ARTICLE
Medical Data Mining 2024;7(3):15. https://doi.org/10.53388/MDM202407015

A machine learning approach for predictings stroke

Yubo Fu[1*]

[1]Hasan School of Business, Colorado State University Pueblo, Pueblo, CO 81001, USA.

[*]**Corresponding to:** Yubo Fu, Hasan School of Business, Colorado State University Pueblo, 2200 Bonforte Blvd, Pueblo, CO 81001, USA. E-mail: yubo.fu@csupueblo.edu.

Author contributions
Yubo Fu analyzed data and wrote this manuscript.
Competing interests
The authors declare no conflicts of interest.
Acknowledgments
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
Peer review information
Medical Data Mining thanks all anonymous reviewers for their contribution to the peer review of this paper.
Abbreviations
AUC, area under the curve; SVM, support vector machine; KNN, K-Nearest Neighbors; BMI, body mass index; SMOTE, Synthetic Minority Over-sampling Technique; HPML, high-performance machine learning.
Citation
Fu Y. A machine learning approach for predictings stroke. *Med Data Min*. 2024;7(3):15. doi: 10.53388/MDM202407015.

Executive editor: Xin-Yun Zhang.
Received: 14 December 2023; **Accepted:** 07 March 2024; **Available online:** 18 March 2024.

Abstract
Background: Stroke is one of the most dangerous and life-threatening disease as it can cause lasting brain damage, long-term disability, or even death. The early detection of warning signs of a stroke can help save the life of a patient. In this paper, we adopted machine learning approaches to predict strokes and identify the three most important factors that are associated with strokes. **Methods:** This study used an open-access stroke prediction dataset. We developed 11 machine learning models and compare the results to those found in prior studies. **Results:** The accuracy, recall and area under the curve for the random forest model in our study is significantly higher than those of other studies. Machine learning models, particularly the random forest algorithm, can accurately predict the risk of stroke and support medical decision making. **Conclusion:** Our findings can be applied to design clinical prediction systems at the point of care.

Keywords: medical decision making; machine learning; predictive modeling; stroke; imbalanced data

Introduction

Stroke, a devastating cardiovascular disease, poses a significant global health burden as the second leading cause of death worldwide [1]. In the United States, it ranks as the fifth leading cause of death among individuals aged 70 years and older [2]. The occurrence of stroke can be attributed to two main mechanisms: either a blockage in the blood supply to the brain or the rupture of a blood vessel within the brain [3]. Consequently, certain regions of the brain may sustain permanent damage, resulting in the loss of specific functions or even mortality. Alarming statistics from the Centers for Disease Control and Prevention [4, 5] reveal that over 795,000 people experience a stroke each year in the United States alone, with a staggering stroke-related fatality occurring every 3.5 minutes. This imposing public health challenge also carries substantial economic implications, exerting a considerable financial burden on individuals, healthcare systems, and society at large [6]. In fact, stroke-related costs account for 1.7% of national health expenditures [7].

However, the Centers for Disease Control and Prevention [8] emphasizes that up to 80% of strokes can be preventable. While blood pressure levels serve as a crucial factor influencing the likelihood of experiencing a stroke, other environmental and lifestyle factors also contribute significantly to this risk. For instance, studies have established a clear association between a diet high in saturated fats, trans fats, and cholesterol and an increased susceptibility to stroke. Furthermore, research has revealed a correlation between the level of stress an individual experiences and their chances of suffering a stroke [9]. Disturbingly, projections indicate that approximately 4% of the adult population in the United States is expected to have experienced a stroke by the year 2030 [7].

In the realm of healthcare, the growing utilization of machine learning, deep learning, and artificial intelligence techniques has revolutionized the accurate diagnosis, treatment, and management of various conditions, including stroke. This paper focuses on leveraging machine learning approaches to predict patient outcomes, enhance stroke management, and improve the overall quality of patient care [10–12]. Specifically, our objective is to investigate the robust predictors for stroke through the application of machine learning approaches and analyze their impact on an individual's risk of developing severe medical conditions.

This paper is organized into five sections. Section 2 provides an overview of related work. In Section 3, we outline the methodology and procedures employed in our study. Subsequently, Section 4 presents the results and discussions derived from our analyses. Finally, the conclusions and future are discussed in the last section of this paper.

Related work

A considerable body of research has employed machine learning techniques to explore and predict strokes. Heo et al. [13] conducted a study aiming to predict long-term outcomes in ischemic stroke patients. They developed three machine learning models, namely random forest, logistic regression, and deep neural network. Notably, their deep neural network model achieved an impressive area under the curve (AUC) of 0.888. Their investigation demonstrated that machine learning algorithms, particularly the deep learning neural network, exhibit the capability to accurately predict long-term outcomes in acute stroke patients. However, their study limited the exploration to only three machine learning methods. Moreover, Wu and Fang [14] developed machine learning models for predicting stroke with imbalanced data in an elderly population in China. Regularized logistic regression, support vector machine (SVM), and random forest methods were used to predict stroke. The AUC for regularized logistic regression reached the maximum value 0.72, and those for SVM and random forest both reached 0.71 using their model. Their study demonstrated that machine learning methods with data balancing techniques were effective tools for stroke prediction with imbalanced dataset. However, their model performances were relatively low even after dealing with imbalanced data. Wang et al. [15] conducted an extensive review of studies published on PubMed and Web of Science from 1990 to March 2019. Their review revealed that the most commonly utilized machine learning methods for stroke prediction were random forest, support vector machines, decision trees, and neural networks. Hence, there is a need to incorporate a wider range of machine learning approaches to assess and compare their performances effectively.

