# Comparative Analysis of Machine Learning Models for Stroke Risk Prediction

Ziqian Gao
*Faculty of Arts and Science, University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3, Canada*

Keywords: Stroke Prediction, Machine Learning, Predictive Modelling.

Abstract: As the volume of medical data continues to grow rapidly, machine learning technologies have shown great promise in predicting the risk of stroke. Stroke remains a leading cause of disability and death worldwide, highlighting the importance of early and accurate risk prediction for effective prevention and management. This study aims to enhance stroke risk prediction by systematically evaluating the performance of various machine learning models, including Logistic Regression, Decision Trees, Random Forest, Gradient Boosting Classifier, and Support Vector Machines. The study systematically compares these models based on metrics such as accuracy, precision, recall, F1-score, and ROC-AUC values obtained from a well-preprocessed dataset. The results show that the Random Forest model outperformed the others, demonstrating higher accuracy and robustness, indicating its potential usefulness in clinical settings for early prediction of stroke risk. Future studies could explore more advanced data analysis techniques and consider incorporating newer models like neural networks to further enhance predictive performance.

## 1 INTRODUCTION

Stroke, a major contributor to serious long-term disability and the second leading cause of mortality globally represents a critical challenge in public health. Defined by the World Health Organization (WHO) as a major cause of mortality, stroke's impact on individuals and healthcare systems worldwide is profound (World Health Organization, 2022). This condition, characterized by the sudden loss of brain function due to disturbances in blood supply, necessitates a nuanced understanding of its multifactorial etiology to develop effective predictive and preventive strategies. The complexity of stroke, with its myriad risk factors ranging from genetic, lifestyle, and environmental to socio-economic and behavioral elements, calls for an integrated approach in research and healthcare practices (Spence JD, 2020).

The urgency of enhancing stroke prediction methodologies cannot be overstated, given its devastating impact on populations worldwide. Recent advancements in machine learning and data analysis have significantly influenced research in stroke prediction, offering new methodologies to assess risk factors and their interactions more comprehensively. The integration of big data analytics into medical research allows for a more profound analysis of risk factors, improving the predictive accuracy of stroke occurrences. Despite these technological advancements, the field continues to grapple with significant challenges. One of the foremost challenges is accurately capturing and integrating the complex interactions among genetic, environmental, and lifestyle factors that contribute to stroke risk. Additionally, there is the issue of class imbalance in datasets, which can skew predictive accuracy and model performance. This study specifically addresses these challenges by employing a variety of advanced machine learning techniques, such as SMOTE for handling class imbalance and ensemble methods like Random Forests and Gradient Boosting Classifier for capturing complex patterns in data. By systematically evaluating these models across multiple performance metrics, this research aims to enhance the robustness and accuracy of stroke prediction models, ultimately improving clinical decision-making and patient outcomes.

This study advances the field of stroke risk prediction through several key contributions. Firstly, it introduces a systematic evaluation of diverse machine learning models, such as logistic regression, decision trees, random forests, gradient boosting classifiers, and support vector machines, tailored



Gao, Z.
Comparative Analysis of Machine Learning Models for Stroke Risk Prediction.
DOI: 10.5220/0012998400004601

In *Proceedings of the 1st International Conference on Innovations in Applied Mathematics, Physics and Astronomy (IAMPA 2024)*, pages 93-101
ISBN: 978-989-758-722-1

specifically for stroke prediction. This evaluation provides a detailed comparison of model performance across multiple metrics—accuracy, precision, recall, and F1-score—offering a nuanced understanding of their practical applications in clinical settings. Secondly, the study employs advanced techniques like SMOTE to effectively address class imbalance, enhancing the reliability of predictions in minority classes. Additionally, the research underscores the importance of integrating various demographic, lifestyle, and medical attributes into predictive models, demonstrating a comprehensive approach to data preprocessing and feature engineering. These contributions collectively enhance the robustness, accuracy, and clinical relevance of stroke prediction models, paving the way for future research to incorporate even more sophisticated methods, such as neural networks, to achieve superior predictive performance and improved patient outcomes.

The arrangement for subsequent papers is as follows. Chapter 2 Review recent literature on stroke prediction. Chapter 3 provides a detailed process for constructing machine learning models. Chapter 4 analyzes the advantages and disadvantages, performance differences, clinical significance, and limitations of different methods from multiple indicators. Finally, a summary was provided for the entire article.

