Original Article

Leveraging multivariate analysis and adjusted mutual information to improve stroke prediction and interpretability

Moutasem S. Aboonq, MD, PhD, Saeed A. Alqahtani, MD, PhD.

ABSTRACT

الأهداف: تطوير نموذج للتنبؤ بخطر الإصابة بالسكتة الدماغية بدقة عالية، مع تحديد أهم عوامل الخطر المرتبطة بها واختيار أفضل خوارزمية تعلم آلي لهذا الغرض.

المنهجية: تم تحليل بيانات 438,693 بالغًا من نظام مسح عوامل الخطر السلوكية لعام 2021. وشملت الدراسة الخصائص الديموغرافية والعوامل السريرية للمشاركين، واستخدمت أساليب إحصائية مختلفة مثل الانحدار اللوجستي والمعلومات المتبادلة لتحديد العلاقات بين العوامل المختلفة وخطر السكتة الدماغية. بالإضافة إلى ذلك، تم بناء وتقييم عدة نماذج تعلم آلي للتنبؤ بالسكتة الدماغية.

النتائج: أظهرت النتائج أن العمر، ومرض السكري، وارتفاع ضغط الدم، وارتفاع الكولسترول، والتاريخ المرضي لأمراض القلب والأوعية الدموية هي عوامل خطر رئيسية للإصابة بالسكتة الدماغية. وكان نموذج الغابة العشوائية هو الأفضل في التنبؤ بخطر السكتة الدماغية، حيث حقق دقة 72.46%.

الخلاصة: يمكن استخدام نموذج الغابة العشوائية للتنبؤ بدقة بخطر الإصابة بالسكتة الدماغية بناءً على المعلومات الديموغرافية والسريرية. وتؤكد هذه الدراسة على أهمية مراقبة وعلاج عوامل الخطر مثل ارتفاع ضغط الدم والسكري لتقليل خطر الإصابة بالسكتة الدماغية.

Objectives: To develop a machine learning model to accurately predict stroke risk based on demographic and clinical data. It also sought to identify the most significant stroke risk factors and determine the optimal machine learning algorithm for stroke prediction.

Methods: This cross-sectional study analyzed data on 438,693 adults from the 2021 Behavioral Risk Factor Surveillance System. Features encompassed demographics and clinical factors. Descriptive analysis profiled the dataset. Logistic regression quantified risk relationships. Adjusted mutual information evaluated feature importance. Multiple machine learning models were built and evaluated on metrics like accuracy, AUC ROC, and F1 score.

Results: Key factors significantly associated with higher stroke odds included older age, diabetes, hypertension, high cholesterol, and history of myocardial infarction or angina. Random forest model achieved the best performance with accuracy of 72.46%, AUC ROC of 0.72, and F1 score of 0.74. Cross-validation confirmed its reliability. Top features were hypertension, myocardial infarction history, angina, age, diabetes status, and cholesterol.

Conclusion: The random forest model robustly predicted stroke risk using demographic and clinical variables. Feature importance highlighted priorities like hypertension and diabetes for clinical monitoring and intervention. This could help enable data-driven stroke prevention strategies.

Neurosciences 2024; Vol. 29 (3): 190-196
doi: 10.17712/nsj.2024.3.20230100

From the Department of Physiology, College of Medicine, Taibah University, Al-Madinah Al-Munawwarah, Kingdom of Saudi Arabia

Received 11th October 2023. Accepted 31st May 2024.

Address correspondence and reprint request to: Dr. Moutasem S. Aboonq, Department of Physiology, College of Medicine, Taibah University, Al-Madinah Al-Munawwarah, Kingdom of Saudi Arabia.
E-mail: aboonq@yahoo.co.uk
ORCID ID: https://orcid.org/0000-0002-9137-941

Stroke is a devastating medical condition and a leading cause of long-term disability worldwide. It is the second most common cause of death and the third most common cause of death and disability combined.[1]

Disclosure. The authors declare no conflicting interests, support or funding from any drug company.

