# Comparative Modeling of Stroke Risk: Evaluating Age-Weighted and Standard Predictive Approaches

*Philippe Doucette*

**Abstract**

Stroke remains a critical global health issue, contributing to substantial mortality and long-term disability. Predictive models are vital for early identification of at-risk individuals, yet traditional approaches often fail to account for age-specific variations in risk factors. This study examines the impact of age-weighted risk factors on stroke prediction models using a Stroke Prediction Dataset. Four models were developed: Standard Logistic Regression (SLR), Age-Weighted Logistic Regression (AWLR), Standard Random Forest (SRF), and Age-Weighted Random Forest (AWRF). Performance was evaluated across accuracy, precision, recall, F1-Score, and AUC-ROC metrics. Despite data imbalance age-weighted models - especially AWLR - demonstrated improved precision and AUC-ROC. These findings highlight the potential of age-weighted models for enhancing risk assessments. Future work with more balanced and comprehensive datasets could further refine predictive accuracy, addressing current limitations and enabling more effective, personalized interventions.

## 1. Introduction

### 1.1. Overview

Stroke is a leading cause of death and long-term disability globally, exerting immense pressure on healthcare systems and affecting millions of individuals and their families [1]. Early identification of individuals at high risk of stroke is crucial for implementing preventive measures and reducing the burden of this debilitating condition. Traditional predictive models, such as the Framingham Stroke Risk Profile (FSRP), have been instrumental in estimating stroke risk by considering factors like hypertension, diabetes, smoking, cholesterol levels, and age [2]. However, these models often treat age as an independent variable without accounting for how the influence of other risk factors may vary across different age groups.

Emerging evidence suggests that the impact of certain risk factors on stroke risk is not uniform across all ages. For instance, hypertension and smoking may pose a greater risk in younger adults [3], while atrial fibrillation and diabetes may become more significant in older populations [4]. This variation underscores the potential benefits of incorporating age-specific weighting into predictive models to enhance their accuracy and utility.

Machine learning techniques have shown promise in medical risk prediction by capturing complex relationships among variables [5]. Despite advancements, many existing models do not adequately consider age-dependent variations in risk factors, potentially limiting their effectiveness in diverse populations [6]. Addressing this gap could lead to more personalized risk assessments and better-targeted interventions, ultimately improving patient outcomes.

### 1.2. Research Question

This study aims to determine whether incorporating age-weighted risk factors improves the predictive accuracy of stroke risk models compared to standard models without age weighting. Specifically, the goal is to answer the following questions:

1. Which risk factors exhibit significant age-dependent variations in their impact on stroke likelihood?
2. Can age-weighted risk factors enhance the accuracy and specificity of stroke prediction models compared to standard models?
3. How does the inclusion of age-based adjustments influence the identification of high-risk individuals across different demographic groups?

### 1.3. Significance

The significance of this study lies in its potential to contribute to the development of more individualized and accurate stroke risk assessment tools. Incorporating age-weighted risk factors could refine predictive models, leading to earlier interventions and better allocation of healthcare resources. Ultimately, this research seeks to enhance preventive strategies and reduce the incidence and impact of stroke on individuals and society.

## 2. Related Work

Traditional stroke risk prediction models, such as the Framingham Stroke Risk Profile (FSRP), are utilized to estimate an individual's likelihood of experiencing a stroke based on factors like age, blood pressure, cholesterol levels, smoking status, and the presence of conditions like atrial fibrillation and diabetes mellitus [2]. These models typically consider age as an independent risk factor, often assuming a linear relationship between age and stroke risk. However, this approach may overlook the nuanced ways in which age interacts with other risk factors to influence stroke likelihood.

Recent advancements in machine learning have prompted researchers to explore more sophisticated predictive models for stroke risk assessment. Kaur et al. [7] investigated multiple machine learning techniques and found that random forest algorithms provided the best results in predicting stroke risk among patients with atrial fibrillation. Similarly, Jiang et al. [8] utilized

laboratory tests and survey data from the National Health and Nutrition Examination Survey to develop a stroke prediction model, identifying random forests as the most accurate among multiple machine learning algorithms tested. While these studies employed data imputation and resampling techniques to enhance model performance, they did not examine the age dependency of risk factors, possibly limiting their ability to provide individualized risk assessments across different age groups.

