# Early Stroke Risk Prediction Using Logistic Regression: A Data-Machine Learning Approach

**[1]Pragatheeswari M*, [2]Hari Preetha T*, [3]Trishitha K, [4]Gopal samy B**

[1,2,3,4] Department of Biotechnology, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India.

[1,2,3]UG Scholars, [4]Professor

[1]mpragatheeswarikavi@gmail.com, [2]haripreethathirumoorthi@gmail.com, [3]kathirvelktrishitha@gmail.com, [4]gopalsamy2k6@gmail.com

***Corresponding author:**

Pragatheeswari M

Email.id:mpragatheeswarikavi@gmail.com

***Abstract-Stroke remains one of the leading causes of mortality and long-term neurological disability worldwide. Timely and accurate prediction of stroke risk is essential for implementing preventive strategies and optimizing patient care. In this study, a logistic regression model was developed to predict stroke occurrence using a comprehensive dataset of 5,110 adults aged 18 and above, sourced from a publicly available healthcare database. The dataset includes demographic, physiological, and lifestyle features such as age, gender, hypertension, heart disease, marital status, work type, residence, smoking status, body mass index (BMI), and average blood glucose level. The data underwent preprocessing, including imputation of missing values, normalization of continuous variables, and encoding of categorical features. Logistic regression was selected for its simplicity, interpretability, and effectiveness in binary clinical classification tasks. The model's performance was evaluated using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC).The model achieved an overall accuracy of 74.64% and a high recall of 85% for stroke prediction, indicating strong sensitivity in identifying true positive cases. Although the precision was relatively low due to class imbalance, the model effectively minimized false negatives an essential consideration in medical diagnostics. These results support logistic regression as a reliable baseline for early stroke risk detection and suggest that performance can be further enhanced through techniques such as resampling or ensemble learning. This study contributes to the development of accessible, data-driven tools for proactive stroke prevention and clinical decision support.***

***Keywords- Stroke prediction, logistic regression, machine learning, health data analytics, imbalanced dataset, risk classification, medical diagnosis, recall, confusion matrix, preventive healthcare.***

**Abbreviations:**

**BMI-** Body Mass index, **ROC-** Receiver Operating Characteristic, AUC- Area Under the Curve, **ML-** Machine Learning, **WHO-** World Health Organization , **SMOTE-** Synthetic Minority Oversampling Technique.

## I. INTRODUCTION

Stroke is a life-threatening medical condition that occurs when blood flow to a part of the brain is obstructed, resulting in neuronal damage, functional impairment, and often long-term disability. Globally, stroke is the second leading cause of death and a major contributor to permanent neurological disability, especially among older adults (World Health Organization [WHO], 2023). The socioeconomic burden of stroke includes direct medical costs, rehabilitation expenses, and a significant reduction in the quality of life for both patients and caregivers. Despite advancements in acute stroke treatment, prevention remains the most effective strategy to reduce its incidence and associated costs (Feigin et al., 2017).

Early identification of individuals at risk for stroke is crucial for implementing preventive measures such as lifestyle modifications, medical therapy, or monitoring. Conventional risk prediction models, such as the Framingham Stroke Risk Profile, rely on a set of predefined clinical variables including age, blood pressure, diabetes, and smoking habits (Goldstein et al., 2006). However, these models often fail to capture non-linear relationships and interactions between variables, which are common in complex diseases like stroke. With the availability of large-scale health data and computational advancements, machine learning (ML) techniques offer more sophisticated, data-driven approaches to improve risk stratification and prediction accuracy (Shickel et al., 2018).

Among various ML models, logistic regression stands out for its balance of simplicity, interpretability, and effectiveness in binary classification problems. It estimates the probability of an event (such as stroke occurrence) by fitting a logistic function to the data, providing not only classification but also insight into the influence of each input feature (Hosmer et al., 2013). In clinical settings where transparency and accountability are critical, logistic regression offers a practical advantage over black-box models like deep

neural networks. Furthermore, logistic regression models are easily deployable in resource-constrained environments due to their low computational overhead (Zhou et al., 2021).