In the United States alone, stroke accounts for one of every 19 deaths.[2] In addition to being a major cause of morbidity and mortality, stroke is also a significant financial burden on the healthcare system. The projected total cost of stroke in the United States in 2035 is expected to be $129.3 billion, including the direct costs of healthcare services and medications, as well as the indirect costs of lost productivity and premature death.[3]

Several demographic and clinical attributes are recognised as risk factors for the development of stroke. Non-modifiable factors established through epidemiological research include advanced age, male sex, and family history of stroke.[4,5] Modifiable risk factors centred around lifestyle and underlying medical conditions have also been identified. The leading risk factors in this category include hypertension, diabetes mellitus, cardiovascular disease, smoking, obesity, physical inactivity, and poor diet.[6,7] A comprehensive understanding of how these attributes interact and jointly influence stroke probability could help guide targeted preventive strategies.

Machine learning presents an opportunity to advance stroke risk prediction through automated detection of complex patterns in large health datasets.[8,9] By considering nonlinear relationships between diverse risk factors, machine learning models show the potential for more accurate risk stratification compared to conventional regression analyses.[10] Although several studies have demonstrated the relevance of machine learning applications to predicting stroke outcomes, further work is still needed that utilises robust datasets and compares model performance. Furthermore, consensus is still evolving on best practices for machine learning workflows in stroke research, including optimal models, feature selection methods, and performance metrics.[11–13]

The study aimed to develop a machine learning model that could accurately predict a patient's risk of stroke based on their demographic and clinical data. The secondary objective was to identify the most important risk factors for stroke and to determine the best machine learning algorithm for stroke prediction. By advancing the prediction of future stroke cases through an automated analysis of risk attributes, this research strives to contribute novel insights with implications for both clinical practice and public health policymaking. Facilitating the early identification of high-risk individuals could empower lifestyle modifications and medical optimisation to reduce stroke occurrences on a population scale. In turn, this may help reduce the immense human and economic burdens associated with strokes worldwide.

Methods.

Data source. The data for this study were sourced from the 2021 Behavioral Risk Factor Surveillance System (BRFSS), a publicly available dataset managed by the U.S. Centers for Disease Control and Prevention and released under the CC0 1.0 Universal (CC0 1.0) Public Domain Dedication license.[14] Due to its open nature, no ethical approval or informed consent was required for its use.

Data collection and preprocessing. Data preprocessing procedures were carried out using the Python programming language within the Google Colab environment. These procedures included data cleansing, feature selection, and feature engineering. Missing values were addressed, and relevant attributes were selected. The dataset's missing values were addressed using Iterative Imputer with a Logistic Regression estimator. This approach predicts missing values based on existing feature relationships. By fitting and transforming the data, missing values were replaced with informed estimations, resulting in a complete and consistent dataset for analysis. This method preserves dataset integrity by leveraging inherent correlations for improved imputation accuracy. Feature engineering involves both the merger of existing features and the creation of new ones. With an initial dataset of 438,693 records, the 'diabetic status' variable was derived by categorizing individuals as 'not diabetic' or 'diabetic,' following the removal of prediabetic and gestational diabetic records from the 'diabetic status' variable. To predict stroke status, we considered categorical variables such as gender, age group, body mass index (BMI) category, smoking status, diabetic status, hypertension status, cholesterol status, myocardial infarction (MI), and angina or coronary heart disease (angina/CHD), **Table 1**.

Descriptive analysis. Descriptive analysis was employed to summarize the categorical variables and their respective group distributions within the dataset. The percentage distribution of groups within each variable was computed to offer a comprehensive overview of the dataset's composition.

Logistic regression analysis. A logistic regression analysis was conducted to examine the relationships between all predictor features and the target variable diabetic status. Odds ratios (ORs), 95% confidence intervals (CIs), and *p*-values were calculated to measure the strengths of these associations and assess their statistical significance, with *p*-values below 0.05 considered significant.

Feature importance assessment. Feature importance in predicting the stroke status target variable was assessed using the Adjusted Mutual Information

(AMI) method. Adjusted Mutual Information, which accounts for chance agreement, measures mutual information between variables while ensuring there is no shared information among the features, enhancing its effectiveness for feature evaluation.