In line with this objective, Khosla et al. [16] proposed a novel machine learning approach that combines Margin-based Censored Regression with Cox models to automatically select robust features capable of predicting stroke risks. Their study applied support vector machines and Margin-based Censored Regression. The resulting technique achieved an AUC of 0.777. Similarly, Chun et al. [17] compared Cox models, machine learning and ensemble models combing both approaches for stroke risk prediction. They concluded that the ensemble approach led higher accuracy (men: 76%, women: 80%). In addition, Chen et al. [18] utilized deep learning method, namely deep neural network and compared it with machine learning-based classifiers, such as logistic regression and support vector machine, to predict patients' 5-year stroke occurrence. Based on their findings, deep neural network and gradient boosting decision tree can achieve similar good predictive results (AUCs of 0.915) while Logistic Regression and SVM both show a less effective performance. Sailasya and Kumari [19] conducted research to construct machine learning models for predicting the likelihood of brain strokes. In their study, they employed various machine learning algorithms, including logistic regression, decision tree, random forest, K-Nearest Neighbors (KNN), and Naïve Bayes. The Naïve Bayes Classifier exhibited the best performance, yielding an accuracy of 82% and an AUC of 82.3. Although their model outperformed the current state-of-the-art in terms of AUC, there is still room for improvement to achieve even better results.

Previous studies [16, 20–23] have identified some significant stroke predictors. These predictors include factors like time to walk 15 feet, creatinine levels, age, the number of correctly coded symbols, total medications, general health, as well as excellent heart and kidney function. Several other factors have also been identified as potential risk factors for stroke. These include the occurrence of a transient stroke, the presence of myocardial infarction, atrial fibrillation, smoking, obesity, alcohol consumption, and estrogen therapy [24–26]. Nonetheless, numerous other risk factors for stroke exist, necessitating further exploration in future studies utilizing machine learning approaches.

Hence, the research question for this study is twofold: can we identify factors that accurately predict the risk of stroke? Moreover, can we develop prediction models that exhibit high accuracy and reliability? By addressing these questions, this research aims to contribute to the advancement of stroke prediction and ultimately improve patient outcomes.

Methodology

Date preprocessing

The dataset utilized in this study was the open-access Stroke Prediction dataset [27], comprising approximately 5110 patients, the age of the participants who are over 18 years old. Each patient's record included multiple attributes, as illustrated in Table 1 along with an indication of whether they had experienced a stroke. As this dataset is publicly available, neither approval from the ethics committee nor informed consent from the study populations is required.

Prior to training the dataset with classifier models, several crucial data cleaning steps were undertaken. First, it was identified that 201 instances had missing values for the body mass index (BMI) attribute. Therefore, these specific rows were removed from the dataset, as BMI serves as a fundamental measure of an individual's overall health.

Moreover, the dataset exhibited a substantial class imbalance, with

95% of individuals not having experienced a stroke, while only 5% had. To address this imbalance and create a more equitable target variable, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE generates artificial samples for the minority class by selecting instances that are proximate in the feature space, using a concept similar to the distance notion often employed in KNN algorithms. By equalizing the distribution of data, SMOTE mitigates the potential bias of the model toward a particular outcome. The results of applying SMOTE are presented in Table 2, which highlights the impact of this technique on the dataset and the achieved class balance.

Following the data cleaning and preprocessing steps, the dataset was partitioned into training and testing data sets. 75% of the data were allocated for training purposes, while the remaining 25% were reserved for testing the trained models. This partitioning facilitated the evaluation of the models' performance on unseen data, providing a reliable assessment of their predictive capabilities.

Next, the training data set was utilized to train 11 different machine learning algorithms, including logistic regression, support vector machines, decision trees, random forests, and neural networks, etc. By employing multiple algorithms, we aimed to explore their individual strengths and weaknesses in predicting stroke risks. The framework diagram in Figure 1 provides an overview of the sequential steps involved in training and testing the machine learning algorithms on the stroke prediction dataset. It visualizes the flow of data, highlighting the key stages of data partitioning, algorithm selection, training, and evaluation. The framework serves as a useful reference for understanding the experimental setup and facilitates the reproducibility of the study.

Table 1 Features given for stroke prediction and their potential values in the dataset

Feature	Potential values
Gender	Male, female
Age	Years
Hypertension	Yes/No
Heart disease	Yes/No
Marital status	Yes/No
Work type	Private, self-employed, children, government job
Residence type	Urban, rural
Average glucose level	mg/dl
BMI	kg/m^2
Smoking status	Never smoked, formerly smoked, smokes, unknown

BMI, body mass index.

Table 2 Comparison of target variables before and after oversampling

Target variable	Before SMOTE	After SMOTE
Stroke patients	249	4861
Non-stroke patients	4861	4861

SMOTE, Synthetic Minority Over-sampling Technique.

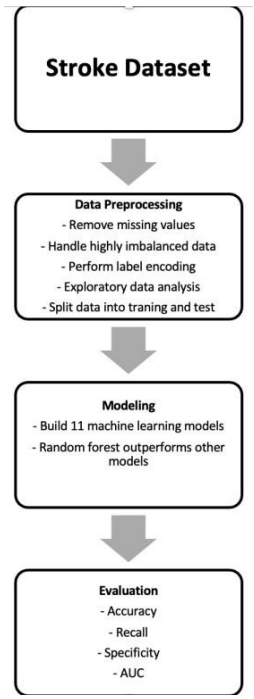

Figure 1 Stroke prediction framework. AUC, area under the curve.

