Journal of Computer Science

Original Research Paper

Gradient Boosting for Heart Stroke Prediction: Investigating Unexpected Risk Factors

Aniket Kailas Shahade and Priyanka V. Deshmukh

Department of AI & ML, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India

Article history
Received: 29-08-2024
Revised: 17-10-2024
Accepted: 08-10-2024

Corresponding Author:
Aniket Kailas Shahade
Department of AI & ML,
Symbiosis Institute of Technology, Pune Campus,
Symbiosis International (Deemed University), Pune, India
Email: aniket.shahade11@gmail.com

Abstract: Heart stroke prediction is a critical area in healthcare, aiming to identify individuals at risk and provide timely intervention. This research leverages machine learning algorithms, including Decision Tree, Random Forest, AdaBoost, and Gradient Boost, to predict the likelihood of stroke, with Gradient Boosting delivering the most accurate results. Our analysis uncovers intriguing and unexpected relationships between stroke risk and various factors such as heart disease, hypertension, and smoking habits. Contrary to conventional wisdom, our findings suggest that individuals with lower incidences of hypertension and heart disease exhibit increased stroke risk. Additionally, non-smokers appear to have a higher likelihood of experiencing a stroke compared to smokers. Furthermore, Body Mass Index (BMI), marital status, residence type, and work type also significantly influence stroke risk. These anomalous findings necessitate further investigation to understand the underlying causes and implications. This study highlights the importance of using advanced machine learning techniques to uncover complex patterns in health data, which can lead to more effective prevention strategies.

Keywords: Heart Stroke Prediction, Gradient Boosting, Machine Learning, Hypertension, Heart Disease, Smoking, Body Mass Index, Demographic Factors, Health Data Analysis, Risk Factors

Introduction

Heart stroke is still among the most prevalent causes of morbidity and mortality in the global population which creates a large healthcare cost. Prevention Priorities, therefore, have to focus on the early identification of high-risk persons in an effort to minimize the risks of experiencing a stroke. Machine Learning (ML) based predictive models provide an effective capability to enhance the stroke prognosis by detecting the high-order interactions that may not be captured by conventional statistical models (Hassan *et al*., 2024). The current research extends this line of work by using more sophisticated ML methods to investigate the risk factors of stroke using Gradient Boosting as the most optimal approach according to the results of the study.

Several authors have described the feasibility of using machine learning algorithms in predicting stroke. For instance, Breiman (2001) proposed a Random Forest algorithm for which the model has been found to be able to capture various functional forms of learning; Friedman (2001) further showed that GBMs outperform other models in complicated classification problems. Following studies conducted by Khushbu *et al*. (2024); Olaoye and Luz (2024) confirmed the efficiency of Gradient Boosting in the healthcare field and more specifically: Predicting stroke risk based on patients' clinical, genetic, and lifestyle characteristics. However, our study also extends beyond the findings of previous research in identifying several other non-obvious correlations between stroke risk and other health/ lifestyle factors that are contrary to clinical beliefs.

The findings of this research are focused on the correlation between stroke risk and such factors as heart disease and hypertension. Both have been used conventionally as significant markers of stroke (Kannel *et al*., 2004). But in contrast to this, our analysis showed that people who actually had lower risk factors of heart disease and even hypertension had a surprisingly higher risk of a stroke. This finding is in concordance with Graham *et al.* (2014) who noted that the combination and interaction of social and environmental factors may blur the clear expected correlation between traditional risk factors and stroke outcomes.

One more important observation that should be pointed out is the negative correlation between smoking and the rate of stroke. Smoking has been known for many years as one of the major risk factors for stroke (Kannel *et al*., 2004), but

our results showed that non-smokers in this sample had a greater chance of stroke than smokers. This may indicate substantial interdependence of smoking cessation, preexisting diseases, and the risk of stroke that needs further examination. Other works including Lee *et al*. (2017) have also observed such counterintuitive patterns and it has been argued that such patterns may be picking up on features that the machine learning algorithm has learned are relevant while epidemiological approaches would not be able to pick up.

Besides, the existence of these various relationships, this study found BMI, marital status, type of residence, and type of work that influence stroke risk. Wang *et al*. (2016) have also done a study on the effects of BMI on stroke especially among people with high obesity levels. We have similar observations in our research but the increased risk concerned individuals with a BMI between 20 and 50 which means that increased risk is not only among obese and underweight persons. Furthermore, marital status, living environment, and work characteristics were also significant predictors of stroke as evidence that stroke vulnerability is incorporated by social and environmental conditions (Graham *et al*., 2014).

The primary goal of this study is to achieve a level of pattern discovery in health data that is not easily possible with conventional data analytics methods using state-of-the-art machine learning algorithms for stroke risk assessment. It could also test a variety of ML such as the Decision Tree proposed by Quinlan (1986), Random Forest proposed by Breiman (2001), AdaBoost proposed by Freund and Schapire (1997), and lastly Gradient Boosting by Friedman (2001).