Model selection and evaluation. Multiple machine learning models were employed to predict stroke status, and their performance was evaluated using various metrics, including accuracy, area under the ROC curve (AUC ROC), precision, recall, and F1 score. The best-performing model underwent retraining and cross-validation to ensure its robustness. The dataset exhibits a significant class imbalance, with the "no stroke" class having 421,479 instances and the "stroke" class having only 17,214 instances. This highlights a ratio of approximately 25:1, where the "stroke" class is severely underrepresented. To ensure our model learns effectively from both classes and makes accurate predictions, addressing this imbalance is crucial. We will employ the Synthetic Minority Over-sampling Technique (SMOTE) to handle this issue. The models encompassed various techniques, including Random Forest, K Nearest Neighbors (KNN), Extreme Gradient Boosting (XGBoost), Neural Network, and Logistic Regression where default parameters were used for all models. SMOTE was employed to mitigate imbalanced data, generating synthetic minority class samples while minimizing overfitting. To rigorously validate the results, a 5-fold cross-validation approach was applied to the best-performing model.

Results.

Descriptive analysis. This study presents a comprehensive analysis of demographic and health-

Table 1 - Variables description.

Variables	Definitions
stroke status	Stroke: 0 (no history of stroke), 1 (has history of stroke)
gender	Gender: 0 (Female), 1 (Male)
Age group in year (y):	Age group: 13-level category (1: 18-24 y, 2: 25-29 y, 3: 30-34 y, 4: 35-39 y, 5: 40-44 y, 6: 45-49 y, 7: 50-54 y, 8: 55-59 y, 9: 60-64 y, 10: 65-69 y, 11: 70-74 y, 12: 75-79 y, 13: 80 y or above)
BMI (Body Mass Index):	Body Mass Index: 1: Underweight (BMI < 18.5 Kg/m^2), 2: Normal weight (BMI 18.5-24.9 Kg/m^2), 3: Overweight (BMI 25 - 29.9 Kg/m^2), 4: Obese (BMI ≥30 Kg/m^2)
smoking status	Smoker: 0 (no), 1 (yes)
diabetic status	Diabetic: 0 (Not Diabetic), 1 (Diabetic)
hypertension status	Hypertension: 0 (No Hypertension), 1 (Hypertension)
cholesterol status	High Cholesterol: 0 (No High Cholesterol), 1 (High Cholesterol)
MI	History of MI: 0 (no), 1 (yes)
angina/CHD	History of angina or CHD: 0 (no), 1 (yes)

BMI: Body mass index, y: year, MI: myocardial infarction, CHD: coronary heart disease

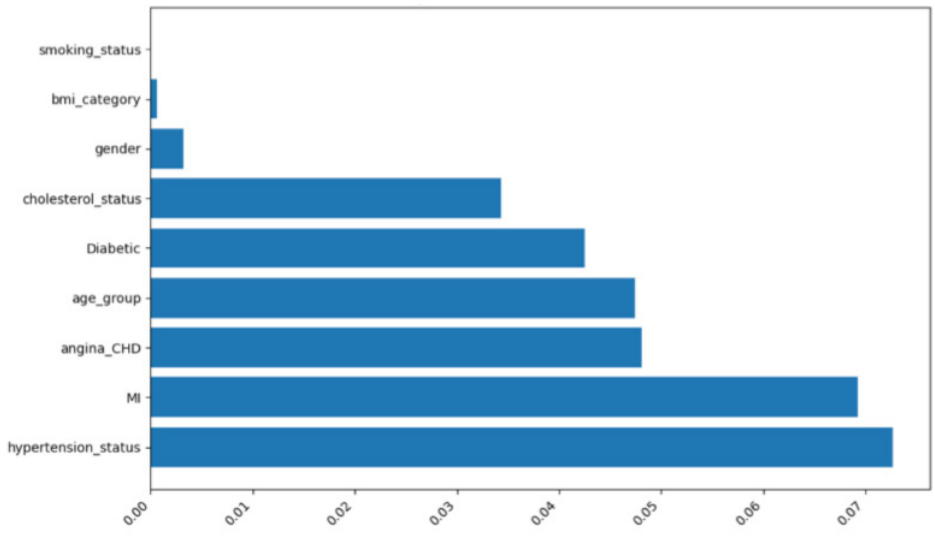

Figure 1 - Adjusted mutual information for the independent variables.