## 2 RELATED WORKS

In recent times, the intersection of machine learning and stroke prediction has seen remarkable advancements, as demonstrated by a variety of studies employing diverse data sources and analytical approaches to improve prediction models and patient outcomes.

One notable area of innovation involves the use of ensemble learning methods, such as Gradient Boosting Machine (GBM) and Extreme Gradient Boosting (XGB), explored by Xie et al. (2019). They specifically utilized these models to integrate clinical, demographic, and imaging data, achieving notable prediction accuracies. This method reflects a growing trend in leveraging complex datasets to refine predictive accuracy in acute medical settings.

Further advancing the field, Islam et al. (2022) introduced the use of EEG data in stroke prediction, applying explainable AI (XAI) frameworks to enhance transparency in AI decision-making processes. This study not only improved prediction accuracy but also provided insights into the model's reasoning, crucial for clinical acceptance. This approach aligns with the broader movement towards interpretability in machine learning, as seen in the work of Bhatt et al. (2023), who integrated federated learning within healthcare IoT frameworks to address data privacy and scalability challenges effectively.

On a different note, Grimaud et al. (2019) focused on the epidemiological aspects of stroke, analyzing how geographical and socio-demographic factors influence stroke outcomes. This study complements clinical and technical approaches by highlighting the importance of environmental and lifestyle factors, also evident in the work of Andersen and Olsen (2018) who examined how social determinants like marital status impact stroke risk. Similarly, another study by Shah et al. (2010) on the direct impact of smoking on stroke incidence reveals how lifestyle choices play a critical role in stroke risk, suggesting that predictive models should integrate these factors for a holistic risk assessment.

Moreover, the comprehensive reviews by Stephan et al. (2017) and Han et al. (2019) provide a broader context by discussing the implications of cognitive impairments and atrial fibrillation in stroke prediction. These studies underscore the necessity of incorporating a wide range of clinical indicators to enhance the specificity and reliability of predictive models.

Collectively, these studies illustrate a shift towards integrating diverse data types—from clinical and demographic data to personal health monitoring and lifestyle factors—into ML models. This integration aims not only to enhance predictive accuracy but also to tailor stroke management strategies to individual patient profiles, thereby advancing personalized medicine in neurology.

Each of these contributions supports a facet of stroke research, from enhancing model accuracy and transparency to incorporating broad epidemiological data, thus paving the way for a more integrated and nuanced approach to stroke prediction and management. The relationship among these studies underscores a comprehensive, multi-disciplinary approach to tackling stroke prediction, which is increasingly recognized as crucial for advancing patient care and outcomes in the field of neurology.

Using the stroke dataset from Kaggle, this essay aims to synthesize these diverse methodologies and data integrations, emphasizing how they collectively enhance the predictive accuracy of stroke outcomes. It seeks to demonstrate how the convergence of machine learning techniques, from the predictive models by Xie et al. (2019) and Islam et al. (2022) to the federated learning approaches by Bhatt et al. (2023), contributes to a more robust understanding of

stroke risks and outcomes, paving the way for advancements in personalized medicine in neurology.

However, despite these advancements, significant challenges remain. Current studies often focus on specific datasets or a limited range of features, which may not fully capture the complex interactions among genetic, environmental, and lifestyle factors contributing to stroke risk. This study addresses these gaps by employing a variety of advanced machine learning techniques to refine stroke prediction models using a comprehensive dataset that includes a wide range of features. By systematically evaluating the performance of different models, including Logistic Regression, Decision Trees, Random Forest, Gradient Boosting Classifier, and Support Vector Machines, this research aims to identify the most effective methods for early stroke risk prediction in clinical settings. The findings from this study are intended to provide a robust foundation for future research, potentially incorporating newer models like neural networks to further enhance predictive performance.

# 3 METHODOLOGY

The methodology deployed in this study includes a comprehensive strategy for predicting stroke risk based on a range of demographic, lifestyle, and medical attributes. It integrates several advanced machine learning techniques to construct and evaluate models capable of effectively identifying individuals at higher risk of stroke. The research presented herein delineates a multifaceted approach to the development of a predictive model for stroke risk, utilizing an array of machine learning techniques. The methodology is segmented into four distinct but interconnected stages: data preprocessing, exploratory data analysis (EDA), model development, and evaluation (Figure 1). This structured approach ensures that each phase builds upon the findings of the previous, culminating in the generation of a reliable predictive tool.

