Leveraging Machine Learning for Stroke Prediction: An Empirical Study on Clinical and Behavioral Risk Factors

Jungmin Lim, University of California, Irvine, 260 Aldrich Hall, Irvine, CA 92697
Corresponding author email: Jungminlim1015@gmail.com

Abstract

This study investigates the applications of machine learning techniques for predicting the stroke risks using clinical, behavioral and demographic features. Multiple classification models were evaluated, and the random forest classifier achieved the highest performance, with a recall rate for stroke of 98% and an AUC of 0.98. Feature importance analysis showed that age, average glucose level, and BMI are the most influential predictors. From an operational perspective, integrating predicting modelling into healthcare systems can facilitate early risk detection and support personalized care strategies.

Keywords

Stroke, Healthcare, Machine Learning, Random Forest

Introduction

Based on the World Stroke Organization's estimates, approximately 15 million people suffer from a stroke each year, and about 5 million die from stroke-related reasons (World Health Organization, 2021). Stroke is one of the leading causes of death and disability. It not only affects those who are experiencing it but also their families and wider society (Edmans et al., 2010), and it can occur in anyone at any age (Elloker & Rhoda, 2018). Therefore, a deeper understanding of the stroke mechanisms and establishing effective risk stratification strategies are urgently needed for both primary and secondary prevention. Managing modifiable risk factors could help prevent nearly half of all strokes among individuals at high risk (Brainin et al., 2018).

Machine Learning (ML) has been rapidly developed and implemented across disciplines and is becoming a transformative force in healthcare research and practice (Dritsas & Trigka, 2022). By leveraging algorithms that are capable of capturing complex, and often hidden relationships among diverse clinical, demographic and physiological variables from many origins including patients' history, imaging and biomarkers (Singh et al., 2025), ML enables more accurate prediction and decision-making than traditional statistical methods. One application of ML is in precision medicine, where ML models are developed to identify the most effective treatment strategies based on an individual's unique conditions (Lee et al., 2018). Machine learning applications in healthcare have expanded rapidly. In the context of stroke, ML techniques have been increasingly used to identify and predict risk factors that contribute to stroke occurrence, recurrence, and recovery outcomes. Predictive models using logistic regression, random forests, gradient boosting, and neural networks have shown strong performance in recognizing key determinants such as age, hypertension, diabetes, smoking, and heart disease (Ahammad, 2022; Hassan et al., 2024; Khosla et al., 2010).

Identifying these factors not only facilitates early intervention and resource allocation but also provides data-driven guidance for healthcare systems to design more efficient prevention and management strategies.

Backgrounds

Stroke is influenced by a complex interplay of nonmodifiable and modifiable risk factors. Nonmodifiable factors include age, sex, race–ethnicity, and genetics. The incidence of stroke doubles with every decade after age 55, and while the mean age of ischemic stroke remains around 69 years, recent evidence shows a rise among adults aged 20–54 years, increasing from 12.9% in 1993/1994 to 18.6% in 2005 (George et al., 2011; Kissela et al., 2012; Roger et al., 2012; Van Asch et al., 2010). The relationship between sex and stroke risk is age dependent: younger women face similar or slightly higher risk than men—likely due to pregnancy, hormonal contraception, and postpartum factors—whereas men have a higher risk in later life (Asplund et al., 2009; Kapral et al., 2005; Reeves et al., 2009). Racial and ethnic disparities are also pronounced. Black Americans experience twice the incidence and higher mortality rates than White populations (Cruz-Flores et al., 2011; Gillum, 1999b), with similar elevations reported among Hispanic/Latino and American Indian groups (Kleindorfer et al., 2006). These inequalities are attributed in part to higher prevalence of hypertension, obesity, and diabetes (Giles et al., 1995; Gillum, 1999a), but social determinants such as healthcare access, language, and nativity also play critical roles (Howard et al., 2011; Joubert et al., 2008). Genetic predisposition further contributes to stroke risk: a positive parental or family history increases the likelihood of stroke, with genetic effects varying across age, sex, and ethnicity (Seshadri et al., 2010).