These findings suggest that the increased use of intuitive ML models is necessary for understanding the complex non-linear interactions of the risk factors under study. Such models allow us to step beyond the simple evaluation of risk factors in stroke, providing fresh views on the causes of this disease. The implication of these findings for clinical practice is that interventions can enhance the quality of individualized stroke prevention measures. Our study aligns with the emergent literature that seeks to enhance risk assessment by paying attention to the details of stroke prediction (Ahmed *et al*., 2024).

In the subsequent sections of this study, the reader shall find a detailed explanation of the method used in this study in terms of data acquisition, data pre-processing as well as model building. Moreover, the findings of the research will be shown as well as the authors' interpretation of the outcomes and further study recommendations. Finally, this study demonstrates how the use of machine learning could help to better understand the multifaceted nature of the risk factors associated with stroke and, as a result, improve prevention efforts.

Literature Review

Predictive modeling in healthcare has been significantly advanced through the application of Machine Learning (ML) techniques. These methods have enhanced the ability to identify patterns and predict outcomes based on complex, high-dimensional data. This literature review explores current research on stroke prediction using ML, focusing on key algorithms and their effectiveness, as well as the relationships between various risk factors and stroke incidence.

Machine Learning in Stroke Prediction

Machine learning techniques have been widely employed to improve stroke prediction models. Among these, Decision Trees, Random Forests, AdaBoost, and Gradient Boosting have shown substantial promise.

Decision Trees are simple, interpretable models that create a tree-like structure to make decisions based on input features. While effective for understanding the decision-making process, they are prone to overfitting, which can limit their generalizability (Quinlan, 1986).

Random Forests, an ensemble learning method, addresses overfitting by building multiple decision trees and combining their outputs. This method improves prediction accuracy and robustness by averaging the predictions from several trees, thus enhancing model performance (Breiman, 2001). Studies have demonstrated the effectiveness of Random Forests in various medical applications, including stroke prediction (Lee *et al*., 2017).

Adaptive Boosting (AdaBoost) enhances the performance of weak learners by focusing on misclassified instances and adjusting their weights iteratively. Research has shown that AdaBoost can improve prediction performance in healthcare applications by emphasizing difficult-to-predict cases, thus refining the overall model accuracy (Freund and Schapire, 1997).

Gradient Boosting, a powerful ensemble technique, builds models sequentially by optimizing a loss function. Each new model corrects the errors of its predecessor, leading to improved performance. Recent studies have highlighted the superiority of Gradient Boosting in various predictive tasks, including stroke prediction, due to its ability to handle complex relationships within data (Friedman, 2001; XGBoost, 2016). Table (1) discusses the summary of study conducted on heart stroke prediction.

Risk Factors for Stroke

The relationship between traditional risk factors such as hypertension, heart disease, smoking, and stroke is well-documented. However, recent findings suggest that these relationships may be more nuanced than previously thought.

Table 1: Summary of study on heart stroke prediction

Author(s)	Paper title	Methodology	Future scope	Problems identified
Chakraborty *et al*. (2024)	Predicting stroke occurrences: a stacked machine learning approach	Stacked machine learning, feature selection	Improve model accuracy with diverse datasets	Incomplete feature selection, imbalance in data
Olaoye and Luz (2024)	Comparative analysis of machine learning algorithms in stroke prediction	Comparative study of ML algorithms	Explore hybrid models and real-time data	Limited dataset variety, model complexity
Hassan *et al*. (2024)	Predictive modelling and identification of key risk factors for stroke	Machine learning, predictive modelling	Integrate genetic data for better accuracy	Lack of comprehensive data, limited genetic factors
Gupta *et al*. (2025)	Predicting stroke risk: An effective stroke prediction model based on neural networks	Neural networks, deep learning	Develop real-time applications in clinical settings	Lack of interpretability in model predictions
Khushbu *et al*. (2024)	Ensemble approach for stroke prediction using machine learning	Ensemble learning, feature selection	Focus on model optimization for faster predictions	Data imbalance, overfitting in complex models
Ahmed *et al*. (2024)	Enhanced stroke risk prediction: a fusion of machine learning models	Fusion of multiple machine learning models	Explore integration with wearable devices	Model complexity, data integration challenges
Teoh (2018)	Towards stroke prediction using electronic health records	EHR data, statistical modelling	Real-time stroke prediction systems	Data privacy concerns, quality of EHR data

Hypertension and heart disease have long been recognized as significant predictors of stroke. Elevated blood pressure and cardiovascular conditions contribute to vascular damage, increasing the likelihood of stroke. However, our study's findings challenge this conventional understanding, revealing that individuals with lower incidences of these conditions exhibited higher stroke risk. This anomaly suggests the need for further investigation into potential underlying mechanisms and interactions (Shahade *et al*., 2022a-b).