## 3.1 Data Preprocessing

Data preprocessing is a critical initial step in the analytical pipeline, focused on converting raw data into an appropriate format that improves the performance of machine learning models. The process began with the importation and cleaning of data, where missing values and inconsistencies were addressed. Specifically, the dataset revealed 201 missing values for the 'BMI' attribute, which were then addressed by median imputation to neutralize the effect of outliers.

Categorical variables such as 'Gender', 'Residence Type', 'Marital Status', and 'Smoking Status' underwent encoding to convert them into numerical formats suitable for machine learning algorithms. This encoding involved replacing categories with designated numerical values, enhancing the dataset's uniformity and suitability for subsequent analysis. Continuous variables were standardized using a StandardScaler to ensure that the model inputs had consistent scales and distributions, thereby preventing any variable from dominating the model's behavior due to its scale.

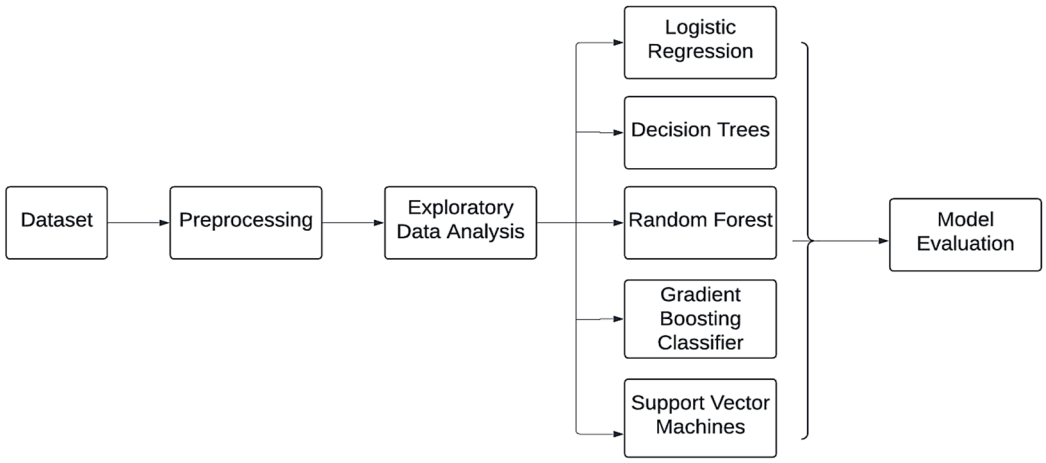


Figure 1: Workflow (Picture credit: Original).

To address the significant class imbalance observed in the stroke dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This method produces synthetic samples from the minority class, thus balancing the class distribution effectively. By doing so, SMOTE enhances the models' ability to detect and predict minority class outcomes, crucial for diseases like stroke where early detection is vital. This approach not only prevents the models from being biased towards the majority class but also improves the sensitivity and specificity of the predictive models used in the study.

## 3.2 Exploratory Data Analysis (EDA)

Exploratory data analysis was performed to unearth underlying structures and detect any anomalies. Histograms for all numerical features were plotted to understand distributional characteristics, which are crucial for the selection of appropriate statistical models and transformation techniques. The age distribution illustrated a fairly uniform distribution with slight right-skewness, indicating a wide range of participants in different age brackets (Figure 2).

A correlation matrix was utilized to identify potential multicollinearity and observe inter-variable relationships (Figure 3). Strong correlations between 'age' and 'ever_married', and between 'hypertension' and 'heart_disease', were noted. These findings were visually corroborated through a heatmap, reinforcing the necessity for careful feature selection to avoid multicollinearity in the predictive models.


Figure 2: Distribution of Stroke Risk Factors (Picture credit: Original).