Table 2 - Multivariate regression analysis.

Features	OR	CI
gender	0.626	0.620-0.632
age group	1.103	1.102-1.104
BMI category	0.613	0.611-0.615
smoking status	1.028	1.013-1.043
diabetic	2.111	2.084-2.138
cholesterol status	1.197	1.185-1.210
MI	3.379	3.317-3.441
angina CHD	1.733	1.701-1.766
hypertension status	2.304	2.279-2.329

related features within the studied population, shedding light on critical characteristics that contribute to our understanding of population health dynamics. The gender distribution within the population is nearly identical, with 53.6% identifying as males and 46.4% as females. The study sample demonstrates a diverse age composition, with the largest age group being (65-69 years), comprising 10.7% of the population. Following closely are the 60-64 and 70-74 age groups, accounting for 10.3% and 10.0%, indicating significant age diversity. Regarding BMI, most of the population falls into overweight (35.4%) and obese groups (33.5%), while a small percentage falls into underweight category (1.6%). Smoking status indicates that 87% of the population does not smoke, while 13% are current or former smokers. Diabetes status reveals that 83.7% of the population does not have diabetes, while 13.2% have diabetes. The analysis of hypertension status within the population indicates a significant distribution. Approximately 60.6% of individuals do not have hypertension, while 39.4% have been diagnosed with hypertension. Cholesterol levels show that 60% of the population has normal cholesterol levels, while 40% have elevated cholesterol levels. Most individuals, accounting for 94.8%, have not experienced an MI, however, 5.2% of individuals have a history of MI. Many individuals, representing 94.7% of the population, have not been diagnosed with angina or CHD. Conversely, 5.3% of individuals have been diagnosed with angina or CHD.

Logistic regression. In the multivariate logistic regression analysis results, various independent variables were examined in relation to the dependent variable, stroke status. The ORs, along with their corresponding 95% CIs, provide insights into the likelihood of stroke occurrence, **Table 2**. Gender demonstrated 37.5% lower odds of stroke for male compared to the female. Age group showed that for each one-unit increase in age category, the odds of experiencing a stroke increased by 10.3%. Different BMI categories were associated with 38.7% lower odds of stroke. Smoking status was linked to a 2.8% increase in stroke risk among smokers. Individuals with diabetes had approximately 110.9% higher odds of experiencing a stroke. Those with a history of MI had approximately 278.9% higher odds, while those with angina or CHD had approximately 73.3% higher odds. Hypertension was associated with approximately 130.3% higher odds of stroke. Conversely, cholesterol status was linked to a 19.7% increase in stroke odds. The relationship between all features and the target variable is statistically significant. These findings provide a quantified understanding of how each factor contributes to the probability of stroke within the studied population.

Table 3 - Predictive models performance.

Models	Accuracy	Precision	Recall	F1 Score	AUC ROC
Random Forest	0.725	0.704	0.775	0.738	0.725
XGBoost	0.721	0.700	0.775	0.735	0.721
Neural Network	0.717	0.688	0.794	0.738	0.717
Logistic Regression	0.706	0.699	0.724	0.711	0.706
KNN	0.690	0.682	0.713	0.697	0.690

Feature importance. The adjusted mutual information scores for various features concerning their relationship with the target variable, stroke status, reveal that the top 6 most informative features collectively account for over 98% of the predictive importance in this analysis, **Figure 1**. These features, namely hypertension status (27.1%), MI (25.85%), angina/CHD (17%), age group (17.7%), diabetic status (15.9%), and cholesterol status (12.9%), display notably higher AMI scores and percentages, indicating a relatively stronger association with the prediction of stroke status. In contrast, gender, BMI category, and smoking status exhibit significantly lower percentages, ranging from 0.01% to 1.2%, suggesting a weaker relationship with the outcome.