This study proposes a logistic regression-based approach for stroke risk prediction using a well-structured dataset of 5,110 individuals aged 18 and above. The dataset includes demographic, physiological, and lifestyle features such as age, gender, hypertension, heart disease, BMI, average glucose level, and smoking status. The objective is to train and evaluate the model on key performance metrics-accuracy, precision, recall, F1-score, and ROC-AUC-to determine its applicability for early detection and clinical decision-making. By using interpretable and accessible modeling techniques, this work aims to contribute toward the development of cost-effective tools for proactive stroke prevention in both urban and rural healthcare settings.

## II. DATASET DESCRIPTION

This study employed a publicly available dataset from Kaggle containing health information for 5,110 individuals aged 18 years and above (Kaggle, 2020). The dataset was curated for stroke prediction using machine learning algorithms and includes demographic, physiological, and lifestyle features relevant to stroke risk.

The features in the dataset are as follows:

- **Age**: All participants were aged 18 or older.
- **Gender**: Out of the total, 2,994 were female and 2,115 were male.
- **Hypertension**: A binary variable indicating whether a participant had been diagnosed with high blood pressure. Approximately 9.75% of the participants were hypertensive.
- **Heart Disease**: Indicates the presence of any heart-related conditions. Around 4.03% reported having heart disease.
- **Ever Married**: Represents the marital status of the individuals; 65.62% were married.
- **Work Type**: Categorized into never worked (0.43%), government job (12.86%), self-employed (16.03%), and private sector (57.24%).
- **Residence Type**: Denotes rural (49.20%) or urban (50.80%) living environments.
- **Average Glucose Level**: Continuous variable measured in mg/dL.
- **Body Mass Index (BMI)**: A numeric feature expressed in kg/m².
- **Smoking Status**: Divided into formerly smoked (17.32%), never smoked (37.03%), and currently smoking (15.44%).
- **Stroke**: Binary target variable indicating whether the participant had previously experienced a stroke. Only 5.53% of the dataset represented positive stroke cases.