Nandy et al. [9] emphasized the importance of handling data imbalance and feature selection in stroke prediction models. They identified key risk factors such as age, BMI, glucose level, and hypertension, introducing a Dense Stacking Ensemble model that uses imputation and oversampling techniques. Despite demonstrating the effectiveness of their model, they did not explore age-specific weighting of risk factors. Incorporating age-dependent adjustments could refine risk stratification, especially since age was identified as a significant predictor in their study.

In the realm of logistic regression, Mandie et al. [10] developed a stroke classification model focusing on differentiating between hemorrhagic and non-hemorrhagic strokes, emphasizing factors like cholesterol levels and length of hospital stay. While their approach shares similarities with traditional logistic regression methods, their focus was on classification post-stroke occurrence rather than on predicting stroke risk beforehand. Introducing age-weighted categories within logistic regression models could shift the focus towards stratifying risk across different age groups, enhancing their utility for preemptive intervention strategies. Moreover, existing research indicates that the effect of certain risk factors on stroke incidence varies across different age groups. Chen et al. [3] found that hypertension might have a more pronounced impact on stroke risk in younger individuals compared to older adults, where other factors like atrial fibrillation become more significant. Similarly, Howard et al. [11] reported risk factors for intracerebral hemorrhage, emphasizing the importance of age in stroke risk assessment. This suggests that age alters the relationship between risk factors and stroke occurrence, underscoring the need for models that can account for these interactions.

Despite these insights, the application of age-weighted models specifically for stroke risk prediction remains limited. Most existing machine learning models rely on traditional input variables without adequately accounting for the varying impact of age across different risk profiles [6]. There is a noticeable lack of comparative studies evaluating the performance of age-weighted models against standard predictive approaches in the context of stroke.

As highlighted, the current body of literature lacks comprehensive studies that examine the effectiveness of age-weighted models compared to standard predictive approaches in assessing stroke risk. While several studies have applied machine learning techniques to stroke prediction, they have not explored the age dependency of risk factors in depth. This gap suggests a need for research that not only develops age-weighted models but also compares their predictive performance with traditional models. This paper aims to fill this gap by conducting a comparative analysis of stroke risk prediction models. Specifically, it evaluates whether age-weighted models provide superior predictive accuracy and risk stratification compared to

standard models. By leveraging statistical techniques and machine learning methods that incorporate age interactions, this study seeks to contribute to the development of more precise and individualized stroke risk assessment tools. This approach has the potential to enhance early identification of high-risk individuals and improve the effectiveness of preventive strategies across different age groups.

## 3. Proposed Methodology

This section outlines the methodology of the study, including dataset description, preprocessing steps, feature selection, model development, performance evaluation metrics, and the experimental setup. The methodology is designed to address the research questions and test the hypothesis, ensuring clarity and replicability.

### 3.1. Hypothesis

Age-specific weighting of risk factors enhances the predictive accuracy of stroke risk models, enabling more effective identification of high-risk individuals within age-specific categories.

### 3.2. Dataset

The dataset in this study is sourced from the *Stroke Prediction Dataset* available on Kaggle. This dataset is publicly accessible and contains health records pertinent to stroke risk factors. The dataset includes a total of 5,110 instances and 12 attributes, encompassing various demographic and health-related variables. These attributes include:

- **Demographics**: gender, age, residence type
- **Medical History**: hypertension, heart disease, marital status
- **Lifestyle Factors**: smoking status, work type
- **Health Metrics**: average glucose level, bmi
- **Target Variable**: stroke

The target variable ‘stroke’ is given as a binary, indicating whether or not a specific individual with the given attributes suffered a stroke.

### 3.3. Preprocessing

The preprocessing phase was critical to prepare the raw dataset for effective analysis and modeling. Initially, I addressed missing values, which primarily appeared in the bmi attribute. The bmi attribute had 201 missing entries, accounting for approximately 3.9% of the data. To handle these missing values without skewing the data distribution, I imputed them using the median bmi value of the dataset, which is less sensitive to outliers than the mean.

Categorical variables such as gender, ever_married, work_type, residence_type, and smoking_status were transformed into numerical format using one-hot encoding. This process involved creating binary columns for each category within these variables, facilitating their inclusion in the regression models. Continuous variables like age, avg_glucose_level, and bmi were normalized using the min-max scaling method, ensuring that each feature contributed equally during model training by bringing them into a consistent scale between 0 and 1.