Models' performance. The comparative analysis of different machine learning models for predicting stroke status reveals interesting insights, **Table 3**. Random Forest and XGBoost exhibit the highest accuracy, at 72.5% and 72.1%, respectively, indicating their overall correctness in predictions. Random Forest stands out in precision, capturing 70.4% of true stroke cases among its positive predictions, closely followed by XGBoost at 70.0%. However, Neural Network excels in recall, correctly identifying 79.5% of actual stroke cases. When considering a balanced metric like the F1 Score, which accounts for both precision and recall, Random Forest and Neural Network lead the pack with scores of 73.8% and 73.75%, respectively. Finally,

the AUC ROC, which measures a model's ability to distinguish between positive and negative cases, aligns closely with accuracy, again placing Random Forest and XGBoost at the forefront. Based on the provided metrics and their comparative analysis, it appears that the Random Forest model is the best-performing model for predicting stroke status. It achieves the highest accuracy, competitive precision, and recall, and it also has a strong F1 score and AUC ROC. To validate the Random Forest model, which performed the best, we used 5-fold cross-validation. This resulted in a mean accuracy score of 0.723 and a standard deviation score of 0.002, which suggests that the model is reliable.

Discussion. This comprehensive analysis of the demographic and health determinants of stroke risk provides critical insights that align with and build upon the existing body of research. Specifically, both the descriptive overview of the population sample and the predictive modelling unveil key patterns and risk relationships with implications for both prevention and care. In this regard, the gender distribution proves notable, with a slightly higher representation of males at 53.6% versus 46.4% females. This approximates the gender ratio within the overall U.S. population and many prior stroke studies, thereby facilitating generalizability.[15,16] Gender emerged as a significant predictor in the regression analysis, with males demonstrating 37.5% lower stroke odds, which suggests that potential physiological or lifestyle factors may be driving higher female susceptibility. Although age and comorbidities can attenuate this difference, a female susceptibility across groups was found to persist in earlier studies.[15,17] This discrepancy warrants a closer look at cohort studies and clinical trials, particularly to ascertain any differential responses to treatments.

The age distribution revealed a high level of diversity, with the most prevalent group being 65–69 years old. This wide range increases the suitability of generalised inferences across ages. It also corresponds with the literature indicating steadily increasing stroke incidence after age 55, peaking in the 60s and 70s.[18–20] Age also exhibited a clear association in the regression, with 10.3% higher odds per age category.

Over two-thirds of the population in this study fell into the overweight or obese BMI classes. While the regression analysis found that people with higher BMIs had a 38.7% lower chance of stroke, other studies have shown mixed results. For instance, some studies have found that obesity increases the risk of stroke, especially in younger people.[21,22] Other studies have found that low body weight also increases the risk of stroke.[23] Additional research into BMI's age interactions may help explain this discrepancy.[24] Regardless, weight's interactions with stroke pathology merit additional probe.

Most individuals who participated in this study did not currently smoke, though 13% reported being former smokers. This aligns with declining national smoking trends.[25] Smoking proved to be a significant but weaker predictor, with just 2.8% higher stroke odds. This corroborates research showing smoking as a less dominant risk factor compared to hypertension or diabetes.[26] Prevalence of diabetes (13.2%) and hypertension (39.4%) closely mirrored nationwide estimates of 10% for diabetes and 43% for hypertension among U.S. adults.[27–29] Our results showed that people with diabetes had 110.9% higher odds of stroke, whereas those with hypertension had 130.3% higher odds. This highlights the significant burden of disease that these conditions place on individuals and society. Over one-third of the population exhibits elevated cholesterol, although its regression OR is more modest at 1.197. This aligns with some studies positioning cholesterol as a significant but weaker metabolic factor than others, such as diabetes, in multivariate analyses.[18] Finally, just 5.2% and 5.3% of patients had an MI or angina/CHD history, respectively. However, these small groups face heavily amplified stroke odds of 278.9% and 73.3%, which underlines the influential risk conferred by these atherosclerotic conditions, especially MI.[30,31]

The adjusted mutual information and feature importance scores provide a powerful synopsis of the multivariate analysis by distilling the most predictive input variables. Hypertension, unsurprisingly, had the strongest association with stroke probability at over 27% explanatory power. Its overwhelming impact here aligns with its prior designation as the single most critical and modifiable stroke risk factor.[32,33] Prior MI and angina/CHD followed hypertension, accounting for over 25% and 17% of the explanatory ability, respectively. This accords with their dramatic ORs and status as major non-modifiable precursors of stroke pathology in the existing literature.[30,31] Diabetes and age also featured prominently, collectively accounting for over 30% of predictive capacity and confirming their role as leading stroke determinants.[18] In contrast, smoking, BMI, and gender displayed AMI scores below 2%, which supports their non-significant or surprising regression relationships in this sample. The feature importance thus efficiently summarises the variables most strongly and directly associated with stroke occurrence, thereby highlighting priorities for clinical intervention and monitoring.