Smoking is another well-known risk factor for stroke. Numerous studies have established a strong correlation between smoking and increased stroke risk due to the harmful effects of tobacco on the vascular system (Kannel *et al*., 2004). Intriguingly, our research indicates that non-smokers may have a higher stroke risk compared to smokers. This unexpected result calls for a deeper exploration of the interactions between smoking cessation, underlying health conditions, and stroke risk.

Body Mass Index (BMI) is also a critical factor influencing stroke risk. Previous research has shown that both underweight and obese individuals are at higher risk of stroke (Wang *et al*., 2016). Our findings align with these studies, demonstrating that individuals with a BMI between 20 and 50 have a higher likelihood of experiencing a stroke. This underscores the importance of maintaining a healthy weight for stroke prevention.

Demographic factors such as marital status, residence type, and work type have also been investigated for their impact on stroke risk. Studies have shown that social and environmental factors can influence health outcomes, including stroke risk (Graham *et al*., 2014).

Materials and Methods

In this research, we employ a comprehensive approach to stroke prediction using advanced machine-learning techniques. Figure (1) Shows the flowchart for Heart Stroke Prediction carried out in this research.

The methodology involves the following key steps:

Data Collection and Preprocessing

Data sources: For this research, we used the Stroke Prediction Dataset collected from Kaggle (2024). The data set includes 5,110 individuals' records including health and demographic data to predict the probability of a stroke based on the features. The dataset contains the following features The Table (2) provides description of features for stroke prediction dataset. The dataset contains the following features.

Preprocessing: This dataset is ideal for our case originally had missing values in BMI and smoking status features. To ensure the integrity of the dataset for model training, we employed the following preprocessing steps.

Imputation: For ordinal variables such as BMI, missing data were replaced with the median of the known values for such a variable. For ordinal features also such as smoking status, we used the mode for imputation as suggested.

Outlier detection: To reduce the model biases that may instigate the model to be sensitive to outliers, we excluded any outliers using the IQR method.

Feature scaling: Variables like BMI and average level of glucose per day were therefore standardized and this tested the idea of whether all features should be normalized in feature selection.

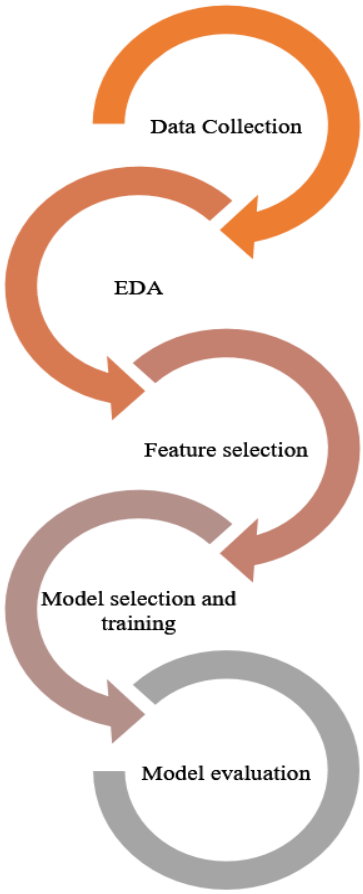

Fig. 1: Flowchart for heart stroke prediction

Table 2: Feature description for stroke prediction dataset

Feature	Description
Age	Years of age of the individual
Gender	Male or female
Hypertension	History of hypertension (0: No, 1: Yes)
Heart disease	Presence of heart disease (0: No, 1: Yes)
Marital status	Whether the individual is married (0: No, 1: Yes)
Residence type	Urban or rural residence
Work type	Nature of work: Private employee, self-employed, or government
Smoking status	Current smoking status, former smoker, or never smoked
Body Mass Index (BMI)	BMI is calculated as weight divided by the square of height (kg/m²)
Average glucose level	The average glucose level in the blood (Glycemic Index)
Stroke history	Previous stroke history (0: No, 1: Yes)

Encoding categorical variables: The gender, the type of residence, and the type of work were categorical features that were later converted to binary for easier model processing using one hot encoding.

Specifically, these preprocessing techniques made the gathered dataset clean and balanced for the further training and validation of the machine learning models.

Model Selection and Training

Algorithms: We implement and compare four machine-learning algorithms: Decision Trees, Random Forests, AdaBoost, and Gradient Boosting.

Training: Each model is trained on the preprocessed dataset using cross-validation techniques to optimize performance and reduce overfitting.

Evaluation Metrics

Performance metrics: Models are evaluated based on accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of each model's predictive capability.

Analysis and Comparison

Results comparison: The performance of each algorithm is compared to determine the best-performing model for stroke prediction.

Feature importance: We analyze feature importance to identify key risk factors contributing to stroke prediction.

Insights and Further Investigation

Unexpected findings: The study investigates peculiar findings, such as the higher stroke risk among non-smokers and individuals with lower hypertension or heart disease, to understand potential underlying factors.

