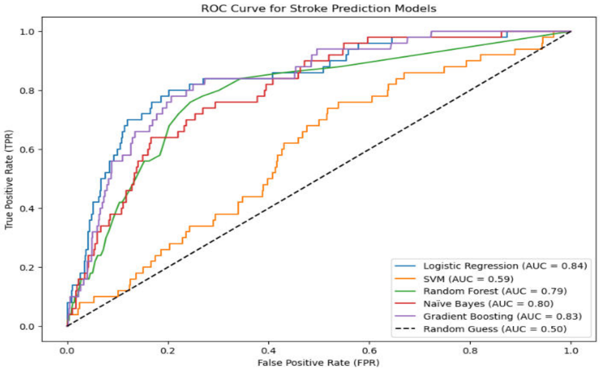


Fig. 3. ROC Curve for Stroke Prediction Models.

Fig. 3 shows the graphical representation of ROC Curve for Stroke Prediction Models.

VII. CONCLUSION

Stroke remains a major global health concern, necessitating initial recognition and prevention approaches in reducing its impact. Self-Learning algorithms offer a powerful approach for predicting stroke risk by analyzing patient medical history, lifestyle factors, also physiological attributes. This study analyzed the performance of four commonly used machine learning models—Random Forest, Decision Tree, Support Vector Machine (SVM), and Logistic Regression—to assess their effectiveness in stroke prediction. Among these, Random Forest achieved the highest accuracy and reliability due to its ensemble learning approach, which integrates multiple decision trees to control variance and enhance generalization. It efficiently handles missing data, identifies complex patterns, and performs well with large datasets, making it a robust choice for medical prediction.

In contrast, Decision Tree, while easy to interpret, is prone to overfitting, leading to lower generalization and reduced accuracy in real-world applications. SVM performed well in distinguishing between high-risk and low-risk patients, particularly in handling non-linear relationships, its drawback like Processing cost of ensemble methods is not suitable for practical for large datasets. Logistic Regression, though useful for baseline classification, lacks the ability to capture intricate feature interactions and non-linear dependencies, making it the least effective among the models tested. Given these findings, Random Forest ranked as the top suitable model for stroke identification, offering high performance and reliability, offering a balanced trade-off between accuracy, robustness, and scalability.

Future research can enhance prediction accuracy by incorporating feature selection techniques, hyperparameter tuning, and deep learning approaches such as neural networks. Additionally, hybrid models that integrate clinical, genetic, and real-time patient data could further refine stroke risk assessment. Deploying machine learning-based stroke prediction systems in cloud-based platforms can also support real-time monitoring and early intervention. In conclusion, Random Forest proves to be the most effective model for stroke prediction, equipping healthcare professionals with a dependable tool for early diagnosis and preventive care, despite its computational complexity.

REFERENCES

[1] Dataset named 'Stroke Prediction Dataset' from Kaggle:
https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.
[2] Singh, M.S., Choudhary, P., Thongam, K.: A comparative analysis for various stroke prediction techniques. In: Springer, Singapore (2020).

[3] Pradeepa, S., Manjula, K. R., Vimal, S., Khan, M. S., Chilamkurti, N., & Luhach, A. K.: DRFS: Detecting Risk Factor of Stroke Disease from Social Media Using Machine Learning Techniques. In Springer (2020).
[4] Vamsi Bandi, Debnath Bhattacharyya, Divya Midhunchakkravarthy: Prediction of Brain Stroke Severity Using Machine Learning. In: International Information and Engineering Technology Association (2020).
[5] Nwosu, C.S., Dev, S., Bhardwaj, P., Veeravalli, B., John, D.: Predicting stroke from electronic health records. In: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE (2019).
[6] Fahd Saleh Alotaibi: Implementation of Machine Learning Model to Predict Heart Failure Disease. In: International Journal of Advanced Computer Science and Applications (IJACSA) (2019).
[7] Ohoud Almadani, Riyad Alshammari: Prediction of Stroke using Data Mining Classification Techniques. In: International Journal of Advanced Computer Science and Applications (IJACSA) (2018).
[8] Kansadub, T., Thammaboosadee, S., Kiattisin, S., Jalayondeja, C.: Stroke risk prediction model based on demographic data. In: 8th Biomedical Engineering International Conference (BMEiCON) IEEE (2015).
[9] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, Honglak Lee:An Integrated Machine Learning Approach to Stroke Prediction. In: Proceedings of the 16th ACM SIGKDD International conference on Knowledge discovery and data mining (2010).
[10] Shanthi, D., Sahoo, G., Saravanan, N.: Designing an artificial neural network model for the prediction of thrombo-embolic stroke. Int. J. Biometric Bioinform. (IJBB) (2009).

INTERNATIONAL JOURNAL
OF INNOVATIVE RESEARCH
IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details