The machine learning models provided predictive power beyond the regression insights, with the top-performing algorithms achieving over 70% predictive accuracy. Among the machine learning algorithms compared, random forest achieved the best performance for stroke prediction based on balanced metrics, including accuracy (72.5%), precision (70.4%), F1 score (73.8%), and AUC ROC (72.5%). Its strength in handling nonlinear relationships and the interactivity of variables is consistent with the complexity of biological systems underlying stroke aetiology.[34] However, to improve the performance of a machine learning model, hyperparameter tuning could be used. Two common approaches supported by the scikit-learn library are Grid Search CV and Randomized Search CV. Validating the optimal random forest model via cross-validation yielded consistent accuracy, which confirmed its reliability for generalisable inferences. Overall, the machine learning models delivered robust predictive analytics while spotlighting specific high-risk factors through feature importance interpretations.

Several other studies have proposed strategies for stroke prediction based on machine learning (ML) algorithms, with excellent results being reported.[35–40] However, there is great diversity in the variables analysed (clinical, molecular markers, or imaging), calibration/training protocols performed, and models implemented (neural networks, tree-based, and kernel-based methods). One study analysed an ML model of stroke prediction at three months using the Hospital's Stroke Registry (BICHUS) on the basis of demographic, clinical, molecular, and neuroimaging variables.[38] The ML classifiers exhibited high performance with over 0.90 AUC in the 3 groups evaluated in relation to the mortality outcome. Another study used ML techniques to predict outcomes after endovascular treatment in stroke patients.[37] In this case, the authors developed a model that could predict the likelihood of a good outcome with an accuracy of 0.81. Another study developed a score that can be used to predict the probability of the patient achieving each of five categories of the Barthel Index score at discharge from rehabilitation.[40] Generally, previous studies on stroke prediction have often been limited by small sample sizes or models that are difficult to interpret. Our study addresses these limitations by using a very large sample size and a combination of 2 statistical methods (AMI and multivariate regression analysis) to explain how each variable contributes to predicting stroke status. This makes our predictive model more reliable and interpretable, which means that it is more likely to be accurate and can be used to better understand the risk factors for stroke.

Despite its value, our study has some limitations that deserve discussion. As this was an observational study, causality between risk factors and stroke occurrence could not be definitively established. Furthermore, residual confounding from unmeasured variables is possible, and model generalizability beyond the study population remains to be validated on independent datasets. The cross-sectional design precluded the capture of time-dependent exposures, such as cumulative smoking pack-years. As another limitation, predictor definitions limited granularity, for instance, by collapsing BMI categories. Predictors focused on classic risk factors and that exclude novel biometrical or omics data could confer added predictive value.

Overall, the multidimensional analytics strongly synthesize established knowledge regarding stroke epidemiology while illuminating new patterns ripe for further investigation. The findings consolidate valuable insights for clinicians to enhance screening and prevention initiatives targeting the most influential risk factors in susceptible populations. They also underscore key variable relationships warranting refined understanding through additional empirical research. The model's predictive capability lays foundation for deployment in clinical decision support and population health management applications.

Acknowledgement. *We would like to thank Scribendi Inc. For English language editing.*