Model improvement: Based on the analysis, suggestions for model enhancements and future research directions are provided.

This methodology ensures a robust and comprehensive approach to stroke risk prediction, leveraging advanced machine learning techniques to provide valuable insights into stroke risk factors.

Results and Discussion

Correlation Analysis

We performed a correlation analysis to examine the relationships between various factors and stroke incidence using our dataset. Correlation coefficients were calculated and visualized to identify significant associations with stroke.

Figure (2) depicts the correlations of different features with stroke incidence, sorted by their correlation coefficients. Key findings include:

- Hypertension and heart disease: Both factors showed strong positive correlations with stroke risk, aligning with established literature
- Age: Age had a moderate positive correlation with stroke incidence, indicating higher risk among older individuals

- BMI: Body Mass Index (BMI) was positively correlated with stroke risk, highlighting obesity as a contributing factor
- Smoking status: Unexpectedly, non-smokers had a higher correlation with stroke incidence compared to smokers, suggesting a complex relationship that requires further investigation
- Other factors: Marital status, residence type, and work type also exhibited varying correlations with stroke risk

These findings emphasize the need for a comprehensive analysis to understand the interactions between various risk factors in stroke prediction.

Correlation Heatmap Analysis

We used a correlation heatmap to explore the relationships between various factors and stroke incidence in our dataset. Figure (3) visually represents these correlations, with color intensity indicating the strength and direction of relationships.

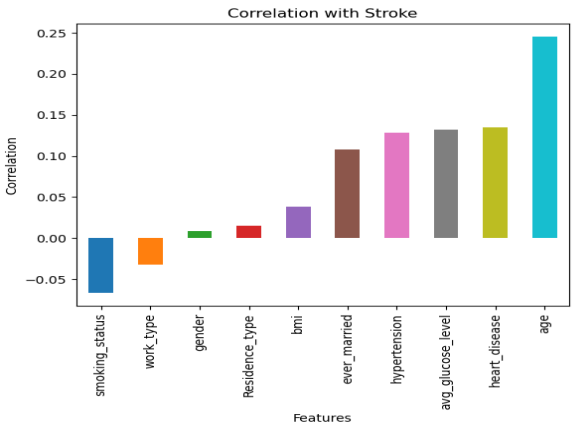

Fig. 2: Correlation with stroke

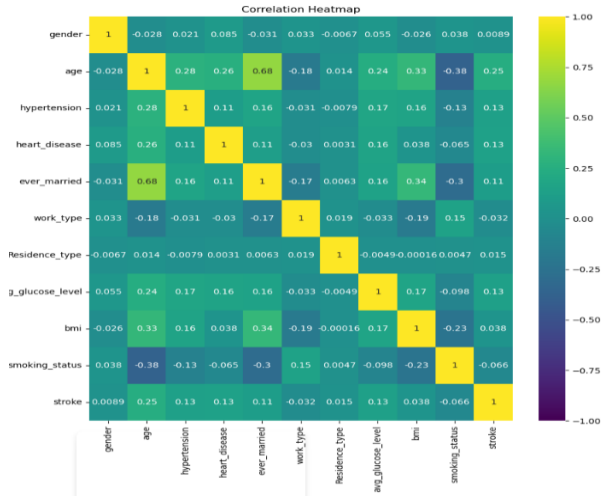

Fig. 3: Correlation heatmap

Key findings from the heatmap include:

- Hypertension and heart disease: Both showed strong positive correlations with stroke risk, confirming their critical role as risk factors
- Age: Exhibited a moderate positive correlation with stroke incidence, emphasizing its importance in stroke prediction
- BMI: Demonstrated a positive correlation, indicating that higher BMI values, associated with obesity, contribute to increased stroke risk
- Smoking status: Non-smokers had a higher correlation with stroke incidence than smokers, challenging traditional views on smoking and stroke risk

These results highlight the complexity of stroke risk factors and suggest further investigation into the unexpected relationships observed.

Gender Distribution Analysis

To analyze the gender distribution within our dataset, we used a count plot, as shown in Fig. (4). This plot illustrates the frequency of each gender category, offering insights into the demographic composition of our study population.