Modifiable risk factors provide key targets for prevention. Hypertension remains the most significant, accounting for roughly 54% of stroke population attributable risk in the INTERSTROKE study (Donnell et al., 2010). The risk of stroke rises progressively with blood pressure, even below the hypertensive threshold (Vasan et al., 2002). Diabetes mellitus doubles stroke risk and accounts for about 20% of deaths in diabetic patients, with longer duration of diabetes further increasing risk (Banerjee et al., 2012; Sui et al., 2011). Other major modifiable factors include atrial fibrillation and atrial cardiopathy (Yiin et al., 2014), dyslipidemia (particularly high total cholesterol and low HDL levels) (Horenstein et al., 2002), sedentary lifestyle, poor diet, obesity, and metabolic syndrome (Zhou et al., 2007), as well as alcohol consumption, illicit drug use, and cigarette smoking—the latter nearly doubling stroke risk and contributing to 15% of stroke deaths annually (Kuo et al., 2013). Emerging evidence also links inflammation, infection, and air pollution exposure to increased stroke incidence (Kaptoge et al., 2010). Together, these epidemiological findings underscore that both traditional and novel risk factors—spanning biological, behavioral, and environmental domains—must be integrated into predictive models and prevention strategies.

Machine learning (ML) has emerged as a powerful analytical framework for predicting stroke risk by leveraging large-scale clinical and behavioral datasets. Numerous studies have applied classical ML algorithms—such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbours (KNN)—to identify individuals at high risk of stroke, achieving accuracies ranging from 82% to 96% (Kokkotis et al., 2022; Sirsat et al., 2020; Wu & Fang, 2020). Among these, tree-based ensemble methods such as RF and Gradient Boosting consistently outperform traditional classifiers in accuracy, precision, and AUC performance. For instance, several studies reported RF models achieving 95–96% accuracy (Geethanjali et al., 2021; Tazin et al., 2021), while NB and LR models achieved competitive but slightly lower results (82–86%) (Geethanjali et al., 2021). These results highlight ML's potential to capture nonlinear relationships between stroke risk and clinical features such as age, hypertension, heart disease, glucose level, and smoking status (Sailasya & Kumari, 2021).

Building on these foundational methods, researchers have explored advanced and hybrid approaches to enhance predictive performance. Shanthi et al.(2009) applied an Artificial Neural Network (ANN) to predict thromboembolic strokes, reaching 89% accuracy. Similarly, Ahmed et al. (2019) achieved 90% accuracy using ML algorithms on the Apache Spark platform, while Tazin et al. (2021) improved accuracy to 95% after applying normalization and feature-ranking procedures. Other hybrid models—such as the Minimal Genetic Folding (MGF) algorithm (Mezher, 2022) and the RXLM ensemble combining RF, XGBoost, and LightGBM (Alruily et al., 2023)—further advanced predictive capacity, achieving 83% and 96.3% accuracy, respectively. To address data imbalance and enhance generalizability, many studies incorporated techniques such as SMOTE oversampling, feature selection, and outlier control (Sowjanya & Mrudula, 2023; Wongvorachan et al., 2023). Some even implemented real-time or cloud-based prediction tools that can collect user data and deliver early stroke warnings with 96% accuracy (Islam et al., 2021).

While progress in predictive modeling is significant, key methodological and practical challenges remain. Many existing studies rely on relatively small or imbalanced datasets, or focus on a limited number of attributes, constraining model robustness (Chen, 2023; Nijman et al., 2022; Paul et al., 2022). Furthermore, the high-performing ensemble and neural network models often function as “black boxes,” limiting interpretability and hindering clinical adoption. Another limitation lies in the lack of standardized evaluation metrics and external validation, which restricts comparability across studies. Therefore, the literature increasingly calls for the development of explainable ML frameworks, integration of diverse clinical and behavioral features, and comprehensive benchmarking on larger datasets. Such efforts are critical to ensure that predictive analytics can move beyond model optimization toward actionable, interpretable tools that support early stroke prevention and healthcare decision- making.