Medical Data Mining 2024;7(3):15. https://doi.org/10.53388/MDM202407015

Exploratory data analysis

Several intriguing insights emerged from the analysis of the dataset, and various visualizations of the feature variables are presented below. Among these visualizations, we will highlight the most noteworthy observations in this section.

Figure 2 demonstrates a distinct pattern regarding the age of stroke patients. It is evident that a majority of stroke cases occur in individuals who are older than 50 years. This finding suggests that age plays a significant role in stroke risk, with advanced age being a notable contributing factor.

Figure 3 reveals an interesting observation: approximately 89% of stroke patients in the dataset were married. However, this association is likely not directly related to the effects of marital status on stroke risk. Instead, it can be attributed to the fact that a significant proportion of married individuals in the dataset were older, thereby aligning with the previous observation regarding age as a risk factor for stroke.

Examining comorbidities, Figure 4 indicates that around 16% of stroke patients had pre-existing heart disease. This finding emphasizes the interplay between heart health and stroke risk. Furthermore, Figure 5 highlights that approximately 24% of stroke patients had hypertension, further underscoring the link between high blood pressure and stroke incidence.

Regarding BMI, Figure 6 demonstrates a highly skewed distribution among stroke patients. The majority of patients fell within the BMI range of 25 to 35, which corresponds to the overweight and obese categories. Notably, it is interesting to observe that several non-stroke patients also fell within the overweight BMI range. However, it is important to acknowledge that BMI has faced criticism within the medical community due to its failure to account for variations in muscle mass, which can potentially lead to skewed results.

These visualizations provide valuable insights into the dataset, shedding light on notable associations between stroke incidence and factors such as age, marital status, comorbidities like heart disease and hypertension, as well as the distribution of BMI among stroke patients.