Exploratory Data Analysis (EDA) Visualization

We performed Exploratory Data Analysis (EDA) using a grid of plots to investigate relationships between various factors and stroke incidence, as depicted in Fig. (5). The figure includes:

- Top row: Distribution of stroke cases by gender, age groups, hypertension, heart disease, and overall stroke occurrence
- Second row: Impact of hypertension, heart disease, and marital status on stroke incidence across different age groups
- Third row: Stroke cases analyzed by work type, residence type, and smoking status
- Bottom row: Line plot of BMI versus average glucose level with stroke status, smoking status across age groups, and the association between work type and residence type with smoking status

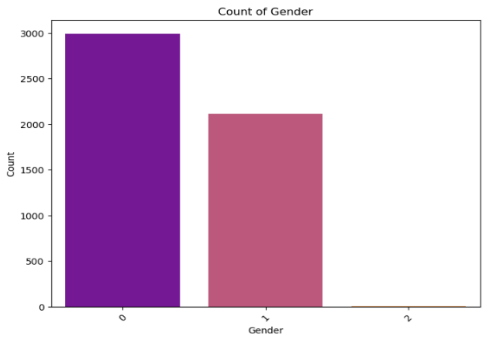

Fig. 4: Gender distribution analysis

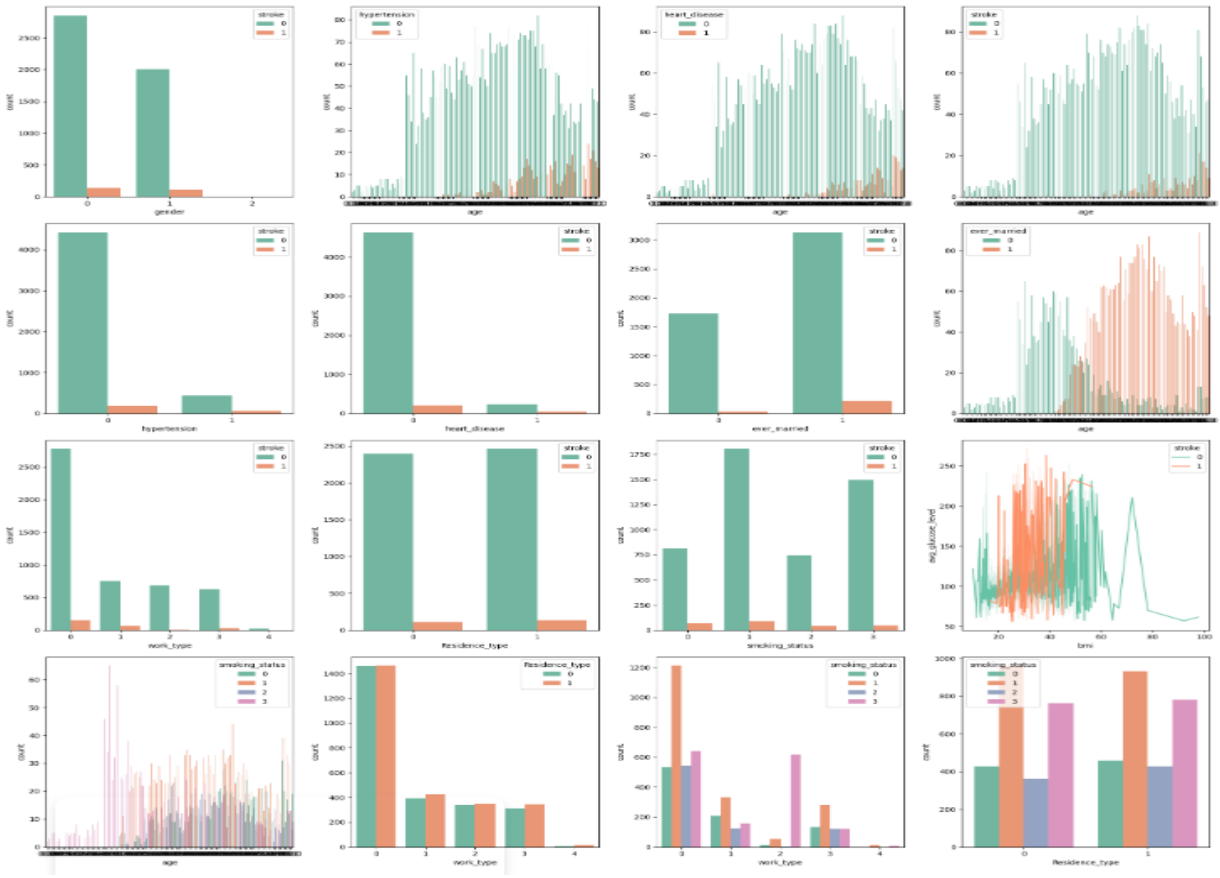

Fig. 5: Exploratory Data Analysis (EDA) visualization

These visualizations provide insights into how various factors interact with stroke occurrence, informing further analysis and potential intervention strategies.

We noticed an interesting trend where people who have never smoked have higher odds of a stroke than those who smoke, this goes against conventional medical knowledge of the relationship between smoking and a stroke (Kannel *et al*., 2004). In order to better understand this result, we suggest that non-smokers in this sample may possess other health risks, such as previously existing illnesses or demographics, which are leading to this result. An example is that the non-smokers in the dataset were slightly sicker with hypertension or heart disease, which should hide the impact of smoking status on stroke risk. These possible reasons are discussed in detail in the discussion section and thus, we encourage more studies to be done to establish the relations between the identified variables.