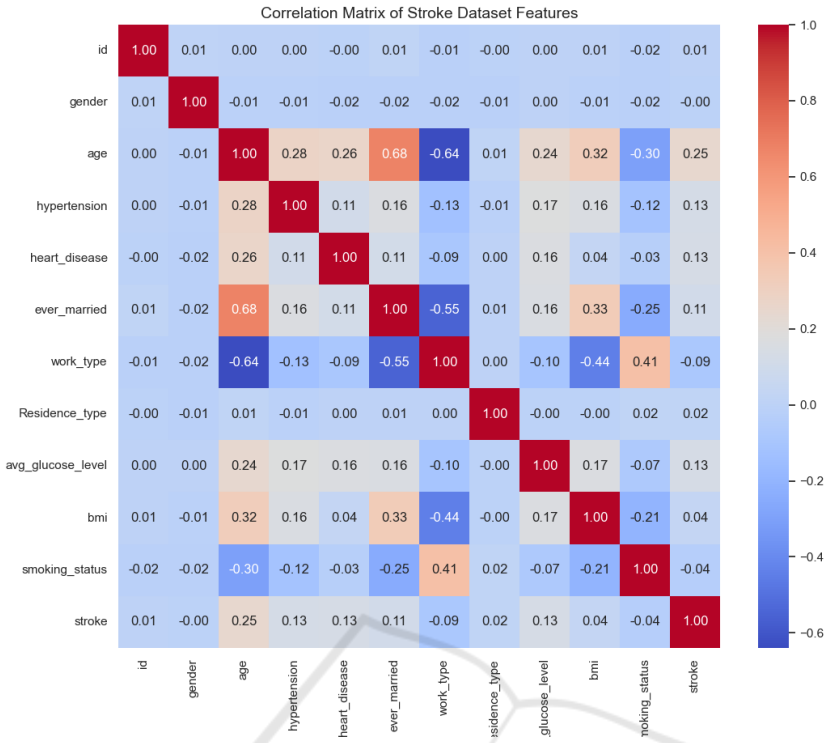


Figure 3: Correlation Matrix of Stroke Dataset Features (Picture credit: Original).

## 3.3 Model Development

The predictive modeling of stroke incidence requires precise and robust machine learning techniques that can effectively handle the intricacies of clinical data characterized by imbalances and high dimensionality. This study employs a suite of models, each chosen for their specific strengths in addressing different aspects of predictive accuracy and model interpretability in medical diagnostics.

Logistic Regression is foundational to the study due to its straightforward output of probability scores that indicate the likelihood of stroke. This model benefits clinical decision-making by providing clear, interpretable results that medical professionals can easily use to assess risk levels. The logistic model excels in situations where the relationship between the binary outcome and the independent variables can be approximated linearly in the logit scale, making it particularly suitable for initial risk assessments.

Decision Trees serve as an intuitive approach for partitioning the data into subsets according to the values of the explanatory variables, which in turn makes the decision-making process transparent. Trees inherently perform feature selection by choosing the most predictive attributes at each node, which simplifies the model by excluding non-informative variables. However, single trees can be prone to overfitting, especially with complex data, which necessitates the use of techniques to prune the trees or limit their growth.

Random Forests address the overfitting tendencies of single decision trees by introducing randomness in the selection of features and instances, and by averaging multiple trees to improve the generalization to new data. This method is highly valued in clinical settings for its robust performance across different types of data and its ability to handle large feature spaces without significant loss of accuracy.

Gradient Boosting Classifiers are included for their capacity to sequentially focus on difficult cases that previous models misclassified. This technique gradually improves the model's performance by combining weak learners into a strong learner, optimizing a differentiable loss function. Gradient boosting is particularly effective in enhancing predictive accuracy, especially in unbalanced datasets typical of medical outcomes like stroke, where negative cases far outnumber positive ones.

SVM are utilized for their effectiveness in classifying non-linearly separable data through the use of kernel functions. This capability allows SVMs to project data into higher-dimensional spaces where a linear separator might exist, making it an excellent tool for complex datasets with intricate patterns.

SVMs are favored for their high accuracy and the flexibility offered by various kernel choices, such as polynomial and radial basis function (RBF), which can be tuned to the specific data characteristics of stroke prediction.

In this study, the ensemble methods, specifically Random Forests and Gradient Boosting, are chosen not only for their high accuracy but also for their ability to provide insights into feature importance and model uncertainties. These properties are crucial for understanding the factors that drive predictions and for refining the models based on domain-specific knowledge in stroke research.