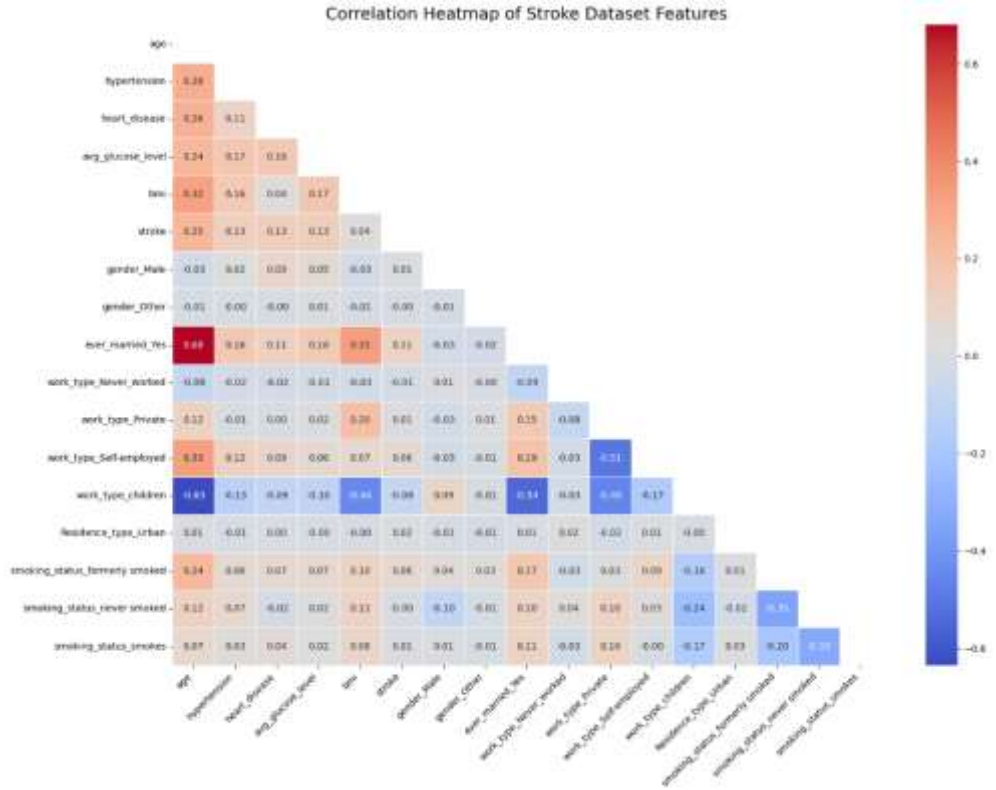


*Figure 1. Correlation heatmap showing relationships among dataset features.*

## III. METHODOLOGY

This section outlines the step-by-step process used to develop a predictive model for stroke classification using logistic regression. The workflow includes data preprocessing, splitting the dataset, model implementation, handling class imbalance, and evaluating model performance.

### 3.1 Data Preprocessing:

Preprocessing is an essential phase in machine learning workflows to guarantee that the data is clean, uniform, and appropriate for modeling. A number of actions were taken to ready the dataset:

- **Missing Values**: The dataset was examined for null entries, especially in the bmi feature. Missing values were imputed using the **median** to minimize distortion caused by outliers (Kotsiantis et al., 2006).
- **Categorical Encoding**: Variables such as gender, work_type, residence_type, and smoking_status were converted into numerical form using **one-hot encoding** to make them compatible with the logistic regression model (Zhang et al., 2021).
- **Feature Scaling**: Continuous variables like age, avg_glucose_level, and bmi were standardized using **z-score normalization** to bring them onto a common scale. This improves optimization performance and prevents dominance of features with larger scales (James et al., 2013).
- **Label Encoding**: The target variable stroke was encoded into a binary format: 0 indicating absence and 1 indicating presence of stroke.

### 3.2 Train-Test Split:

To evaluate the generalizability of the logistic regression model, the dataset was partitioned into training and testing subsets using the train_test_split() function from the scikit-learn library (Pedregosa et al., 2011). An 80:20 split ratio was implemented, where 80% of the data was used to train the model and learn the underlying patterns, while the remaining 20% served as a test set to evaluate the model's performance on previously unseen data. This division made sure the evaluation metrics showed how well the model could work with new data, which helped lower the chance of overfitting.

### 3.3 Logistic Regression Model:

Logistic regression is a type of classifier that works with straight lines to predict the chance that an input falls into a specific group. It uses a special curve called the sigmoid function to turn the total of multiplied input values into a number that shows how likely something is, ranging from 0 to 1.. The sigmoid function is defined as:

$$P(y=1) = \frac{1}{1+e^{-z}} where\ z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

This probability can be thresholded (commonly at 0.5) to classify an instance as stroke or non-stroke. The model was implemented using Python's LogisticRegression() function from the sklearn.linear_model module. The solver was set to 'liblinear', which works well for smaller datasets and can handle both L1 and L2 types of regularization. This choice ensures computational efficiency and reliable convergence during model training.

### 3.4 Addressing Class Imbalance:

The dataset exhibited significant class imbalance, with stroke cases accounting for only 5.53% of the total records. Such imbalance can lead to biased models that disproportionately favor the majority class (non-stroke), thereby reducing sensitivity to the minority class (stroke). To solve this, the logistic regression model was trained with the class_weight parameter set to 'balanced'. This setting automatically changes how much importance is given to each class, based on how often they appear in the data. As a result, the model places greater emphasis on learning patterns associated with the minority class without needing external resampling techniques.

### 3.5 Model Evaluation Metrics:

The performance of the logistic regression model was assessed using several well-established metrics in binary classification, particularly suited for medical diagnosis tasks. Although overall accuracy-defined as the proportion of correct predictions was reported, it is not sufficient on its own due to the class imbalance in the dataset. So, more measurements were looked at to give a better overall assessment. . Precision was used to quantify the proportion of true stroke cases among all instances predicted as stroke, while recall (or sensitivity) measured the model's ability to correctly identify actual stroke cases. In clinical applications, high recall is especially important to minimize the risk of missing true positive cases. The F1-score, which is the harmonic mean of precision and recall, offered a balanced metric that accounted for both false positives and false negatives. Finally, the area under the Receiver Operating Characteristic curve (ROC-AUC) was used to evaluate the model's overall discriminative ability across various classification thresholds. A higher AUC indicates better performance in distinguishing between stroke and non-stroke cases, making it a valuable metric in imbalanced classification settings.