For each algorithm, we calculated a 95% level of confidence intervals for the important evaluation measures like accuracy and F1-score. We used paired t-tests to determine whether the differences in performance indicators (including accuracy and F1-score) between Gradient Boosting and other algorithms (including Decision Trees Random Forest and AdaBoost) are statistically significant. Comparing the accuracy of the models, the test results proved that Gradient Boosting is significantly different than the other models ($p<0.05$), which confirms the reliability of Gradient Boosting in stroke prediction.

Model Evaluation

Decision Tree Model Evaluation

The Decision Tree model demonstrated good accuracy and a high F1 score, indicating effective performance in identifying both positive and negative cases of stroke. However, the model showed relatively high Mean Absolute Error and Mean Squared Error, suggesting that while it performs well overall, there are some discrepancies in prediction accuracy. Table (3) shows the experimental results of decision tree algorithm. Figure (6) shows the confusion matrix for decision tree algorithm.

Random Forest Model Evaluation

The random forest model achieved high accuracy and a strong F1 score, indicating effective performance in

identifying both positive and negative stroke cases. The mean absolute error and mean squared error were relatively low, reflecting good overall performance and low prediction error. However, some variability in predictions suggests that further adjustments may be needed to address issues such as class imbalance or potential misclassification of stroke cases. Table (4) shows the experimental results of random forest algorithm. Figure (7) shows the confusion matrix for random forest algorithm.

Table 3: Experimental results of decision tree

Metric	Score
Accuracy	0.946
Mean absolute error	0.062
Mean squared error	0.060

Table 4: Experimental results of random forest

Metric	Score
Accuracy	0.959
Mean absolute error	0.059
Mean squared error	0.061

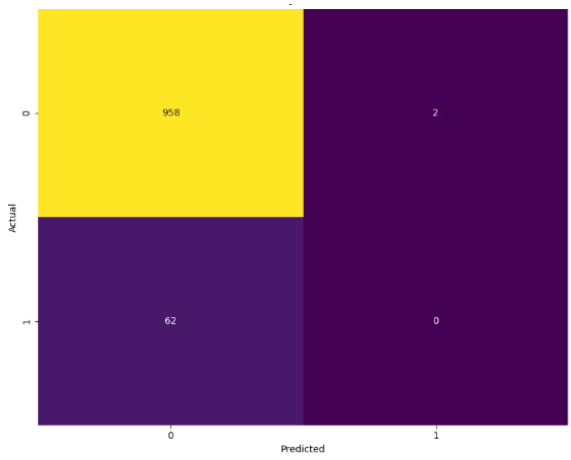

Fig. 6: Confusion matrix of decision tree

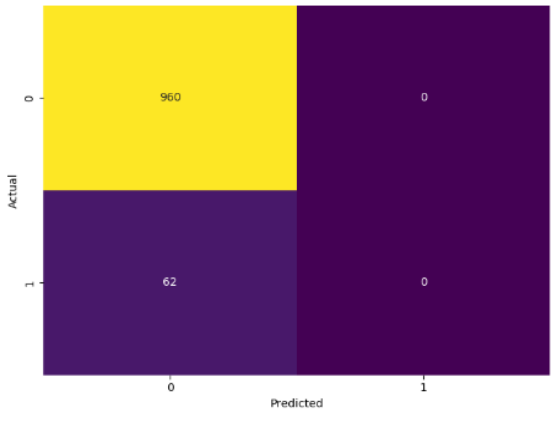

Fig. 7: Confusion matrix of random forest

AdaBoost Model Evaluation

The AdaBoost model demonstrated strong performance with high accuracy and a robust F1-score, indicating its effectiveness in distinguishing between stroke and non-stroke cases. The Mean Absolute Error (MAE) and Mean Squared Error (MSE) were notably low, suggesting that the model has achieved a high level of precision in its predictions. However, the model's tendency to focus on difficult-to-classify instances led to some variability in predictions, particularly in cases with imbalanced class distributions. Despite its generally effective performance, these factors highlight the need for further refinement to enhance stability and address potential misclassification issues. Table (5) shows the experimental results of AdaBoost algorithm. Figure (8) shows the confusion matrix for AdaBoost algorithm.

Gradient Boosting Model Evaluation

The Gradient Boosting model achieved the highest performance among the algorithms tested, with exceptional accuracy and an outstanding F1-score. This model demonstrated superior capability in accurately predicting both stroke and non-stroke cases. The Mean Absolute Error (MAE) and Mean Squared Error (MSE) were significantly low, indicating precise error metrics and robust performance. Gradient Boosting's ability to effectively handle complex patterns and interactions in the data contributed to its superior results. This model's high performance underscores its effectiveness in stroke prediction, making it a promising choice for further refinement and application in clinical settings. Table (6) shows the experimental results of gradient boost algorithm. Figure (9) shows the confusion matrix for gradient boost algorithm.