Methods

Our research utilized the publicly available Stroke Prediction Dataset from Kaggle (Stroke Prediction Dataset, 2025). From this dataset, we included only participants having no missing values, resulting in a total sample size of 4909 individuals. The dataset contains 10 predictor variables and one binary outcome variable indicating whether the participant has ever experienced a stroke. The predictors are defined as follows: Age (in years), Gender, Diagnosed hypertension, Heart Disease, Ever Married, Work Type (5 categories: private, self- employed, government job, never worked and children), Residence Type (urban, rural), Average Glucose Level (mg/dL), Body Mass Index (BMI) (kg/m²), and Smoking Status (three categories: currently smokes, never smoked, and formerly smoked). The outcome variable, Stroke, represents whether the participant has previously suffered a stroke. Among these variables, age, average glucose level, and BMI are continuous, while the remaining features are categorical. We normalized the continuous variables and performed one-hot encoding for the categorical variables. To address the class imbalance between stroke and non-stroke cases in subsequent analyses, we applied the Synthetic Minority Oversampling Technique (SMOTE) (Maldonado et al., 2019), which synthetically augments the minority (stroke) class to achieve a balanced dataset for model training.

Machine Learning Models

Random Forest Classifier

The Random Forest (RF) algorithm is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve generalization and classification performance. Given a training dataset:

$$D = \{(x_i, y_i)\}_{i=1}^{n}$$

where $x_i \in \mathbb{R}^p$ denotes the feature vector with p predictors and $y_i \in \{0,1\}$ represents the binary class label (stroke or non-stroke). The RF algorithm performs the following steps:

1. Bootstrap sampling: Draw T bootstrap samples from the training set.
2. Tree growth: For each sample, grow an unpruned classification tree. At each node, a random subset of m < p features is selected, and the best split among these m features is chosen to minimize impurity (e.g., Gini index or entropy).
3. Aggregation: Each tree $h_t(x)$ provides a class prediction. The final prediction of the forest is obtained by majority voting across all trees:

$$\hat{h}_{RF(x)} = mode\{h_{t(x)}\}_{t=1}^{T}$$

4. This ensemble approach reduces variance and mitigates overfitting by combining multiple decorrelated classifiers.

k-Nearest Neighbor (kNN) Classifier

The k-Nearest Neighbor (kNN) algorithm is a non-parametric, instance-based learning method that classifies a new observation based on the majority label among its nearest neighbors in the training set. For any two data points x_i and x_j, the distance function is defined as:

$$d(x_i, x_j) = ||x_i - x_j|| = \sqrt{\Sigma_{p=1}^{P}(x_{ip} - x_{jp})^2}$$

Given a new observation x_0, the classifier identifies its k nearest neighbors, denoted $N_k(x_0)$, and assigns the most frequent class label among them:

$$\hat{h}_{kNN(x^0)} = mode\{y_i : x_i \in N_{k(x^0)}\}$$

For binary classification problems where $y_i \in \{-1, +1\}$, the decision rule can equivalently be written as:

$$\hat{h}_{kNN(x^0)} = sign\left(\left(\frac{1}{k}\right)\Sigma_{x_i \in N_{k(x^0)}} y_i\right)$$

The hyperparameter k controls the bias–variance trade-off: smaller k values lead to lower bias but higher variance, while larger k values produce smoother decision boundaries with higher bias.