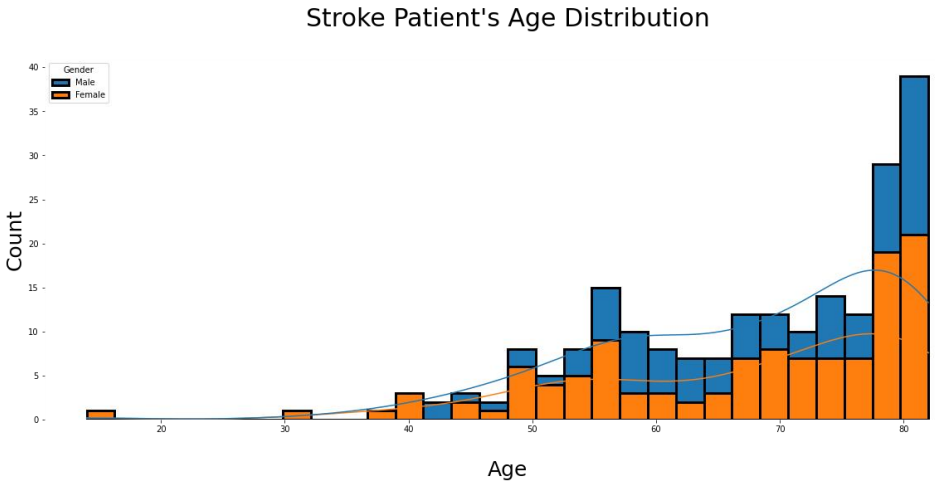

Figure 2 Age distribution of stroke patients

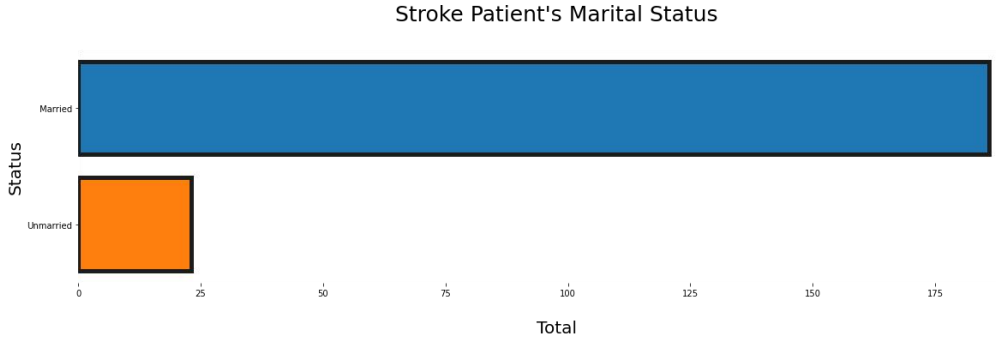

Figure 3 Marital status of stroke patients

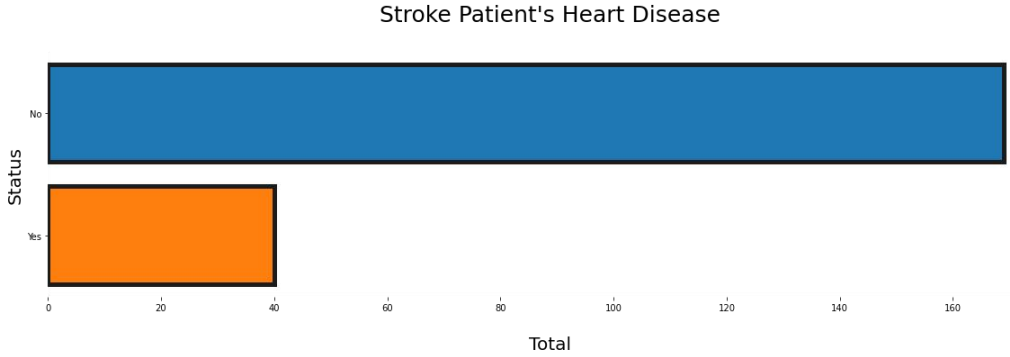

Figure 4 Stroke patients with heart disease

Medical Data Mining 2024;7(3):15. https://doi.org/10.53388/MDM202407015

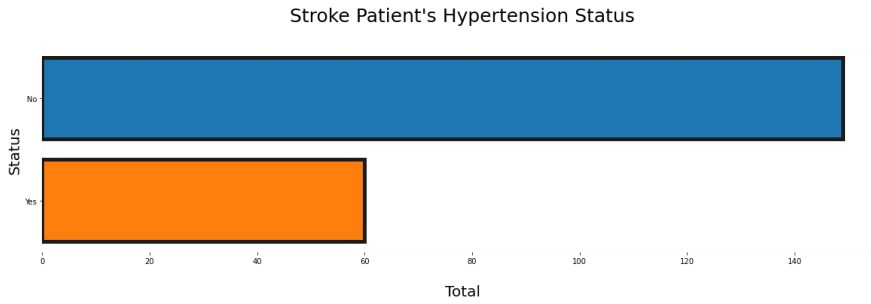

Figure 5 Stroke patients with hypertension

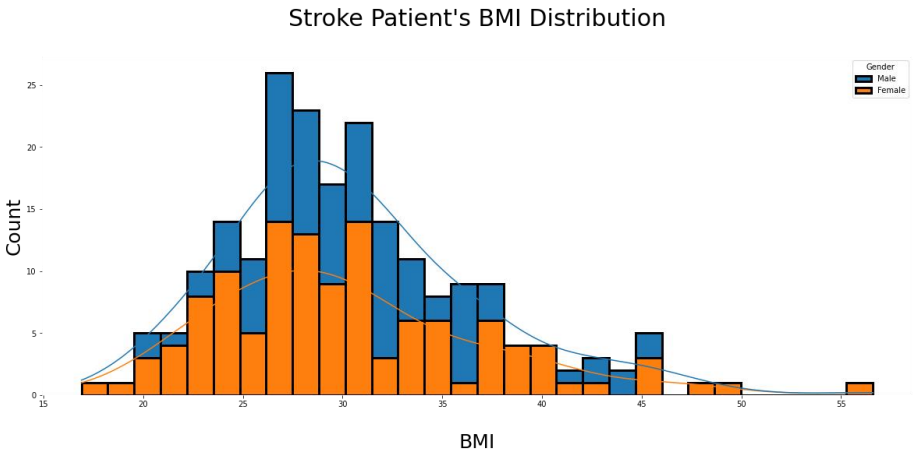

Figure 6 BMI distribution of stroke patients. BMI, body mass index.

The presence of diabetes among stroke patients was also observed in our dataset. A considerable number of stroke patients were found to have diabetes, and some exhibited high glucose values exceeding 200 mg/dl as shown in Figure 7. This observation aligns with existing literature [28, 29], which has consistently shown that comorbidities can have a significant impact on the development of heart disease and stroke.

Considering the complex interplay between diabetes, heart disease, and stroke, it becomes evident that comorbidities can significantly influence an individual's susceptibility to these conditions. The presence of diabetes among stroke patients in our dataset underscores the importance of addressing and managing comorbidities to mitigate the risk of stroke and improve overall cardiovascular health.

Modeling

In this study, we aimed to explore and evaluate the performance of various machine learning models in predicting stroke risks. Specifically, we trained a comprehensive set of 11 machine learning algorithms, each known for its effectiveness in handling classification tasks. The models utilized in this study include logistic regression, KNN, support vector machine, decision tree, random forest, adaptive boosting, gradient boosting, linear discriminant analysis, quadratic discriminant analysis, eXtreme Gradient Boosting, and categorical boosting.

While previous research in this domain often focused on a limited selection of machine learning techniques, our study sought to address this limitation by incorporating a broader range of high-performance machine learning (HPML) methods. By leveraging a diverse set of algorithms, we aimed to identify the most suitable models for stroke risk prediction. By examining the performance of these various models on our dataset, we sought to contribute to the existing literature by expanding the repertoire of machine learning techniques employed in stroke prediction research. Our research endeavors to bridge the gap by investigating and adopting additional HPML models, such as adaptive boosting, support vector machine, and eXtreme Gradient Boosting, which have shown promising results in other classification tasks.

Through this comprehensive analysis of machine learning algorithms, we aimed to identify the models that exhibit the highest predictive accuracy and reliability in predicting stroke risks. This research approach allows us to shed light on the potential benefits of utilizing a wider range of machine learning techniques in stroke prediction, enhancing the overall effectiveness of stroke risk assessment and facilitating informed decision-making in clinical settings.

Results and discussions

To analyze the factors contributing the most to stroke, we have printed out the correlation values for each feature variable, along with their respective relationship to stroke. The correlation values are presented in Table 3 below.

The dataset analysis revealed interesting insights regarding the factors associated with stroke risk. It was observed that individuals living in urban areas exhibited a slight negative correlation with the risk of stroke. At first glance, this finding may appear counter-intuitive considering the fast-paced nature and potential stress associated with urban living, particularly in corporate settings. However, further research indicates that differences in obesity rates may account for this observation. Rural areas tend to have a higher prevalence of obesity, with approximately 34.2% of adults classified as obese, compared to 28.7% in urban areas. This disparity in obesity rates could contribute to the lower stroke risk among urban residents.