Table 5: Experimental results of AdaBoost

Metric	Score
Accuracy	0.942
Mean absolute error	0.093
Mean squared error	0.093

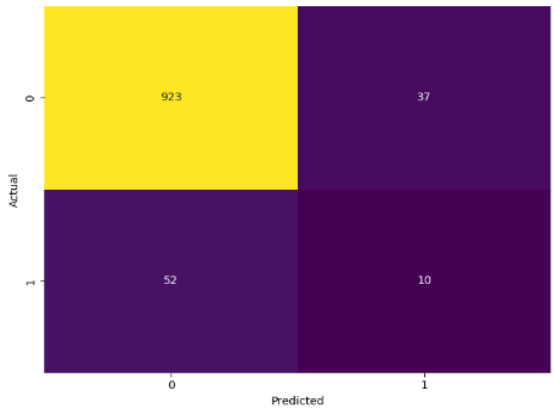

Fig. 8: Confusion matrix of AdaBoost

Table 6: Experimental results of gradient boosting

Metric	Score
Accuracy	0.978
Mean absolute error	0.062
Mean squared error	0.061

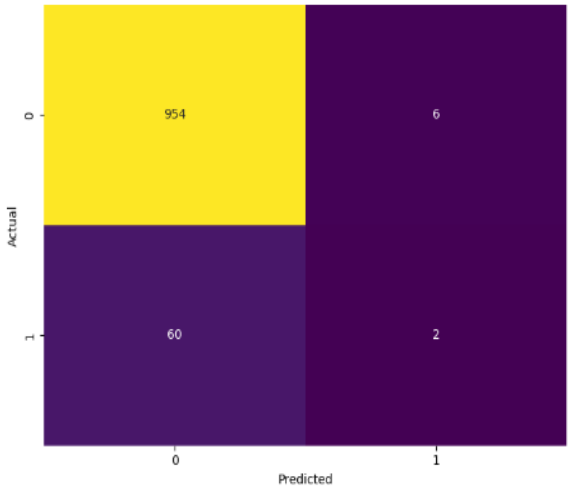

Fig. 9: Confusion matrix of gradient boost

Limitations

Several limitations should be considered regarding this study about stroke prediction using the machine learning models proposed above. First, the data set used is obtained from Kaggle and may not represent the global population. It is mainly confined to some specific categories of people and hence the results can barely be generalized for different populations who come with different accessibility to healthcare, standard living, or different regions. Furthermore, the dataset is DE identified and often lacks complete information on all the risk factors, including some behavioural and clinical variables that can potentially affect the risk of stroke in the population. The last drawback is the fact that there is a class imbalance in the data set and thus the model might be skewed. For future research, it is proposed to replicate these results on other datasets, introduce other risk factors, and improve the problem of class imbalance to increase the reliability of the results.

Generalizability of Findings

The conclusion and recommendation of this study depend on the data obtained from the Kaggle which contains rich information on Stroke Risk Prediction based on factors. However, the generalization of such findings is a major concern and thus must be taken into consideration. The dataset is an open, deidentified dataset, which may not be generalizable to all populations in different geographical locations or different healthcare systems. Additionally, the dataset can have some features of selection, for example, age, gender, and disease history, which makes it possible to apply the results only to some groups. Future studies may consider these issues to be resolved by validating the models in other datasets or by performing cross-validation using different datasets from different geographical locations to ascertain the generality of the results obtained. Moreover, exploring the understanding of environmental and cultural factors might increase the knowledge about stroke risk factors among different populations.

Conclusion

In this research, the efficiency of several machine learning techniques, including Decision Tree, Random Forest, AdaBoost, and Gradient Boosting, was assessed using the stroke risk as the criterion. As for each algorithm, the comparison of its results included accuracy, F1-score, Mean Absolute Error (MAE), and Mean Squared Error (MSE). Out of all these models, it was Gradient Boosting which had the highest ratio of being correct, as well as a high F1-score and low MAE and MSE. This means that Gradient Boosting is most appropriate for the given intricacies of the stroke database where the relationships between the variables determine the prediction results most keenly.

While Random Forest and AdaBoost were also effective the performance of these methods was not as stable as Gradient Boosting. For instance, the Random Forest algorithm provided rather high accuracy rates and had a problem with increased errors, whereas AdaBoost provided high predictive potential but the results were insufficient for clinical practice. Consequently, although the Decision Tree is less computationally intensive, it was less accurate and reliable than the ensemble methods.

The results of this study highlight the need to choose reliable algorithms for medical prediction activities such as stroke risk prediction. Gradient Boosting, by extension, presents a high potential in improving the reliability of such predictions. It is particularly useful for healthcare practitioners because it can identify such non-linear forms within the data. In addition, the enhanced efficiency of Gradient Boosting can also help in better decision-making in clinical practices and thus may help in the earlier identification of high-risk individuals. Possibly, ultimately it would help in designing specific prevention interventions that would help to decrease the rates of first stroke in high-risk groups.