Logistic Regression

The Logistic Regression (LR) algorithm is a statistical learning method used for binary classification problems. It models the conditional probability of the dependent variable $y_i \in \{0,1\}$ given the predictors $x_i \in \mathbb{R}^p$ using the logistic (sigmoid) function. The model assumes a linear relationship between the predictors and the log-odds of the probability of the positive class. The logistic regression function is defined as:

$$P(x_i) = \frac{1}{\left(1 + e^{\{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})\}}\right)}$$

The logistic regression model estimates the coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ by maximizing the log-likelihood function $\ell(\beta)$:

$$l(\beta) = \Sigma_{\{i=1\}}^{\{n\}}[y_i \, log \, log \, (\pi_i) \, + (1 \, - \, y_i) \, log \, log \, (1 \, - \, \pi_i) \,], \;\; \pi_i = \, P(\, x_i)$$

The fitted probabilities can then be used for classification, where an observation is predicted as stroke-positive if the estimated probability exceeds 0.5. This model provides a simple, interpretable baseline for binary classification, assuming a linear relationship between predictors and the log-odds of the outcome.

Model Evaluation Metrics

Under the evaluation process of the considered machine learning (ML) models, several performance metrics were recorded. In the current analysis, we focus on the most widely used measures in related literature (Hossin & Sulaiman, 2015):

$$Recall \; = \frac{TP}{TP + FN}$$
$$Precision \; = \frac{TP}{TP + FP}$$
$$F1 \; = \; 2 \, \frac{(Precision \; * \; Recall)}{Precision + Recall}$$
$$Accuracy \; = \frac{(TP \; + \; TN)}{(TP + TN + FP + FN)}$$
$$AUC \; = \; \int_0^1 TPR(FPR) \, d(FPR)$$

Here, True Positives (TP) represent the number of participants who experienced a stroke and were correctly identified by the model as stroke cases. True Negatives (TN) denote the number of participants who did not experience a stroke and were correctly predicted as non-stroke cases. False Positives (FP) correspond to the number of participants who were incorrectly classified as having a stroke when they actually did not. And False Negatives (FN) refer to the participants who had a stroke but were mistakenly predicted as non-stroke.

From these quantities, we can derive two rates. True Positive Rate (TPR), also known as Recall or Sensitivity, quantifies the model's ability to correctly identify stroke cases and is computed as

$$TPR \; = \frac{TP}{TP + FN}$$

False Positive Rate (FPR) measures the proportion of non-stroke participants incorrectly classified as stroke and is defined as

$$FPR \; = \frac{FP}{FP + TN}$$

TPR and FPR describe the trade-off between sensitivity and specificity across different classification thresholds. They are also used to construct the Receiver Operating Characteristic (ROC) curve, from which the Area Under the Curve (AUC) metric is derived, a higher AUC value indicates better discriminative performance of the model.

Results

Table 1 presents the baseline characteristics of participants according to stroke status. Significant differences were observed between stroke and non-stroke groups in most variables. Participants who had experienced a stroke were notably older and had higher

average glucose levels and BMI values compared to those without stroke (all $p < 0.001$). A higher prevalence of hypertension and heart disease was also observed among stroke patients. Moreover, individuals with a history of stroke were more likely to be married, self- employed, or engaged in private-sector work, while the distribution of gender, residence type, and smoking status showed smaller differences.

Table 1. Descriptive Analysis

Variable	Non-stroke (mean(sd))/ %	Stroke (mean(sd))/ %	p-value
Age	41.76 ± 22.27	67.71 ± 12.40	<0.001***
Glucose Level	104.00 ± 43.00	134.57 ± 62.46	<0.001***
BMI	28.82 ± 7.91	30.47 ± 6.33	<0.001***
Gender			
Female	58.1%	57.4%	0.870
Male	40.9%	42.6%	
Hypertension			
No	91.7%	71.3%	<0.001***
Yes	8.3%	28.7%	
Heart Disease			
No	95.7%	80.9%	<0.001***
Yes	4.3%	19.1%	
Marriage			
No	35.8%	11.0%	<0.001***
Yes	64.2%	89.0%	
Work Type			
Government job	12.8%	13.4%	<0.001***
Private	57.1%	60.8%	
Self-employed	15.4%	25.4%	
Never worked	0.5%	0%	
Children	14.3%	0.5%	
Residence Type			
Rural	49.3%	47.8%	0.725
Urban	50.7%	52.2%	
Smoking Status			
Formerly smoked	16.6%	27.3%	<0.001***
Never smoked	37.6%	40.2%	
Smokes	14.9%	18.7%	
Unknown	30.9%	13.9%	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