By integrating these diverse methodologies, the research aims to construct a comprehensive predictive model that leverages the unique strengths of each method. The ensemble approaches enhance model stability and accuracy, logistic regression offers simplicity and interpretability, decision trees provide a clear visualization of the decision paths, gradient boosting focuses on improving predictions iteratively, and SVMs offer robust classification capabilities. This multifaceted approach ensures that the predictive model is not only accurate but also adaptable to the complexities and variabilities inherent in medical data related to stroke.

# 4 EXPERIMENTAL SETUP AND RESULTS

The purpose of this section is to delineate the experimental setup utilized for model training and the subsequent results, which were pivotal in ascertaining the efficacy of various machine learning algorithms for stroke prediction.

## 4.1 Experimental Setup

The experimental framework was meticulously designed to provide an unbiased and rigorous assessment of the models. The dataset was partitioned into training and testing subsets using an 80:20 split, ensuring adequate data for model training while reserving a subset for assessment. Model training was conducted on a controlled computational environment to maintain consistency across experiments.

Model hyperparameters were selected based on preliminary tests and literature precedents to optimize each algorithm's performance. Each model was evaluated using cross-validation techniques on the training set to tune the parameters and prevent overfitting. The Python programming language, along with libraries such as Scikit-learn and imbalanced-learn, was employed to implement the algorithms and handle data manipulation and analysis tasks.

## 4.2 Model Training and Evaluation

Following model development, rigorous evaluation metrics were applied to assess each model's performance. Accuracy, a fundamental metric, offered an initial estimation of model performance. However, accuracy alone can be misleading, particularly in the presence of class imbalance. Therefore, confusion matrices were utilized to provide a more nuanced assessment, offering insights into the models' abilities to correctly predict each class.

Classification reports provided a detailed account of the precision, recall, and F1 scores for each model, allowing for the evaluation of models beyond mere accuracy. These scores are particularly critical in medical diagnostics, where the costs of false negatives and false positives can have significant implications.

Finally, ROC curve analysis was conducted for an aggregate evaluation of model performance across various threshold levels. Each model's AUC score was calculated, serving as a singular metric encapsulating the model's capacity to differentiate between classes. Models were ranked based on their AUC scores, with higher scores indicating superior performance in stroke prediction.

The evaluation process also considered the practical implications of model implementation. The complexity of the model, interpretability of results, and computational efficiency were factored into the selection of the most appropriate model for deployment in a clinical setting.

## 4.3 Results and Analysis

The effectiveness of the models was assessed through various metrics, including accuracy, precision, recall, and the F1-score. Each metric provides insight into different aspects of performance, crucial for a nuanced understanding of each model's strengths and limitations in the context of stroke prediction (Table 1 and figure 4).

Table 1: Result of stroke prediction.

| Method | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Logistic Regression | 75.24% | 93% | 75% | 82% |
| Decision Tree | 87.08% | 89% | 87% | 88% |
| Random Forest | 92.27% | 89% | 92% | 91% |
| Gradient Boosting Classifier | 89.63% | 91% | 90% | 90% |
| SVM | 73.68% | 93% | 74% | 80% |