## IV. RESULTS

This section presents the outcomes of the logistic regression model evaluated on the test dataset and interprets the results in terms of classification performance, confusion matrix analysis, and clinical relevance.

### 4.1 Model Accuracy and Classification Report:

The logistic regression model achieved an overall **accuracy of 74.64%** on the test set. Detailed performance metrics are shown below:

Table 1. Classification metrics of the logistic regression model

| *Class* | *Precision* | *Recall* | *F1-Score* | *Support* |
|---|---|---|---|---|
| ***0 (No Stroke)*** | 0.99 | 0.74 | 0.85 | 929 |
| ***1 (Stroke)*** | 0.16 | 0.85 | 0.27 | 53 |
| ***Accuracy*** | | | **0.75** | 982 |
| ***Macro Average*** | 0.57 | 0.79 | 0.56 | 982 |
| ***Weighted Avg*** | 0.94 | 0.75 | 0.82 | 982 |

**Observations:**

- **Recall for stroke (1)** is **very high (0.85)**, which means that the model successfully identified the majority of actual stroke cases.
- **Precision for stroke** is **low (0.16)**, indicating a high false positive rate-the model predicts many non-stroke cases as stroke.
- **F1-score for stroke** is **0.27**, reflecting the imbalance between precision and recall.

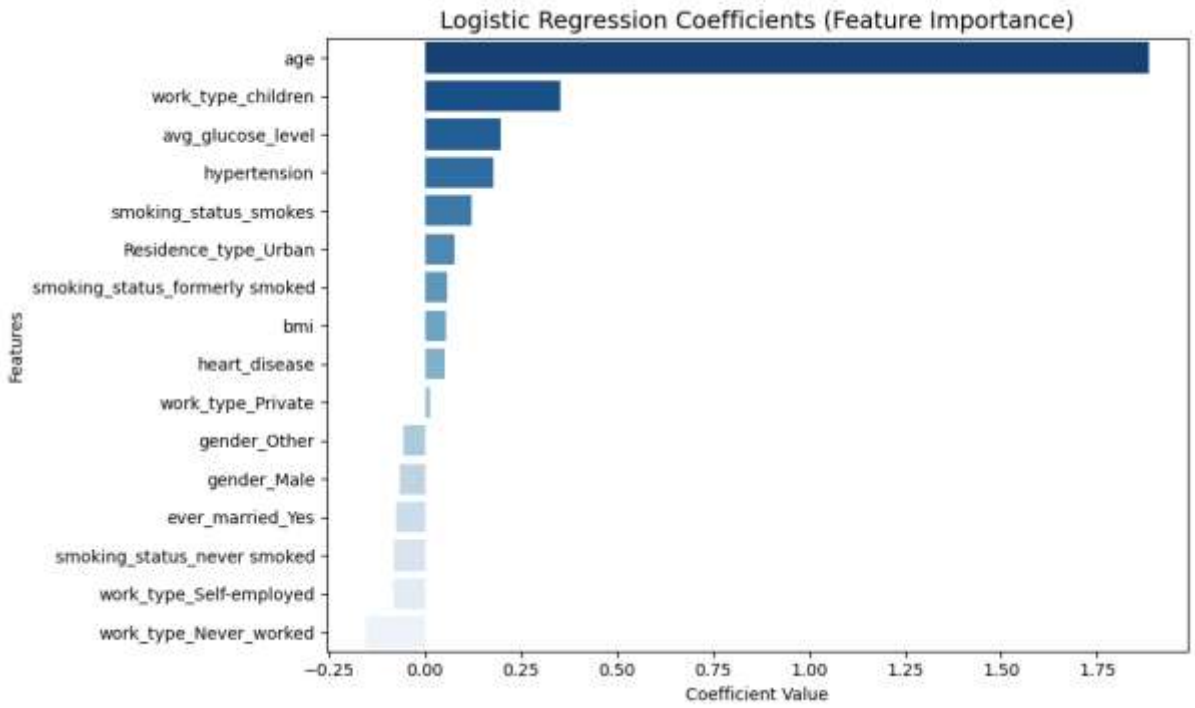


*Figure 2. Logistic regression model coefficients indicating the relative importance of features. Age, work type (children), and glucose level contribute most to stroke prediction.*

### 4.2 Confusion Matrix Analysis:

This section presents the confusion matrix to illustrate the model's classification results in terms of true positives, true negatives, false positives, and false negatives. It provides a clear visualization of how the model performs on both stroke and non-stroke classes.