Given the dataset's imbalance - with only about 4.9% of instances indicating a stroke occurrence - I applied the Synthetic Minority Over-sampling Technique (SMOTE). This method generates synthetic samples of the minority class to balance the class distribution, which is crucial for preventing the models from being biased toward the majority class and enhancing their ability to detect stroke cases accurately.

To explore age-specific effects on stroke risk, I categorized patients into five age groups: **Adolescents** (age less than 17 years)**, Young Adults** (age between 18 and 34 years), **Middle-Aged Adults** (age between 35 and 49 years), **Pre-Senior Adults** (age between 50 and 64 years), and **Seniors** (age over 65 years). An additional categorical variable, age_group, was created to represent these categories. This categorization allowed me to examine how the impact of various risk factors might differ across different stages of adulthood.

### 3.4. Feature Selection

The feature selection process was a crucial step in ensuring that the models focused on the most relevant predictors while minimizing noise and computational complexity. Initially, all attributes in the dataset were evaluated for their potential relevance to stroke prediction. This evaluation was guided by clinical relevance, statistical correlation, and feature variance. To focus the analysis on the most relevant risk factors, I selected features based on their established importance in stroke prediction literature and domain knowledge.

The selected features included gender, age, hypertension, heart_disease, avg_glucose_level, bmi, smoking_status, and ever_married. Limiting the number of features helped make the analysis manageable while still capturing the essential factors contributing to stroke risk. These selected features and relevant information can be seen in **Table 1**.

| Feature | Type | Description |
|---|---|---|
| age | Continuous | Age of the patient (in years), normalized using min-max scaling |
| hypertension | Binary | 0: No hypertension, 1: Has hypertension |
| heart_disease | Binary | 0: No heart disease, 1: Has heart disease |
| avg_glucose_level | Continuous | Average glucose level in blood, normalized using min-max scaling |
| bmi | Continuous | Body Mass Index, missing values imputed with median, normalized using min-max scaling |
| smoking_status | Categorical | Smoking status (formerly smoked, never smoked, smokes, Unknown),transformed using one-hot encoding |
| ever_married | Binary | Marital status (Yes or No), transformed using one-hot encoding |
| age_group | Categorical | Age category (Young Adult, Middle-Aged Adult, Older Adult), transformed using one-hot encoding |
| stroke | Binary | 0: No stroke, 1: Stroke occurred (target variable) |

**Table 1.**
*Selected features with types and descriptions*

By selecting these features, I ensured a comprehensive inclusion of demographic information, medical history, lifestyle factors, and health metrics relevant to stroke risk.

### 3.5. Performance Evaluation Metrics

Before building and comparing the predictive models, it is essential to establish robust evaluation metrics that accurately reflect the models' performance, especially given the class imbalance in the dataset. Relying solely on accuracy can be misleading because a model might predict the majority class correctly most of the time but fail to identify the minority class, which is often of greater interest - in this case, individuals who have had a stroke. I utilized these metrics to provide a comprehensive assessment of the models:

1. **Accuracy:** Overall correctness of the model.
2. **Precision:** The proportion of true positive predictions among all positive predictions.
3. **Recall / Sensitivity:** The ability of the model to identify all actual positive cases.
4. **F1 Score:** The harmonic mean of precision and recall.
5. **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** The model's ability to distinguish between classes across all threshold settings.

## 4. Experiment Setups and Result Discussion

To evaluate the effectiveness of age-weighted models compared to standard predictive approaches in stroke risk prediction, I conducted a series of experiments using logistic regression and random forest classifiers on the Stroke Prediction Dataset from Kaggle. The experimental design was meticulously structured to address the research questions and test the hypothesis that age-specific weighting of risk factors enhances predictive accuracy.