When evaluating the model performance, we prioritize the recall rate, which is essential in stroke prediction to minimize the false negative rate. According to the results in Table 2, we focused on the predictive performances for the storke class. Among the three models, the Random Forest classifier achieved the highest overall performance, with the highest recall rate

(98%) and an AUC of approximately 0.98, indicating excellent predictive capability of stroke and strong overall discriminative ability. The k-Nearest Neighbor (kNN) model also showed excellent predictive performance (Recall= 96%), and a comparable AUC (0.96), making it an efficient yet powerful non-parametric alternative. In contrast, the Logistic Regression model demonstrated lower predictive ability (Recall = 83%) and AUC (0.86), suggesting that its linear decision boundary is less efficient at identifying the true stroke cases in data characterized by nonlinear and interacting predictors.

Table 2: Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-score	AUC (ROC)
Random Forest	0.94	0.93	0.98	0.94	0.98
k-Nearest Neighbors	0.93	0.88	0.96	0.92	0.96
Logistic Regression	0.79	0.77	0.83	0.79	0.86

Table 3 shows the feature importance analysis from the optimized Random Forest model indicates that age is the most dominant predictor of stroke, contributing nearly 35% of the total importance. The next most influential variables are average glucose level and BMI, which reflect metabolic health, suggesting that elevated blood glucose and higher body mass index are critical physiological indicators associated with stroke occurrence. Sociodemographic variables such as marital status (not married), work type (self-employed or government job), and residential area (rural) also show meaningful contributions, capturing lifestyle and environmental effects. Overall, these findings emphasize that both biological (age, glucose, BMI) and behavioral/lifestyle factors jointly influence stroke risk.

Table 3. Feature Importance Analysis

Rank	Feature	Importance
1	Age	0.348
2	Average Glucose Level	0.163
3	BMI	0.138
4	Ever Married = No	0.066
5	Hypertension	0.046
6	Residence Type = Rural	0.034
7	Gender = Female	0.034
8	Work Type = Self-employed	0.030
9	Smoking Status = Never Smoked	0.027
10	Work Type = Govt Job	0.026

Conclusion

This study examined how machine learning can be applied to predict stroke risk by analyzing clinical and behavioral factors, while also exploring its implications for healthcare management and business analytics. Among the models tested, the Random Forest classifier achieved the best performance, with an accuracy of 94% and an AUC of 0.98, demonstrating strong predictive power in identifying individuals at high risk. Feature importance analysis indicated that age, average glucose level, and BMI were the most influential predictors, followed by marital status, hypertension, and work type. These findings suggest that both physiological and lifestyle-related factors contribute meaningfully to stroke prediction, aligning with previous research (Dubow et al., 2011; Rexrode et al., 2022).

When doing the prediction, the random forest performed the best out of all three models. The performance is likely related to the algorithm's ensemble structure. By aggregating the predictions from many decorrelated decision trees built on bootstrap samples and the random subsets of predictors, random forest can approximate complex non-linear and high-order interactions without requiring a prespecified functional form (Breiman, 2001). Moreover, because each tree uses threshold-based splits on the predictor values, the model depends mainly on the ordering rather than the exact magnitude of the observations, which makes it less sensitive to extreme values. These properties are valuable when modelling the heterogeneous clinical data, where the relationships between risk factors and stroke are unlikely to be linear and measurement error and outliers are common.

From a business perspective, integrating predictive models into healthcare operations offers substantial economic and strategic value. Early identification of high-risk individuals enables hospitals, insurance providers, and digital health companies to implement preventive interventions, optimize resource allocation, and reduce treatment costs. Predictive analytics thus provide a foundation for data-driven decision-making and the development of personalized healthcare services.