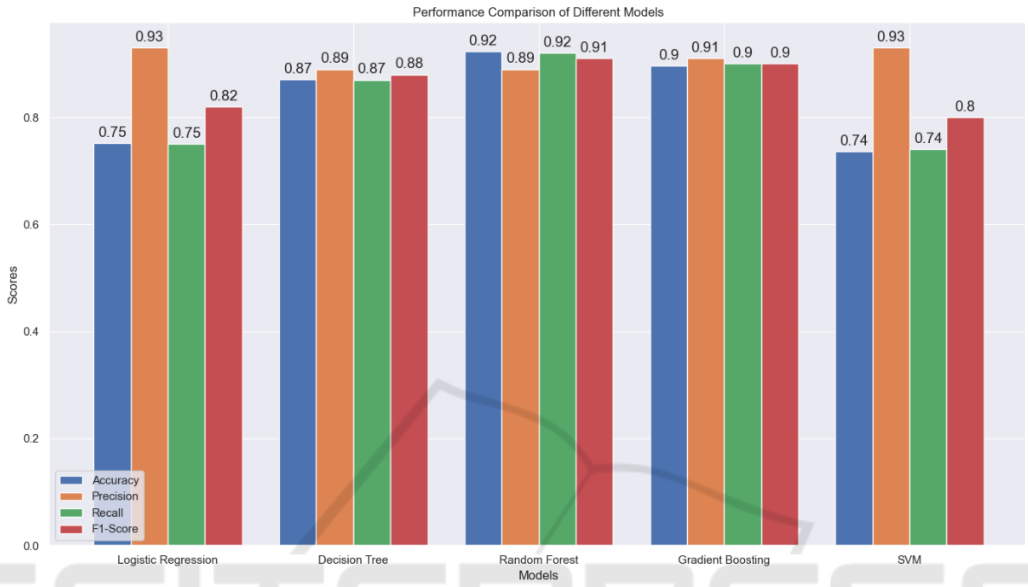


Figure 4: Performance Comparison of Stroke Prediction Models (Picture credit: Original).

In terms of accuracy, Random Forest (92.27%) is the highest-performing model due to its ensemble nature that combines multiple decision processes to reduce overfitting and biases. Decision Tree (87.08%) and Gradient Boosting Classifier (89.63%) follow, with the latter using a sequential corrective approach for classification refinement. Logistic Regression (75.24%) and SVM (73.68%) have lower accuracy, possibly due to their linear nature struggling with non-linearity or class imbalances in stroke data.

Precision measures a model's exactness in positive predictions. Logistic Regression and SVM score high in precision (93%), possibly from conservative prediction strategies sacrificing recall. Gradient Boosting Classifier (91%) and Random Forest (89%) maintain high precision without significant recall trade-off due to their complex structures detecting subtle data patterns.

Recall is critical in medical diagnostics. Random Forest leads with 92% recall, capturing a broad range of positive cases. Gradient Boosting Classifier (90%) focuses on prior errors to enhance sensitivity iteratively. Decision Tree's recall (87%) may be due to its unpruned nature capturing more positives at an overfitting risk. Logistic Regression and SVM have recall rates of 75% and 74%, needing additional measures for class imbalances or complex interactions.

The F1-score balances precision and recall, with Gradient Boosting Classifier (90%) as the top performer, followed by Random Forest (91%). Logistic Regression and SVM have F1-scores of 82% and 80%, indicating less effectiveness in identifying true positives.

The ROC-AUC score, measuring a model's class discrimination, is high for Logistic Regression and SVM (0.85), followed by Random Forest and Gradient Boosting Classifier (0.82 and 0.83). Decision Tree lags with an AUC of 0.54 due to simplicity and vulnerability to noise (Figure 5).

In summary, each model's performance is influenced by dataset challenges like class imbalance and feature dependencies. Ensemble methods excel in integrating multiple decision processes, crucial in stroke prediction, while simpler models may need advanced techniques for improved performance.

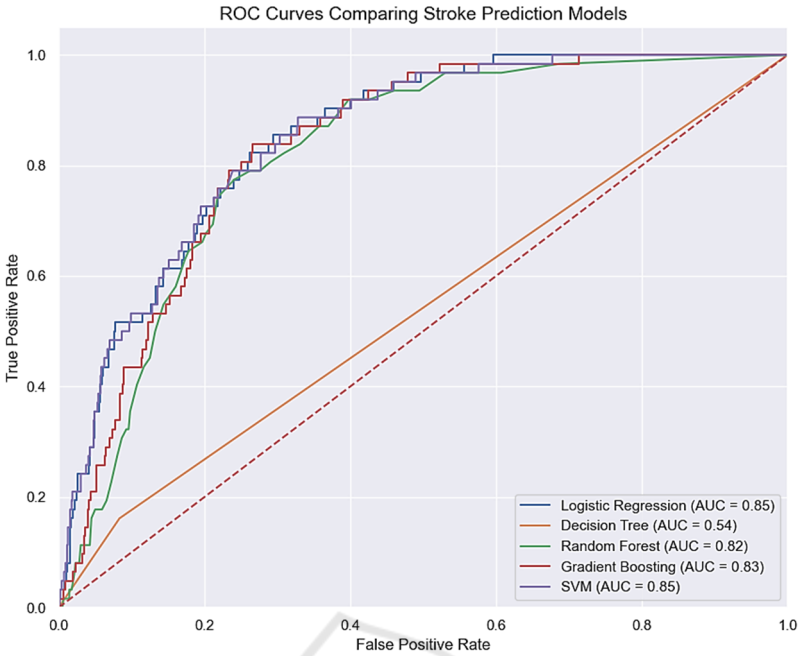


Figure 5: ROC Curves Comparing Stroke Prediction Models (Picture credit: Original).