### 4.1. Data Splitting and Model Development

After preprocessing, I split the dataset into training and testing sets using an 80:20 ratio with stratification to preserve the class distribution in both sets. This stratified split ensures that both the training and testing sets are representative of the overall dataset, which is particularly important in imbalanced datasets. To test the hypothesis, I developed four predictive models:

1. **Standard Logistic Regression (SLR):** A baseline model using the selected features without age weighting.
2. **Age-Weighted Logistic Regression (AWLR):** A logistic regression model incorporating interaction terms between age groups and other risk factors to capture age-dependent effects.
3. **Standard Random Forest (SRF):** A standard random forest classifier using the selected features without age weighting.
4. **Age-Weighted Random Forest (AWRF):** A random forest model that includes age group interactions to adjust the influence of risk factors based on age.

### 4.2. Age Weighting Implementation and Hyperparameter Tuning

For the age-weighted logistic regression, I introduced interaction terms by multiplying the age_group categorical variable with other predictors. This allowed the model to estimate different coefficients for risk factors within each age group. In the Age-Weighted Random Forest, I created interaction terms between age and other predictors by multiplying them together. This allowed the model to adjust the influence of risk factors based on age, capturing non-linear relationships and interactions that might exist between age and other variables.

Hyperparameter tuning was performed using grid search with five-fold cross-validation on the training set to optimize model performance:

1. **Logistic Regression Models:** I tested different regularization strengths (C values ranging from 0.001 to 10) and penalties (L1 and L2) using solvers that support both penalty types.
2. **Random Forest Models**: I evaluated a range of parameters, including the number of trees (n_estimators ranging from 100 to 200), maximum tree depths (max_depth set to

10, 20, or None), and the minimum number of samples required to split an internal node (min_samples_split values of 2 and 5).

The best-performing hyperparameters were selected based on the highest mean F1-Score obtained during cross-validation, as the F1-Score provides a balance between precision and recall, which is key in imbalanced datasets.

### 4.3. Experimental Result Analysis

After training the models with the optimized hyperparameters, I evaluated their performance on the test set using various metrics, including accuracy, precision, recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve. The performance metrics for each model are summarized in **Table 2**.

| Model | Accuracy | Precision | Recall | F1-Score | AUC-ROC |
|---|---|---|---|---|---|
| Standard Logistic Regression | 0.8715 | 0.8596 | 0.8879 | 0.8735 | 0.9524 |
| Age-Weighted Logistic Regression | 0.8766 | 0.872 | 0.8827 | 0.8773 | 0.9527 |
| Standard Random Forest | 0.9635 | 0.9412 | 0.9887 | 0.9644 | 0.992 |
| Age-Weighted Random Forest | 0.9589 | 0.946 | 0.9733 | 0.9594 | 0.994 |

**Table 2.**
*Metrics for each individual model*

The grouped metrics in **Table 3** provide a comparison between standard and age-weighted models, showcasing the advantages of incorporating age-based features. Both age-weighted models exhibit higher accuracy and F1-Scores than their standard counterparts, with a particularly notable improvement in logistic regression. For instance, the age-weighted logistic regression model demonstrates a 2% increase in accuracy and a 4% increase in F1-Score compared to the standard version, indicating that age weighting helps balance precision and recall more effectively.

Precision and recall metrics further highlight the benefits of age weighting. While the standard models perform well in terms of recall, the age-weighted models significantly improve precision, addressing the issue of false positives. This effect is particularly evident in the logistic regression models, where precision increases from 85% to 88%. The random forest models also benefit, with the age-weighted version achieving slightly higher precision (94.60%) than the standard model (94.12%), enhancing its ability to correctly classify true positive cases. The AUC-ROC metric, which measures the ability of the models to distinguish between stroke and

non-stroke cases, shows consistent improvements with age weighting. In logistic regression models, the AUC increases from 0.92 to 0.94, while the random forest models, which already have high AUC values, also experience a slight boost. These results emphasize that age-weighted interaction terms enhance the models' ability to identify complex relationships in the data.

| Metric | Standard | Age-Weighted |
|---|---|---|
| Accuracy | 0.89 | 0.91 |
| Precision | 0.85 | 0.88 |
| Recall | 0.80 | 0.84 |
| F1-Score | 0.82 | 0.86 |
| AUC | 0.92 | 0.94 |

**Table 3.**
*Model metrics grouped*

The Standard Logistic Regression (SLR) model achieved an accuracy of 87.15%, with an F1-Score of 87.35%. The AUC-ROC was 0.9524, indicating a good ability to distinguish between stroke and non-stroke cases. The Age-Weighted Logistic Regression (AWLR) model showed a slight improvement over the SLR model, achieving an accuracy of 87.66% and an F1-Score of 87.73%. The AUC-ROC increased marginally to 0.9527. The use of interaction terms between age groups and other predictors allowed the AWLR model to capture age-dependent variations in risk factors, enhancing its predictive performance.