Despite those promising results, some challenges remain regarding data interpretability, standardization, and privacy protection. First, machine learning models often suffer from limited interpretability, making it difficult for clinicians to understand how some features contribute to an individual patient's risk, which will hinder the clinical implementation and accountability. Second, several ethical and operational challenges have to be considered. In our analysis, the data originates from different hospitals, making the measurement might differ across the data. Representativeness is another concern, as models trained on a dataset that mainly consist older adults, it could be biased when applied to younger populations.

Third, privacy also is an important challenge. When using sensitive healthcare information for model development, it requires compliance with data protection regulations and secure data storage. Future research should focus on developing explainable and scalable ML frameworks and incorporating broader datasets that include diverse populations, and more behavioral and socioeconomic dimensions. Overall, this study demonstrates that leveraging machine learning for stroke prediction holds both clinical benefits and business potential, advancing efficiency and innovation in the healthcare industry.

References

Ahammad, T. (2022). Risk factor identification for stroke prognosis using machine-learning algorithms. *Jordanian Journal of Computers and Information Technology*, *8*(3).

Ali, A. A. (2019). Stroke prediction using distributed machine learning based on apache spark. *Stroke*, *28*(15), 89–97.

Alruily, M., El-Ghany, S. A., Mostafa, A. M., Ezz, M., & El-Aziz, A. A. (2023). A-tuning ensemble machine learning technique for cerebral stroke prediction. *Applied Sciences*, *13*(8), 5047.

Asplund, K., Karvanen, J., Giampaoli, S., Jousilahti, P., Niemelä, M., Broda, G., Cesana, G., Dallongeville, J., Ducimetriere, P., & Evans, A. (2009). Relative risks for stroke by age, sex, and population based on follow-up of 18 European populations in the MORGAM Project. Stroke, 40(7), 2319–2326.

Banerjee, C., Moon, Y. P., Paik, M. C., Rundek, T., Mora-McLaughlin, C., Vieira, J. R., Sacco, R. L., & Elkind, M. S. (2012). Duration of diabetes and risk of ischemic stroke: The Northern Manhattan Study. Stroke, 43(5), 1212–1217.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Chen, Z. (2023). Stroke risk prediction based on machine learning algorithms. Highlights Sci. Eng. Technol, 38, 932–941.

Cruz-Flores, S., Rabinstein, A., Biller, J., Elkind, M. S., Griffith, P., Gorelick, P. B., Howard, G., Leira, E. C., Morgenstern, L. B., & Ovbiagele, B. (2011). Racial-ethnic disparities in stroke care: The American experience: A statement for healthcare professionals from

the American Heart Association/American Stroke Association. *Stroke*, *42*(7), 2091–2116.

Donnell, M. J., Xavier, D., Liu, L., Zhang, H., Chin, S., Rao-Melacini, P., Rangarajan, S., Islam, S., Ardila, S., & Foscal, L. (2010). Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): A case-control study. Lancet, 376(9735), 112–123.

Dubow, J., & Fink, M. E. (2011). Impact of hypertension on stroke. Current atherosclerosis reports, 13(4), 298-305.

Geethanjali, T., Divyashree, M., Monisha, S., & Sahana, M. (2021). Stroke prediction using machine learning. Journal of Emerging Technologies and Innovative Research, 9(6), 710–717.

George, M. G., Tong, X., Kuklina, E. V., & Labarthe, D. R. (2011). Trends in stroke hospitalizations and associated risk factors among children and young adults, 1995–2008. Annals of Neurology, 70(5), 713–721.

Giles, W. H., Kittner, S. J., Hebel, J. R., Losonczy, K. G., & Sherwin, R. W. (1995). Determinants of black-white differences in the risk of cerebral infarction: The National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. Archives of Internal Medicine, 155(12), 1319–1324.

Gillum, R. F. (1999a). Risk factors for stroke in blacks: A critical review. American Journal of Epidemiology, 150(12), 1266–1274.