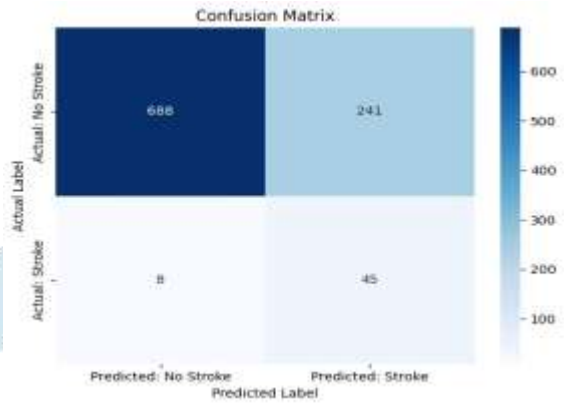


*Figure 3. Confusion matrix showing the performance of the logistic regression model in predicting stroke and non-stroke cases.*

Table 2. Confusion matrix of the logistic regression model

| | ***Predicted: No Stroke*** | ***Predicted: Stroke*** |
|---|---|---|
| ***Actual: No Stroke (0)*** | 688 | 241 |
| ***Actual: Stroke (1)*** | 8 | 45 |

- **True Negatives (688)**: Correctly identified non-stroke cases.
- **True Positives (45)**: Correctly identified stroke cases.
- **False Negatives (8)**: Stroke cases missed by the model.
- **False Positives (241)**: Non-stroke individuals incorrectly flagged as stroke-positive.

The model demonstrates a strong ability to **detect strokes (high sensitivity)** but at the cost of **a large number of false alarms** (low specificity for stroke).

### 4.3 Predicted Probability Distribution:

To further understand how the model assigns prediction probabilities across classes, a probability distribution plot was generated. This visualizes the spread of predicted stroke probabilities for both stroke and non-stroke cases, offering insight into why the model performs well in terms of recall but poorly in precision.

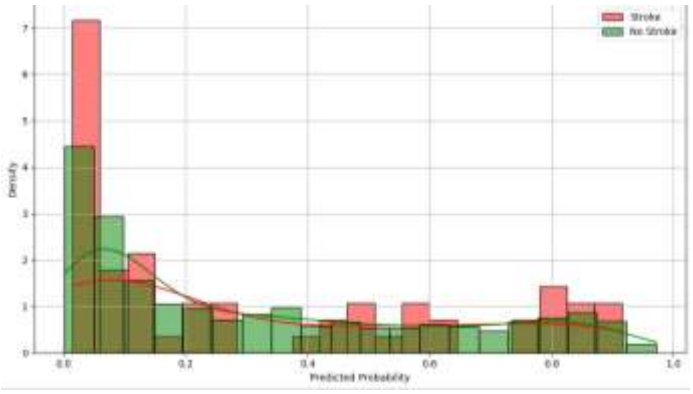


*Figure 4. Probability distribution of predicted stroke risk. The model assigns higher probabilities to stroke cases, although some overlap remains due to class imbalance.*

### 4.4 ROC Curve and Model Discrimination:

In addition to the confusion matrix and classification metrics, the Receiver Operating Characteristic (ROC) curve was used to assess the model's ability to distinguish between stroke and non-stroke cases across all classification thresholds. The ROC curve and the associated Area Under Curve (AUC) provide a comprehensive view of the model's overall discriminative power.