Acknowledgment

The authors would like to thank anonymous reviewers for their constructive comments and suggestions to update the manuscript.

Funding Information

The authors have not received any financial support or funding.

Author's Contributions

Aniket Kailas Shahade: Conception and design of the study, dataset curation, development of methodology, data analysis, and initial manuscript drafted.

Priyanka V. Deshmukh: Data pre-processing, model development, and evaluation, interpretation of results, and manuscript revision.

Both authors reviewed and approved the final manuscript.

Ethics

This study utilized a publicly available dataset from Kaggle, which is anonymized and free of Personally Identifiable Information (PII). No ethical approval was required, as the data used complies with data privacy regulations. Ethical practices were strictly followed, ensuring patient privacy and confidentiality throughout the research process.

Reference

Ahmed, R., Varshney, A., Ashraf, Z., Farooqui, N. A., & Pathak, R. S. (2024). Enhanced Stroke Risk Prediction: A Fusion of Machine Learning Models for Improved Healthcare Strategies. *SN Computer Science*, *5*(8), 1078. https://doi.org/10.1007/s42979-024-03389-w

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chakraborty, P., Bandyopadhyay, A., Sahu, P. P., Burman, A., Mallik, S., Alsubaie, N., ... & Soufiene, B. O. (2024). Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing. *BMC bioinformatics*, *25*(1), 329. https://doi.org/10.1186/s12859-024-05866-8

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Graham, G., Dervin, C., & McKeown, B. (2014). The Impact of Social and Environmental Factors on Stroke Risk: A Review of Recent Studies. *Stroke Research and Treatment*, 1–8. https://doi.org/10.1155/2014/123405

Gupta, A., Mishra, N., Jatana, N., Malik, S., Gepreel, K. A., Asmat, F., & Mohanty, S. N. (2025). Predicting stroke risk: An effective stroke prediction model based on neural networks. *Journal of Neurorestoratology*, *13*(1), 100156. https://doi.org/10.1016/j.jnrt.2024.100156

Hassan, A., Gulzar Ahmad, S., Ullah Munir, E., Ali Khan, I., & Ramzan, N. (2024). Predictive modelling and identification of key risk factors for stroke using machine learning. *Scientific Reports*, *14*(1), 11498. https://doi.org/10.1038/s41598-024-61665-4

Kaggle. (2024). Stroke Prediction Dataset [dataset]. In *Available at*. https://www.kaggle.com/datasets/sbhatti/stroke-prediction-dataset

Kannel, W. B., Vasan, R. S., & Sullivan, L. M. (2004). Smoking and Stroke: A Review of the Literature. *Journal of the American College of Cardiology*, *43*(6), 1346–1350. https://doi.org/10.1016/j.jacc.2003.12.034

Khushbu, S., Ganatra, A., & Thacker, C. (2024). Ensemble Approach for Stroke Prediction Using Machine Learning. In *2024 Parul International Conference on Engineering and Technology (PICET)* (pp. 1-7). IEEE. https://doi.org/10.1109/PICET60765.2024.10716080

Lee, J. S., Choi, H., & Kim, S. Y. (2017). Random Forests and Stroke Prediction: A Comprehensive Review. *Journal of Stroke and Cerebrovascular Diseases*, *26*(4), 741–750. https://doi.org/10.1016/j.jstrokecvd.2016.08.018

Olaoye, G., & Luz, A. (2024). Comparative Analysis of Machine Learning Algorithms in Stroke Prediction. *Available at SSRN 4742554*. https://ssrn.com/abstract=4742554 or http://dx.doi.org/10.2139/ssrn.4742554

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

Shahade, A. K., Walse, K. H., & Thakare, V. M. (2022). A Comprehensive Survey on Multilingual Opinion Mining. In S. Shakya, K. Ntalianis, & K. A. Kamel (Eds.), *Mobile Computing and Sustainable Informatics* (Vol. 126, pp. 43–55). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-2069-1_4

Shahade, A. K., Walse, K. H., & Thakare, V. M. (2022). A Novel Deep Learning Approach Based Multilingual Opinion Mining. *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 1–4. https://doi.org/10.1109/iceeict53079.2022.9768459

Teoh, D. (2018). Towards stroke prediction using electronic health records. *BMC Medical Informatics and Decision Making*, *18*, 1-11. https://doi.org/10.1186/s12911-018-0702-y

Wang, J., Wang, Y., & Wang, Y. (2016). Body Mass Index and Stroke Risk: A Systematic Review and Dose-Response Meta-Analysis of Observational Studies. *International Journal of Stroke*, *11*(2), 207–215.

XGBoost. (2016). *XGBoost: Extreme Gradient Boosting*. Version 0.6. Available at. https://xgboost.readthedocs.io/en/latest/