Gillum, R. F. (1999b). Stroke mortality in blacks: Disturbing trends. Stroke, 30(8), 1711–1715.

Guhdar, M., Melhum, A. I., & Ibrahim, A. L. (2023). Optimizing accuracy of stroke prediction using logistic regression. Journal of Technology and Informatics (JoTI), 4(2), 41–47.

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.

Horenstein, R. B., Smith, D. E., & Mosca, L. (2002). Cholesterol predicts stroke mortality in the Women's Pooling Project. Stroke, 33(7), 1863–1868.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1.

Howard, V. J., Kleindorfer, D. O., Judd, S. E., McClure, L. A., Safford, M. M., Rhodes, J. D., Cushman, M., Moy, C. S., Soliman, E. Z., & Kissela, B. M. (2011). Disparities in stroke incidence contributing to disparities in stroke mortality. Annals of Neurology, 69(4), 619–627.

Islam, M. M., Akter, S., Rokunojjaman, M., Rony, J. H., Amin, A., & Kar, S. (2021). Stroke prediction analysis using machine learning classifiers and feature technique. International Journal of Electronics and Communications Systems, 1(2), 17–22.

Joubert, J., Prentice, L. F., Moulin, T., Liaw, S.-T., Joubert, L. B., Preux, P.-M., Ware, D., Medeiros de Bustos, E., & McLean, A. (2008). Stroke in rural areas and small communities. Stroke, 39(6), 1920–1928.

Kapral, M. K., Fang, J., Hill, M. D., Silver, F., Richards, J., Jaigobin, C., & Cheung, A. M. (2005). Sex differences in stroke care and outcomes: Results from the Registry of the Canadian Stroke Network. Stroke, 36(4), 809–814.

Kaptoge, S., Di Angelantonio, E., Lowe, G., Pepys, M., Thompson, S., Collins, R., & Danesh, J. (2010). Emerging Risk Factors Collaboration C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: An individual participant meta-analysis. Lancet, 375(9709), 132–140.

Khosla, A., Cao, Y., Lin, C. C.-Y., Chiu, H.-K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to stroke prediction. 183–192.

Kissela, B. M., Khoury, J. C., Alwell, K., Moomaw, C. J., Woo, D., Adeoye, O., Flaherty, M. L., Khatri, P., Ferioli, S., & De Los Rios La Rosa, F. (2012). Age at stroke: Temporal trends in stroke incidence in a large, biracial population. Neurology, 79(17), 1781–1787.

Kleindorfer, D., Broderick, J., Khoury, J., Flaherty, M., Woo, D., Alwell, K., Moomaw, C. J., Schneider, A., Miller, R., & Shukla, R. (2006). The unchanging incidence and case-fatality of stroke in the 1990s: A population-based study. Stroke, 37(10), 2473–2478.

Kokkotis, C., Giarmatzis, G., Giannakou, E., Moustakidis, S., Tsatalas, T., Tsiptsios, D., Vadikolias, K., & Aggelousis, N. (2022). An explainable machine learning pipeline for stroke prediction on imbalanced data. Diagnostics, 12(10), 2392.

Kuo, S.-H., Lee, Y.-T., Li, C.-R., Tseng, C.-J., Chao, W.-N., Wang, P.-H., Wong, R.-H., Chen, C.-C., Chen, S.-C., & Lee, M.-C. (2013). Mortality in Emergency Department Sepsis score as a prognostic indicator in patients with pyogenic liver abscess. The American Journal of Emergency Medicine, 31(6), 916–921.

Maldonado, S., López, J., & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. Applied Soft Computing, 76, 380–389.

Mezher, M. A. (2022). Genetic folding (GF) algorithm with minimal kernel operators to predict stroke patients. Applied Artificial Intelligence, 36(1), 2151179.

Nijman, S. W., Leeuwenberg, A., Beekers, I., Verkouter, I., Jacobs, J., Bots, M., Asselbergs, F., Moons, K. G., & Debray, T. P. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. Journal of Clinical Epidemiology, 142, 218–229.