### 4.4 Discussion

The experimental results revealed that while some models achieved high overall accuracy, their application in a clinical setting requires careful consideration of the trade-offs between various performance metrics. High accuracy may not always translate to clinical utility, particularly when the cost of false negatives is substantial, as in the case of stroke prediction. The ROC analysis provided a more comprehensive understanding, suggesting that Logistic Regression and SVM, despite their limitations in precision for stroke cases, offered a balanced discriminative ability across thresholds.

The ramifications of these results are significant. In clinical practice, the ability to accurately predict stroke cases could save lives and prevent long-term disabilities. Therefore, the selection of the appropriate model is not solely based on statistical performance but also on the clinical context and the consequences of predictive errors.

## 5 CONCLUSION

In summary, this study has successfully evaluated the performance of multiple machine learning models for stroke risk prediction, focusing on models including Logistic Regression, Decision Trees, Random Forest, Gradient Boosting Classifier, and Support Vector Machines. Among these, the Random Forest model stood out for its superior effectiveness, attributed to its robust handling of complex and imbalanced datasets crucial in medical diagnostics. The model's ability to aggregate multiple decision trees helps mitigate biases and overfitting, producing a more reliable and generalizable prediction tool.

Moving forward, future research could explore real-time data analytics and continuous monitoring of physiological parameters to better capture the temporal progression of risk factors. These advancements may lead to the development of dynamic prediction models that adjust predictions based on new data, potentially enhancing accuracy and clinical utility. By leveraging the findings of this study and incorporating cutting-edge research, future efforts can work towards transforming the landscape of stroke prevention. This proactive and personalized approach could significantly improve patient outcomes by predicting and mitigating risks more effectively.

## REFERENCES

World Health Organization. Stroke, Cerebrovascular accident. WHO EMRO; 2022. Available from: https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html.

Spence JD. Stroke Prevention. Stroke. 2020;51(7):2255–2262.

Xie Y, Jiang B, Gong E, Li Y, Zhu G, Michel P, Wintermark M, Zaharchuk G. Use of Gradient Boosting Machine Learning to Predict Patient Outcome in Acute Ischemic Stroke on the Basis of Imaging, Demographic, and Clinical Information. AJR Am J Roentgenol. 2019 Jan;212(1):44-51.

Islam MS, Hussain I, Rahman MM, Park SJ, Hossain MA. Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal. Sensors (Basel). 2022;22(24):9859.

Bhatt H, et al. Artificial Neural Network-Driven Federated Learning for Heart Stroke Prediction in Healthcare 4.0 Underlying 5G. Concurrency and Computation: Practice and Experience. 2023;36(3). doi: 10.1002/cpe.7911.

Grimaud O, Lachkhem Y, Gao F, Padilla C, Bertin M, Nowak E, Timsit S. Stroke Incidence and Case Fatality According to Rural or Urban Residence. Stroke. 2019;50(10):2661–2667.

Andersen KK, Olsen TS. Married, unmarried, divorced, and widowed and the risk of stroke. Acta Neuro Scandinavica. 2018;138(1):41–46. .

Shah RS, Cole JW. Smoking and stroke: the more you smoke the more you stroke. Expert Rev Cardiovasc Ther. 2010;8(7):917–932.

Stephan BCM, Richardson K, Savva GM, Matthews FE, Brayne C, Hachinski V. Potential Value of Impaired Cognition in Stroke Prediction: A U.K. Population-Based Study. J Am Geriatr Soc. 2017;65(8):1756–1762.

Han L, Askari M, Altman RB, Schmitt SK, Fan J, Bentley JP, Narayan SM, Turakhia MP. Atrial Fibrillation Burden Signature and Near-Term Prediction of Stroke. Circ Cardiovasc Qual Outcomes. 2019;12(10).

Stroke Prediction Dataset [Internet]. Kaggle. Available from:https://www.kaggle.com/code/mennatallah77/stroke-prediction-with-99-accuracy.