References

1. Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. ***Int J Stroke*** 2022; 17: 18-29.
2. Stroke Facts. National Center for Chronic Disease Prevention and Health Promotion; About the Division for Heart Disease and Stroke Prevention 2024. https://www.cdc.gov/stroke/data-research/facts-stats/index.html
3. Tsao CW, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, et al. Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association. ***Circulation*** 2023; 147: e93–e621
4. Stroke Risk Factors, Genetics, and Prevention | Circulation Research [Internet]. [cited 2022 Dec 29]. Available from: https://www.ahajournals.org/doi/10.1161/circresaha.116.308398
5. Diaz MA, Rosendale N. Exploring Stroke Risk Factors and Outcomes in Sexual and Gender Minority People. ***Neurol Clin Pract*** 2023; 13: 1-10.
6. Bukhari S, Yaghi S, Bashir Z. Stroke in Young Adults. ***J Clin Med*** 2023; 12: 4999.
7. Delgardo M, Rabin G, Tudor T, Tang AJ, Reeves G, Connolly ES. Monitoring risk and preventing ischemic stroke in the very old. ***Expert Rev Neurother*** 2023; 23: 791-801.
8. Hong C, Pencina MJ, Wojdyla DM, Hall JL, Judd SE, Cary M, et al. Predictive accuracy of stroke risk prediction models across black and white race, sex, and age groups. ***JAMA*** 2023; 329: 306-317.