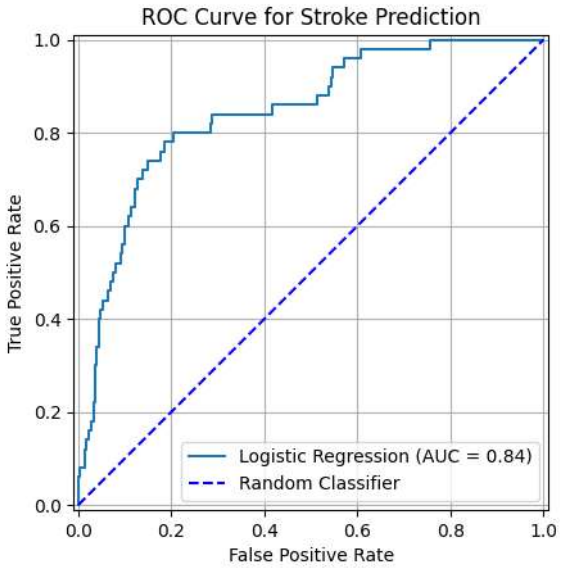


*Figure 5. ROC curve showing model performance in distinguishing stroke vs. non-stroke cases. The AUC of 0.84 indicates strong discriminatory power.*

## V. DISCUSSION

The findings of this study demonstrate that logistic regression is a viable and interpretable model for predicting stroke risk based on routinely collected clinical and demographic features. The model achieved an overall accuracy of 74.64%, but more notably, it attained a high recall of 85%, indicating a strong ability to correctly identify individuals who have experienced a stroke. This high sensitivity is very important in making medical decisions, because not finding a positive case could lead to serious problems. However, the model's precision was relatively low at 16%, reflecting a high number of false positives. While this may result in unnecessary follow-up procedures for some patients, such false alarms are generally considered more acceptable than false negatives in preventive medicine, particularly for life-threatening conditions such as stroke.

Further insight is provided by the probability distribution plot, which shows that stroke cases tend to receive higher predicted probabilities, with many clustering above 0.8. Nonetheless, some overlap with non-stroke cases in the mid-probability range (0.3–0.7) contributes to the lower precision. This trade-off between sensitivity and specificity is also evident in the confusion matrix, which reveals that while the model correctly identified most stroke cases, it also misclassified a significant number of non-stroke individuals. Despite this, the ROC curve yielded an AUC of 0.84, demonstrating that the model performs well overall in distinguishing between stroke and non-stroke cases across all classification thresholds.

The imbalance in the dataset, where only 5.53% of cases are stroke-positive, played a significant role in this precision-recall trade-off. Although internal weighting through the class_weight='balanced' parameter helped mitigate this imbalance, logistic regression may still fall short in capturing more complex, non-linear relationships present in real-world clinical data. Future work could explore resampling techniques like SMOTE, or more advanced models such as Random Forest, Gradient Boosting, or Neural Networks to enhance predictive performance. Nonetheless, the transparency and interpretability of logistic regression make it a practical choice in clinical settings, where understanding model decisions is just as important as accuracy. The insights derived from feature coefficients also offer clinicians guidance on key stroke risk factors such as age, glucose level, hypertension, and heart disease, further supporting its value in proactive, data-driven healthcare.

## VI. CONCLUSION

This study presented a logistic regression-based approach for early stroke risk prediction using demographic and clinical health data. By focusing on model simplicity and interpretability, it demonstrates how even foundational machine learning methods can support proactive healthcare efforts.

Rather than aiming for perfect classification, the model emphasizes sensitivity, making it suitable as a screening tool in clinical settings where early detection is essential. The work reinforces the importance of data-driven solutions in medical diagnostics and lays the groundwork for more advanced, precision-focused models in future research.

## ACKNOWLEDGMENT

The authors gratefully acknowledge the Department of Biotechnology, KIT–Kalaignarkarunanidhi Institute of Technology, Coimbatore, for providing the necessary facilities and resources to carry out this research work. We extend our sincere thanks to our research supervisor, Prof. Gopal Samy B, for his constant guidance, encouragement, and valuable suggestions throughout the project. We also thank our faculty and peers for their support during the development and validation of the machine learning model used in this study.

## AUTHOR CONTRIBUTIONS

Pragatheeswari M: Writing – Original Draft, Data Curation, Resources, Visualisation.

Hari Preetha T: Writing – Review & Editing, Formal Analysis.

Trishitha K: Investigation, Validation, Writing – Review & Editing.