Additionally, a positive correlation was observed between marital status and the likelihood of experiencing a stroke. This association can be primarily attributed to the fact that married individuals tend to be older on average. However, it is worth noting that a substantial proportion of marriages end in divorce, implying that many long-term marriages may involve individuals who experience chronic unhappiness and higher stress levels, consequently increasing their risk of stroke. Age was identified as a strong predictor of stroke risk, further emphasizing its significance in determining an individual's susceptibility to stroke.

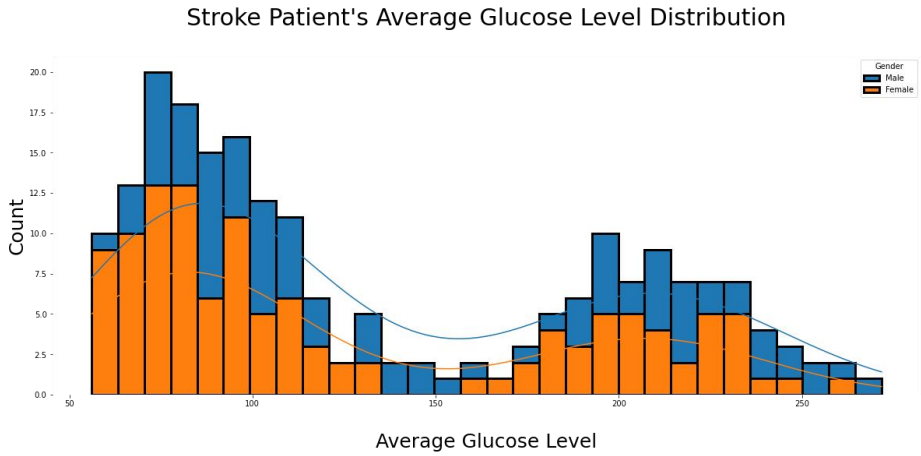

Figure 7 Glucose level distribution of stroke patients

Table 3 Correlation values (r) for each feature with respect to stroke

Independent predictor	Stroke correlation coefficient (r)
Age	0.59
Average glucose level	0.31
BMI	0.12
Male gender	–0.24
Married	0.17
Private career	–0.22
Self employed	–0.11
Children career	–0.27
Urban residence	–0.23

BMI, body mass index.

Having explored the factors influencing stroke risk, the subsequent focus was on assessing the predictive power of the machine learning models employed in this study. The model performance metrics, including accuracy, precision, recall, and F1-score, were used to evaluate their effectiveness. The results, presented earlier, demonstrated that the models consistently exhibited robust predictive capabilities, accurately identifying individuals at risk of stroke.

These findings contribute to our understanding of the factors influencing stroke risk and highlight the potential of machine learning models in predicting stroke occurrence. The observed negative correlation between urban living and stroke risk, when considering the influence of obesity rates, challenges initial assumptions and emphasizes the importance of considering various factors holistically. Similarly, the positive correlation between marital status and stroke risk underscores the complex interplay between age, marital satisfaction, and stress in determining susceptibility.

We next show our model's predictive power. By showcasing the predictive power of the models, this study establishes their potential applicability in clinical settings. These models can aid healthcare professionals in accurately identifying individuals at risk of stroke, facilitating timely preventive interventions and improving patient outcomes.

As shown in Table 4, all of our machine learning models exhibited excellent performance in predicting stroke risk. Accuracy, which measures the frequency of correct predictions by the machine learning models regarding stroke occurrence, resulted in high values. Recall assesses the proportion of correctly predicted records out of the total number of records, while specificity indicates the correct prediction of negative instances (i.e., no stroke) among all negative records. Both recall and specificity demonstrated high values in our models. It is worth noting that initially, the models were trained on an unbalanced dataset. However, after employing the SMOTE technique and using a balanced dataset, we observed improved accuracy across all models. The AUC score is a widely used metric to evaluate classifier performance, with higher scores indicating better performance. In our study, the Random Forest Classifier achieved the highest AUC score of 0.99, indicating its superior predictive ability.

In real-world applications, where preventative treatment decisions are crucial, selecting a model with high recall would be desirable. Recall measures a model's capability to identify all relevant cases within a dataset. Among our models, the KNN classifier exhibited the highest recall. Therefore, if we were to prioritize identifying all individuals at risk of stroke, the KNN classifier would be a suitable choice.

Table 5 presents a comparison of the results obtained from our random forest model with those reported in recent studies by Sailasya (2021) and Dev et al. (2022). Our model achieved an accuracy of 96.56%, a recall of 94.74%, a specificity of 98.36%, and an impressive AUC of 0.99. In contrast, the reference study by Sailasya achieved lower performance metrics, with an accuracy of 73%, a recall of 73.5%, a specificity of 72%, and an AUC of 0.73. In Dev et al.'s study, the model's performance was also lower with respect to accuracy (75%), recall (74%), specificity (76%), and AUC (0.75). These results clearly demonstrate the superior performance of our model compared to previous studies in accurately predicting stroke risk.

Furthermore, we generated a relative graph to visualize the importance of each feature in predicting stroke, as presented below in Figure 8. Notably, age, average glucose level, and BMI emerged as the top three most influential factors. Age, being a well-established risk factor for stroke, unsurprisingly demonstrated a high level of importance in our model. The average glucose level, reflecting the individual's blood sugar level, also exhibited strong predictive power. BMI, a measure of body mass index, was another significant predictor, suggesting the association between obesity and increased stroke risk.