9. Zafeiropoulos N, Mavrogiorgou A, Kleftakis S, Mavrogiorgos K, Kiourtis A, Kyriazis D. Interpretable Stroke Risk Prediction Using Machine Learning Algorithms. In: Nagar AK, Singh Jat D, Mishra DK, Joshi A, editors. Intelligent Sustainable Systems. Singapore: Springer Nature Singapore; 2023. p. 647–56. Available from: https://link.springer.com/10.1007/978-981-19-7663-6_61
10. Khosravi B, Weston AD, Nugen F, Mickley JP, Kremers HM, Wyles CC, et al. Demystifying statistics and machine learning in analysis of structured tabular data. ***J Arthroplasty*** 2023; 38: 1943-1947.
11. Chandrabhatla AS, Kuo EA, Sokolowski JD, Kellogg RT, Park M, Mastorakos P. Artificial Intelligence and Machine Learning in the Diagnosis and Management of Stroke: A Narrative Review of United States Food and Drug Administration-Approved Technologies. ***J Clin Med*** 2023; 12: 3755.
12. Sheth SA, Giancardo L, Colasurdo M, Srinivasan VM, Niktabe A, Kan P. Machine learning and acute stroke imaging. ***J Neurointerventional Surg*** 2023; 15: 195-199.
13. Daidone M, Ferrantelli S, Tuttolomondo A. Machine learning applications in stroke medicine: advancements, challenges, and future prospectives. ***Neural Regen Res*** 2024; 19: 769-773.
14. CDC - 2021 BRFSS Survey Data and Documentation [Internet]. 2023 [cited 2023 Oct 1]. Available from: https://www.cdc.gov/brfss/annual_data/annual_2021.html
15. Reeves MJ, Bushnell CD, Howard G, Gargano JW, Duncan PW, Lynch G, et al. Sex differences in stroke: epidemiology, clinical presentation, medical care, and outcomes. ***Lancet Neurol*** 2008; 7: 915-926.
16. Ahnstedt H, McCullough LD, Cipolla MJ. The Importance of Considering Sex Differences in Translational Stroke Research. ***Transl Stroke Res*** 2016; 7: 261-273.
17. Shajahan S, Sun L, Harris K, Wang X, Sandset EC, Yu AY, et al. Sex differences in the symptom presentation of stroke: A systematic review and meta-analysis. ***Int J Stroke*** 2023; 18: 144-153.
18. Chen RL, Balami JS, Esiri MM, Chen LK, Buchan AM. Ischemic stroke in the elderly: an overview of evidence. ***Nat Rev Neurol*** 2010; 6: 256-265.
19. Yousufuddin M, Young N. Aging and ischemic stroke. ***Aging*** 2019; 11: 2542.
20. Popa-Wagner A, Petcu EB, Capitanescu B, Hermann DM, Radu E, Gresita A. Ageing as a risk factor for cerebral ischemia: underlying mechanisms and therapy in animal models and in the clinic. ***Mech Ageing Dev*** 2020; 190: 111312.
21. Strazzullo P, D'Elia L, Cairella G, Garbagnati F, Cappuccio FP, Scalfi L. Excess Body Weight and Incidence of Stroke: Meta-Analysis of Prospective Studies With 2 Million Participants. ***Stroke*** 2010; 41: e418–e426. Available from: https://www.ahajournals.org/doi/10.1161/STROKEAHA.109.576967
22. Horn JW, Feng T, Mørkedal B, Aune D, Strand LB, Horn J, et al. Body Mass Index Measured Repeatedly over 42 Years as a Risk Factor for Ischemic Stroke: The HUNT Study. ***Nutrients*** 2023;15: 1232.
23. Hamatani Y, Ogawa H, Uozumi R, Iguchi M, Yamashita Y, Esato M, et al. Low body weight is associated with the incidence of stroke in atrial fibrillation patients–Insight from the Fushimi AF Registry. ***Circ J*** 2015; 79: 1009-1017.
24. Ye J, Hu Y, Chen X, Yin Z, Yuan X, Huang L, et al. Association between the weight-adjusted waist index and stroke: a cross-sectional study. ***BMC Public Health*** 2023; 23: 1689.
25. Kasza KA, Tang Z, Xiao H, Marshall D, Stanton C, Gross A, et al. National longitudinal tobacco product discontinuation rates among US youth from the PATH Study: 2013–2019 (waves 1–5). ***Tob Control*** 2024; 33: 186-192.
26. Stanaway JD, Afshin A, Gakidou E, Lim SS, Abate D, Abate KH, et al. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. ***The Lancet*** 2018; 392: 1923-1994.
27. Chiu N, Chiu L, Aggarwal R, Raber I, Bhatt DL, Mukamal KJ. Trends in Blood Pressure Treatment Intensification in Older Adults With Hypertension in the United States, 2008 to 2018. ***Hypertension*** 2023; 80: 553-562.
28. Diabetes Statistics. DRIF. [cited 2023 Sep 29]. Available from: https://diabetesresearch.org/diabetes-statistics/
29. CDC. By the Numbers: Diabetes in America. Centers for Disease Control and Prevention. 2022 [cited 2023 Sep 29]. Available from: https://www.cdc.gov/diabetes/health-equity/diabetes-by-the-numbers.html
30. Bierbower E, Griffith N, Raman VK, Brar V, Roseman J, Deedwania P, et al. Risk of Stroke in Older Adults With Heart Failure. ***Am J Cardiol*** 2023; 189: 70-75.
31. Ho JSY, Sia CH, Zheng H, Tan BYQ, Ho AFW, Yeo LLL, et al. Interplay between post-myocardial infarction ejection fraction and atrial fibrillation: implications for ischemic stroke. ***Eur Heart J*** 2023; 44: ehac779-060.
32. Wajngarten M, Silva GS. Hypertension and Stroke: Update on Treatment. ***Eur Cardiol Rev*** 2019; 14: 111-115.
33. Fishman B, Bardugo A, Zloof Y, Bendor CD, Libruder C, Zucker I, et al. Adolescent Hypertension Is Associated With Stroke in Young Adulthood: A Nationwide Cohort of 1.9 Million Adolescents. ***Stroke*** 2023; 54: 1531-1537.
34. Charmilisri A, Harshi I, Madhushalini V, Raja L. Enhanced Stroke Prediction through Recursive Feature Elimination and Cross-Validation in Machine Learning. In: 2023 8th International Conference on Communication and Electronics Systems (ICCES) [Internet]. IEEE; 2023. p. 1075–1080. Available from: https://ieeexplore.ieee.org/abstract/document/10192685/
35. Kaur M, Sakhare SR, Wanjale K, Akter F. Early Stroke Prediction Methods for Prevention of Strokes. ***Behav Neurol*** 2022; 2022: 7725597.
36. Alanazi EM, Abdou A, Luo J. Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models. ***JMIR Form Res*** 2021; 5: e23440.
37. 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: 1263-1265.
38. Fernandez-Lozano C, Hervella P, Mato-Abad V, Rodríguez-Yáñez M, Suárez-Garaboa S, López-Dequidt I, et al. Random forest-based prediction of stroke outcome. ***Sci Rep.*** 2021; 11: 10071.
39. Lee J, Park KM, Park S. Interpretable machine learning for prediction of clinical outcomes in acute ischemic stroke. ***Frontiers in Neurology*** 2023; 7: 1234046.
40. Stinear CM, Smith MC, Byblow WD. Prediction Tools for Stroke Rehabilitation. ***Stroke*** 2019; 50: 3314-3322.