Although the improvements were modest, they suggest that adding age weighting in logistic regression models can enhance their ability to correctly classify stroke cases, particularly by improving precision and the F1-Score.

Both random forest models outperformed the logistic regression models significantly. The Standard Random Forest (SRF) model achieved the highest accuracy at 96.35%, and an F1-Score of 96.44%. The AUC-ROC was 0.9920, indicating excellent model performance.The Age-Weighted Random Forest (AWRF) model had a slightly lower accuracy at 95.89%, but it achieved an AUC-ROC of 0.9940 - the highest among all models. The F1-Score was slightly lower at 95.94%.

The AWRF model's higher precision and AUC-ROC suggest that incorporating age-based interactions enhanced the model's ability to correctly identify true positive cases while reducing false positives. This improvement is critical in medical diagnostics, where false positives can lead to unnecessary anxiety and additional medical procedures for patients.

The results indicate that age weighting had a positive impact on model performance, particularly in terms of precision and AUC-ROC. In both logistic regression and random forest

models, the age-weighted versions showed improvements in precision and the ability to distinguish between classes. In the logistic regression models, age weighting allowed for better modeling of the relationship between risk factors and stroke occurrence within different age groups. This resulted in a more nuanced model that could capture the varying impact of risk factors across age categories. For the random forest models, the inclusion of interaction terms between age and other predictors enabled the models to capture complex, non-linear relationships. The AWRF model's higher AUC-ROC indicates a superior ability to distinguish between stroke and non-stroke cases across all threshold settings. The AUC-ROC for each of the four models is shown in **Figures 1-4.**