Medical Data Mining 2024;7(3):15. https://doi.org/10.53388/MDM202407015

Table 4 Accuracy, recall, specificity, and AUC scores of each machine learning model

Classifier	Accuracy	Recall	Specificity	AUC
LogisticRegression	94.4987147	0.90309278	0.98666667	0.97818239
KNeighborsClassifier	96.0925450	0.95773196	0.96410256	0.97782818
SVC	92.4935733	0.85567010	0.99384615	0.97810785
DecisionTreeClassifier	94.0359897	0.94020619	0.94051282	0.94035950
RandomForestClassifier	96.5552699	0.94742268	0.98358974	0.99323817
AdaBoostClassifier	93.2133676	0.91030928	0.95384615	0.97929791
GradientBoostingClassifier	95.0128535	0.92680412	0.97333333	0.98638858
LinearDiscriminantAnalysis	93.5218509	0.87216495	0.99794872	0.97772773
QuadraticDiscriminantAnalysis	88.8431877	0.83195876	0.94461538	0.88828707
XGBClassifier	94.7043702	0.92268041	0.97128205	0.98614750
CatBoostClassifier	96.4524422	0.95051546	0.97846154	0.99212583

AUC, area under the curve; SVC, support vector machines.

Table 5 Results comparison

Ours (random forest)	Sailasya's (random forest)	Dev et al.'s (random forest)
Accuracy: 96.56%	Accuracy: 73%	Accuracy: 75%
Recall: 94.74%	Recall: 73.5%	Recall: 74%
Specificity: 98.36%	Specificity: 72%	Specificity: 76%
AUC: 0.99	AUC: 0.73	AUC: 0.75

AUC, area under the curve.

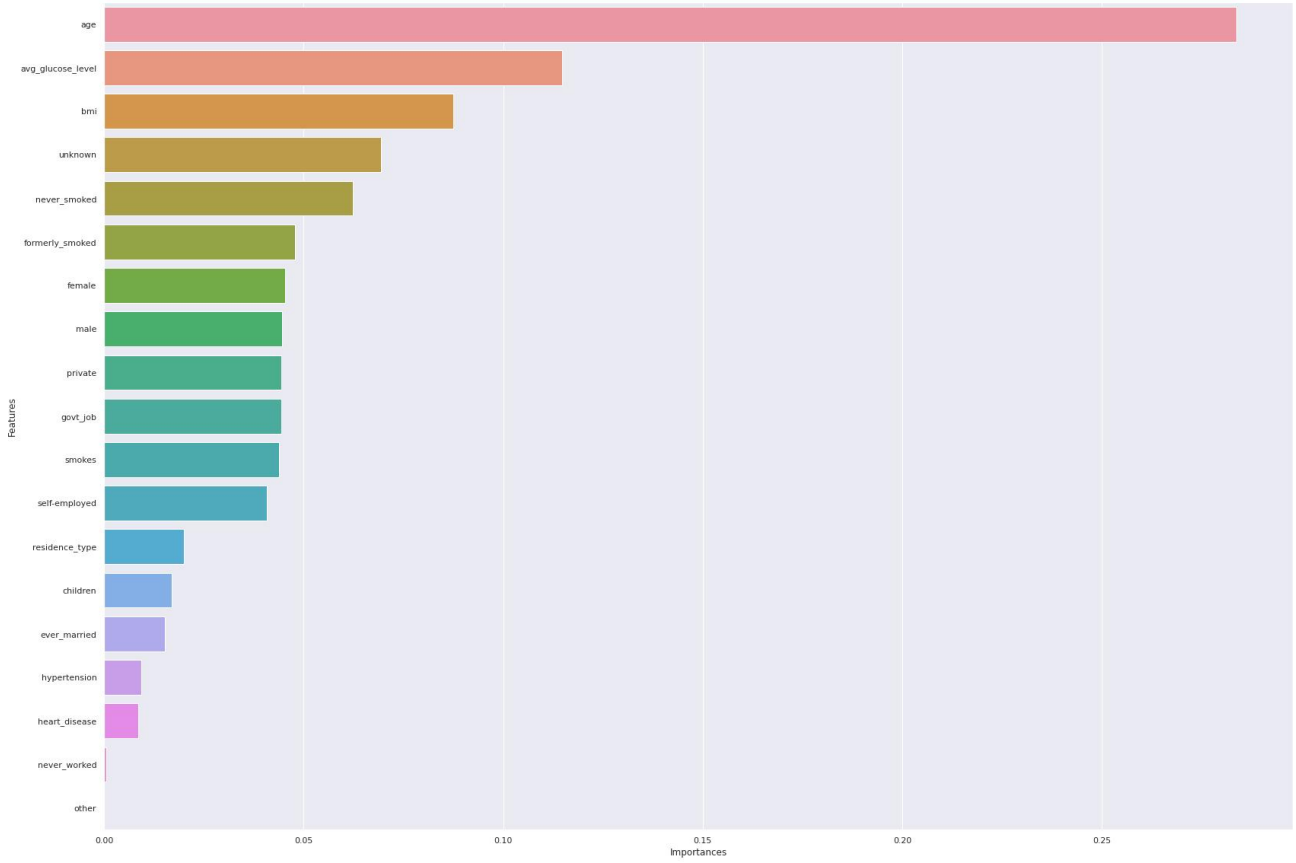

Figure 8 Relative importance of each feature for stroke prediction. BMI, body mass index.