Gopal Samy B: Supervision, Reviewing & Project Administration.

**DECLARATION OF COMPETING INTEREST**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**VII. REFERENCES**

[1]WHO. (2023). Stroke: Key facts. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/stroke

[2] Feigin, V. L., Norrving, B., & Mensah, G. A. (2017). Global burden of stroke. Circulation Research, 120(3), 439–448. https://doi.org/10.1161/CIRCRESAHA.116.308413

[3] Goldstein, L. B., Bushnell, C. D., Adams, R. J., Appel, L. J., Braun, L. T., Chaturvedi, S., ... & Stroke Council. (2011). Guidelines for the primary prevention of stroke: A guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke, 42(2), 517–584. https://doi.org/10.1161/STR.0b013e3181fcb238

[4] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep HER: A survey of recent advances in deep learning techniques for electronic health record (HER) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

[5] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). John Wiley & Sons.https://scholar.google.com/citations?user=780ULboAAAAJ&hl=en&oi=sra

[6]Zhou, Y., Zhang, J., Wang, X., & Tang, X. (2021). Application of logistic regression and machine learning models to predict stroke risk: A comparative study. BMC Medical Informatics and Decision Making, 21(1), 51. https://doi.org/10.1186/s12911-021-01406-6

[7]Zheng, Q., Zhao, A., Wang, X., Bai, Y., Wang, Z., Wang, X., … & Dong, G. (2025). Machine learning algorithms to predict stroke in China based on causal inference of time series analysis. BMC neurology, 25(1), 236. https://doi.org/10.48550/arXiv.2503.14512

[8] Le, N. B., Pham, T. T. H., Nguyen, S. H., Nguyen, N. M., & Nguyen, T. N. (2024). AI-powered predictive model for stroke and diabetes diagnostic. Int. J. Intell. Syst. Appl.(IJISA), 16(1), 24-40. https://doi.org/10.5815/ijisa.2024.01.03

[9] Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. https://doi.org/10.1214/15-AOAS848

[10] Lee, M., Yeo, N. Y., Ahn, H. J., Lim, J. S., Kim, Y., Lee, S. H., … & Kim, C. (2023). Prediction of post-stroke cognitive impairment after acute ischemic stroke using machine learning. Alzheimer's Research & Therapy, 15(1), 147. https://doi.org/10.1186/s13195-023-01289-4

[11] Li, L. (2024). Stroke Prediction Base on Logistic Regression Model. Highlights in Science, Engineering and Technology, 123, 574-578. https://doi.org/10.54097/cx2f3j88

[12] Mbarek, L., Chen, S., Jin, A., Pan, Y., Meng, X., Yang, X., … & Wang, Y. (2024). Predicting 3-month poor functional outcomes of acute ischemic stroke in young patients using machine learning. European Journal of Medical Research, 29(1), 494. https://doi.org/10.1186/s40001-024-02056-3

[13] Vodencarevic, A., Weingärtner, M., Caro, J. J., Ukalovic, D., Zimmermann-Rittereiser, M., Schwab, S., & Kolominsky-Rabas, P. (2022). Prediction of recurrent ischemic stroke using registry data and machine learning methods: the Erlangen stroke registry. Stroke, 53(7), 2299-2306. https://doi.org/10.1161/STROKEAHA.121.036557

[14] Wang, Y., Zhang, Z., Zhang, Z., Chen, X., Liu, J., & Liu, M. (2025). Traditional and machine learning models for predicting haemorrhagic transformation in ischaemic stroke: a systematic review and meta-analysis. Systematic Reviews, 14(1), 4. https://doi.org/10.1186/s13643-025-02771-w

[15]Wolfe, C. D. (2000). The impact of stroke. British Medical Bulletin, 56(2), 275–286. https://doi.org/10.1258/0007142001903120

[16] Yang, C., Hu, R., Xiong, S., Hong, Z., Liu, J., Mao, Z., & Chen, M. (2024). Development of machine learning-based models for predicting risk factors in acute cerebral infarction patients: a clinical retrospective study. BMC neurology, 24(1), 306. https://doi.org/10.1186/s12883-024-03818-6