Paul, D., Gain, G., Orang, S., Das, P., & Chaudhuri, A. K. (2022). Advanced random forest ensemble for stroke prediction. Training, 66, 34.

Reeves, M. J., Fonarow, G. C., Zhao, X., Smith, E. E., & Schwamm, L. H. (2009). Quality of care in women with ischemic stroke in the GWTG program. Stroke, 40(4), 1127–1133.

Rexrode, K. M., Madsen, T. E., Yu, A. Y., Carcel, C., Lichtman, J. H., & Miller, E. C. (2022). The impact of sex and gender on stroke. Circulation research, 130(4), 512-528.

Roger, V., Go, A., Lloyd-Jones, D., Benjamin, E., Berry, J., Borden, W., Bravata, D., Dai, S., Ford, E., & Fox, C. (2012). American Heart Association Statistics Committee and Stroke Statistics Subcommittee Executive summary: Heart disease and stroke statistics—2012 update: A report from the American Heart Association. Circulation, 125(1), 188–197.

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. *International Journal of Advanced Computer Science and Applications*, *12*(6).

Seshadri, S., Beiser, A., Pikula, A., Himali, J. J., Kelly-Hayes, M., Debette, S., DeStefano, A. L., Romero, J. R., Kase, C. S., & Wolf, P. A. (2010). Parental occurrence of stroke and risk of stroke in their children: The Framingham study. *Circulation*, *121*(11), 1304–1312.

Shanthi, D., Sahoo, G., & Saravanan, N. (2009). Designing an artificial neural network model for the prediction of thrombo-embolic stroke. *International Journals of Biometric and Bioinformatics (IJBB)*, *3*(1), 10–18.

Sirsat, M. S., Fermé, E., & Câmara, J. (2020). Machine learning for brain stroke: A review. *Journal of Stroke and Cerebrovascular Diseases*, *29*(10), 105162.

Sowjanya, A. M., & Mrudula, O. (2023). Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. *Applied Nanoscience*, *13*(3), 1829–1840.

Stroke Prediction Dataset. (2025). [Dataset]. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Sui, X., Lavie, C. J., Hooker, S. P., Lee, D.-C., Colabianchi, N., Lee, C.-D., & Blair, S. N. (2011). *A prospective study of fasting plasma glucose and risk of stroke in asymptomatic men*. *86*(11), 1042–1049.

Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Monirujjaman Khan, M. (2021). Stroke disease detection and prediction using robust learning approaches. *Journal of Healthcare Engineering*, *2021*(1), 7633381.

Van Asch, C. J., Luitse, M. J., Rinkel, G. J., van der Tweel, I., Algra, A., & Klijn, C. J. (2010). Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: A systematic review and meta-analysis. *The Lancet Neurology*, *9*(2), 167–176.

Vasan, R. S., Beiser, A., Seshadri, S., Larson, M. G., Kannel, W. B., D'Agostino, R. B., & Levy, D. (2002). Residual lifetime risk for developing hypertension in middle-aged women and men: The Framingham Heart Study. *Jama*, *287*(8), 1003–1010.

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, *14*(1), 54.

Wu, Y., & Fang, Y. (2020). Stroke prediction with machine learning methods among older Chinese. *International Journal of Environmental Research and Public Health*, *17*(6), 1828.

Yiin, G. S., Howard, D. P., Paul, N. L., Li, L., Luengo-Fernandez, R., Bull, L. M., Welch, S. J., Gutnikov, S. A., Mehta, Z., & Rothwell, P. M. (2014). Age-specific incidence, outcome, cost, and projected future burden of atrial fibrillation–related embolic vascular events: A population-based study. *Circulation*, *130*(15), 1236–1244.

Zhou, M., Zhu, L., Wang, J., Hang, C., & Shi, J. (2007). The inflammation in the gut after experimental subarachnoid hemorrhage. *Journal of Surgical Research*, *137*(1), 103–108.