Interestingly, the feature labeled "unknown" ranked fourth in terms of importance. However, it should be noted that this feature refers to missing data regarding the patient's smoking habits. Given that approximately 30% of the smoking status data was missing, it likely introduced uncertainty and potential confusion in the model. As a result, we recommend disregarding this feature in the analysis.

In contrast, the importance ranking of having a history of hypertension and heart disease was relatively low, which may appear surprising considering the well-established relationship between high blood pressure and stroke risk [30, 31]. However, it is important to consider that the predictive power of these features can be influenced by various factors, such as the specific characteristics of the dataset and the interactions between different variables. Further investigation and analysis should be undertaken to explore the underlying reasons for the observed results.

The graph of feature importance provides valuable insights into the relative contribution of each variable in predicting stroke. By identifying the most influential factors, healthcare professionals and policymakers can prioritize interventions and strategies that target these specific areas, such as promoting healthy aging, managing glucose levels, and addressing obesity. It is essential to continuously refine and update our understanding of the predictors of stroke to enhance risk assessment and preventive measures effectively.

Conclusions and future work

This study serves as compelling evidence for the accurate prediction of stroke risks through the application of machine learning. Among the various algorithms trained in this study, the random forest algorithm emerged as the most effective, exhibiting an impressive accuracy of 96.56%. The superior performance of the random forest algorithm suggests its suitability for stroke prediction. The significant contributions of this research can be summarized as follows: 1) the exploration of several HPML algorithms, leading to a substantial improvement in model performance, and 2) the provision of a framework that has the potential to assist in medical decision-making regarding stroke prediction.

Furthermore, this study identified the key factors that contribute to predicting the risk of stroke, including age, blood glucose levels, and BMI. These factors can be utilized as independent variables to construct descriptive theory models, with stroke as the dependent variable. By collecting new data from stroke patients across different regions, correlations can be identified through multiple regression analysis. It is also crucial to consider cultural nuances and tailor the models to specific regions. Such descriptive theory models would greatly aid in the design of applications that provide support to stroke patients. Moreover, a normative theory can be developed using longitudinal data, following inductive-deductive cycles of theory building.

While this study provides valuable insights into stroke prediction, it is essential to recognize several limitations. Firstly, the dataset used for analysis comprised only 10 attributes, which might limit the exploration of all potential predictors associated with stroke. Future research endeavors should prioritize the inclusion of additional predictive attributes such as drinking behaviors, blood pressure, and atrial fibrillation to enhance the comprehensiveness of the model.

Moreover, the reliance on experimental data solely from a publicly available database means that the study did not contribute any original laboratory findings or clinic observations. Addressing this limitation, future research will include plans to generate and incorporate original data into the existing database. This approach aims to enrich the dataset and strengthen the predictive capabilities of the model.

Additionally, the study's validation process relied exclusively on internal validation through training and test splits. To ensure the model's robustness and generalizability in real-world clinical practice, external validation is imperative. Ongoing efforts include deploying the prediction model in a local heart clinic to validate its performance in clinical settings. Furthermore, leveraging a larger dataset could further enhance the model's generalizability and applicability in diverse clinical contexts.

In conclusion, this study presented an integrated approach that combined data preprocessing and modeling techniques within the framework of machine learning. The findings and methodology described in this paper aim to inspire and encourage the wider application of machine learning methods for predicting stroke risks and automating stroke diagnosis. By leveraging the power of machine learning, the accuracy of stroke prediction can be significantly improved, leading to better patient outcomes and informed medical decision-making.

References

1. Katan M, Luft A. Global Burden of Stroke. *Semin Neurol.* 2018;38(02):208–211. Available at: http://doi.org/10.1055/s-0038-1649503
2. Kochanek KD, Murphy SL, Xu J, Arias E. Mortality in the United States, 2013. *NCHS Data Brief.* 2014;178:1–8. Available at: https://www.cdc.gov/nchs/data/databriefs/db178.pdf
3. Murphy SJX, Werring DJ. Stroke: causes and clinical features. *Medicine* (*Baltimore*). 2020;48(9):561–566. Available at: http://doi.org/10.1016/j.mpmed.2020.06.002
4. Stroke Facts. Centers for Disease Control and Prevention. [Internet]. cdc.gov. [cited 2023 Dec 10] Available at: https://www.cdc.gov/stroke/facts.htm
5. Preventing Stroke Deaths. Centers for Disease Control and Prevention. [Internet]. archive.cdc.gov. [cited 2023 Dec 10] Available at: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/vitalsigns/stroke/index.html
6. Taylor TN, Davis PH, Torner JC, Holmes J, Meyer JW, Jacobson MF. Lifetime Cost of Stroke in the United States. *Stroke.* 1996;27(9):1459–1466. Available at: http://doi.org/10.1161/01.STR.27.9.1459
7. Ovbiagele B, Goldstein LB, Higashida RT, et al. Forecasting the Future of Stroke in the United States. *Stroke.* 2013;44(8):2361–2375. Available at: http://doi.org/10.1161/STR.0b013e31829734f2
8. Know Your Risk for Stroke. Centers for Disease Control and Prevention. [Internet]. cdc.gov. [cited 2023 Dec 10] Available at: https://www.cdc.gov/stroke/risk_factors.htm
9. Kotlęga D, Gołąb-Janowska M, Masztalewicz M, Ciećwież S, Nowacki P. The emotional stress and risk of ischemic stroke. *Neurol Neurochir Pol.* 2016;50(4):265–270. Available at: http://doi.org/10.1016/j.pjnns.2016.03.006
10. Kamal H, Lopez V, Sheth SA. Machine Learning in Acute Ischemic Stroke Neuroimaging. *Front Neurol.* 2018;9:945. Available at: http://doi.org/10.3389/fneur.2018.00945
11. Feng R, Badgeley M, Mocco J, Oermann EK. Deep learning guided stroke management: a review of clinical applications. *J NeuroIntervent Surg.* 2017;10(4):358–362. Available at: http://doi.org/10.1136/neurintsurg-2017-013355
12. Lee E-J, Kim Y-H, Kim N, Kang D-W. Deep into the Brain: Artificial Intelligence in Stroke Imaging. *J Stroke.* 2017;19(3):277–285. Available at: http://doi.org/10.5853/jos.2017.02054
13. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine Learning–Based Model for Prediction of Outcomes in Acute Stroke. *Stroke.* 2019;50(5):1263–1265. Available at: http://doi.org/10.1161/STROKEAHA.118.024293
14. Wu Y, Fang Y. Stroke Prediction with Machine Learning Methods among Older Chinese. *Int J Environ Res Public Health.* 2020;17(6):1828. Available at: http://doi.org/10.3390/ijerph17061828
15. Wang W, Kiik M, Peek N, et al. A systematic review of machine learning models for predicting outcomes of stroke with

structured data. *PLoS One.* 2020;15(6):e0234722. Available at: https://doi.org/10.1371/journal.pone.0234722
16. Khosla A, Cao Y, Lin CC-Y, Chiu H-K, Hu J, Lee H. An integrated machine learning approach to stroke prediction. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2010:183–192. Available at: http://doi.org/10.1145/1835804.1835830
17. Chun M, Clarke R, Cairns BJ, et al. Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. *J Am Med Inform Assoc.* 2021;28(8):1719–1727. Available at: https://doi.org/10.1093/jamia/ocab068
18. Hung C-Y, Chen W-C, Lai P-T, Lin C-H, Lee C-C. Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. *Annu Int Conf IEEE Eng Med Biol Soc.* 2017;2017:3110–3113. Available at: http://doi.org/10.1109/EMBC.2017.8037515
19. Sailasya G, Kumari GLA. Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. *IJACSA.* 2021;12(6):539–545. Available at: http://doi.org/10.14569/IJACSA.2021.0120662
20. Manolio TA, Kronmal RA, Burke GL, O'Leary DH, Price TR. Short-term Predictors of Incident Stroke in Older Adults. *Stroke.* 1996;27(9):1479–1486. Available at: http://doi.org/10.1161/01.STR.27.9.1479
21. Longstreth WT Jr, Bernick C, Fitzpatrick A, et al. Frequency and predictors of stroke death in 5,888 participants in the Cardiovascular Health Study. *Neurology.* 2001;56(3):368–375. Available at: http://doi.org/10.1212/WNL.56.3.368
22. Lumley T, Kronmal RA, Cushman M, Manolio TA, Goldstein S. A stroke prediction score in the elderly. *J Clin Epidemiol.* 2002;55(2):129–136. Available at: http://doi.org/10.1016/S0895-4356(01)00434-6
23. McGinn AP, Kaplan RC, Verghese J, et al. Walking Speed and Risk of Incident Ischemic Stroke Among Postmenopausal Women. *Stroke.* 2008;39(4):1233–1239. Available at: http://doi.org/10.1161/STROKEAHA.107.500850
24. Boden-Albala B, Sacco RL. Lifestyle factors and stroke risk: Exercise, alcohol, diet, obesity, smoking, drug use, and stress. *Curr Atheroscler Rep.* 2000;2(2):160–166. Available at: http://doi.org/10.1007/s11883-000-0111-3
25. Kwon Y, Norby FL, Jensen PN, et al. Association of Smoking, Alcohol, and Obesity with Cardiovascular Death and Ischemic Stroke in Atrial Fibrillation: The Atherosclerosis Risk in Communities (ARIC) Study and Cardiovascular Health Study (CHS). *PLoS One.* 2016;11(1):e0147065. Available at: https://doi.org/10.1371/journal.pone.0147065
26. Viscoli CM, Brass LM, Kernan WN, Sarrel PM, Suissa S, Horwitz RI. A Clinical Trial of Estrogen-Replacement Therapy after Ischemic Stroke. *N Engl J Med.* 2001;345(17):1243–1249. Available at: http://doi.org/10.1056/NEJMoa010534
27. Stroke prediction dataset (2020). Kaggle. [Internet]. kaggle.com. [cited 2023 Dec 10]. Available at: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
28. Bruno A, Biller J, Adams HP Jr, et al. Acute blood glucose level and outcome from ischemic stroke. *Neurology.* 1999;52(2):280–284. Available at: http://doi.org/10.1212/WNL.52.2.280
29. Cipolla MJ, Liebeskind DS, Chan S-L. The importance of comorbidities in ischemic stroke: Impact of hypertension on the cerebral circulation. *J Cereb Blood Flow Metab.* 2018;38(12):2129–2149. Available at: http://doi.org/10.1177/0271678X18800589
30. Lawes CMM, Bennett DA, Feigin VL, Rodgers A. Blood Pressure and Stroke. *Stroke.* 2004;35(3):776–785. Available at: http://doi.org/10.1161/01.STR.0000116869.64771.5A
31. Willmot M, Leonardi-Bee J, Bath PMW. High Blood Pressure in Acute Stroke and Subsequent Outcome. *Hypertension.* 2004;43(1):18–24. Available at: http://doi.org/10.1161/01.HYP.0000105052.65787.35