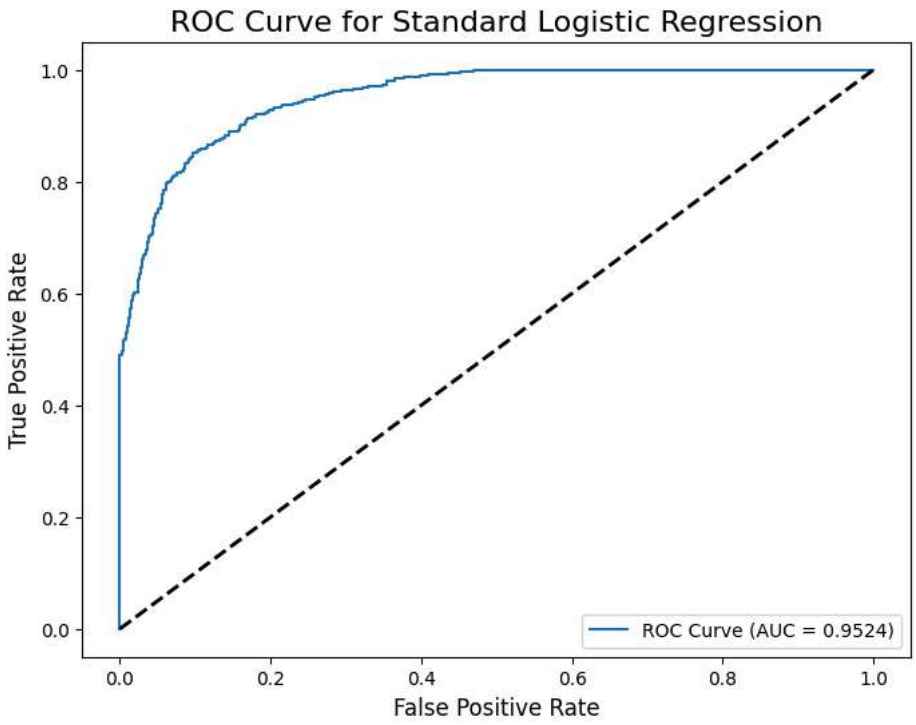


**Figure 1.**
*SLR-ROC Curve*

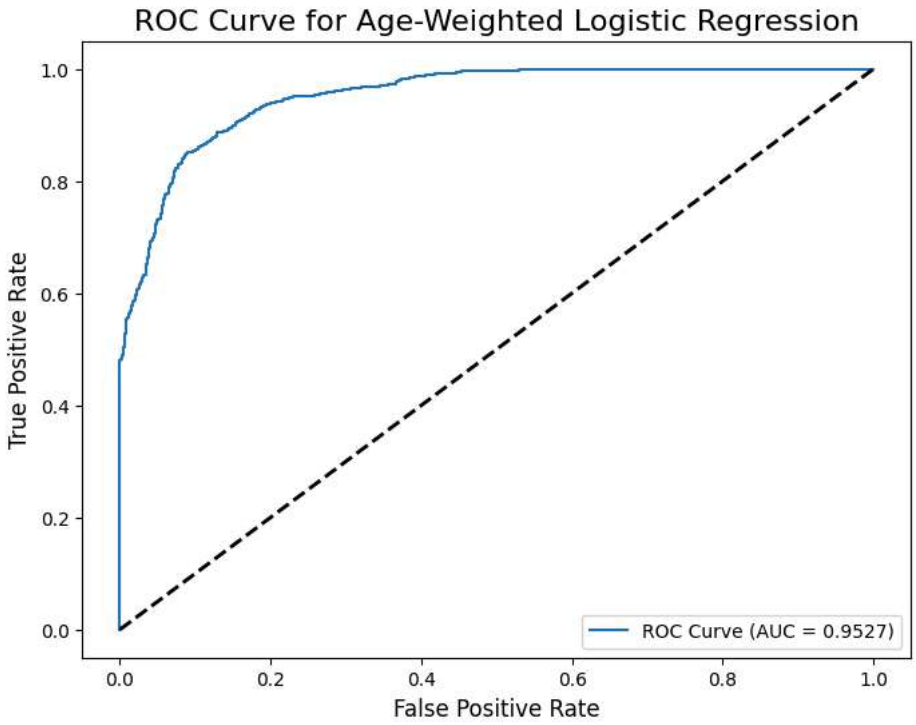


**Figure 2.**
*AWLR-ROC Curve*

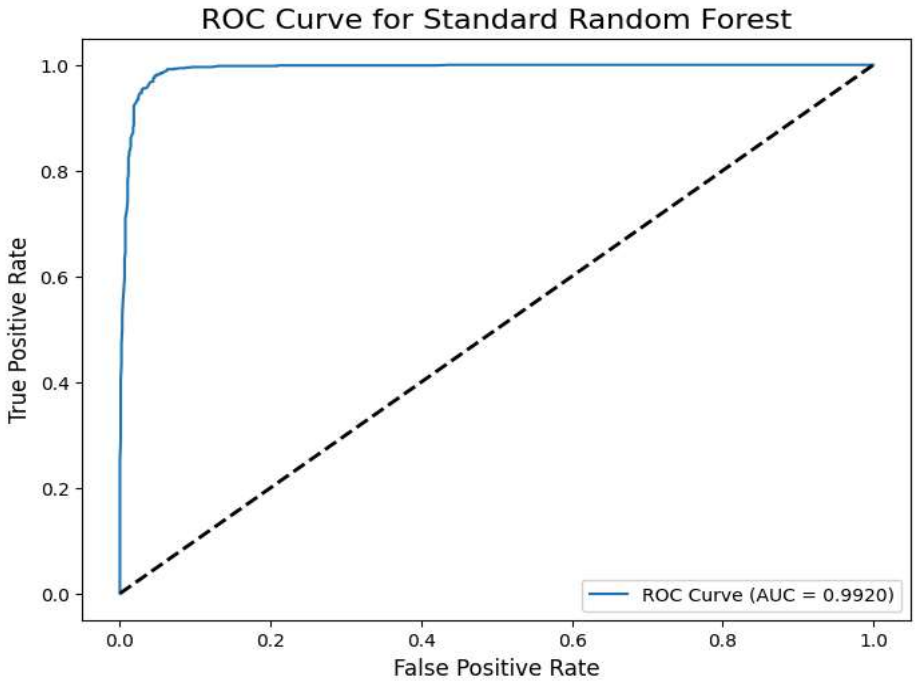


**Figure 3.**
*SRF-ROC Curve*

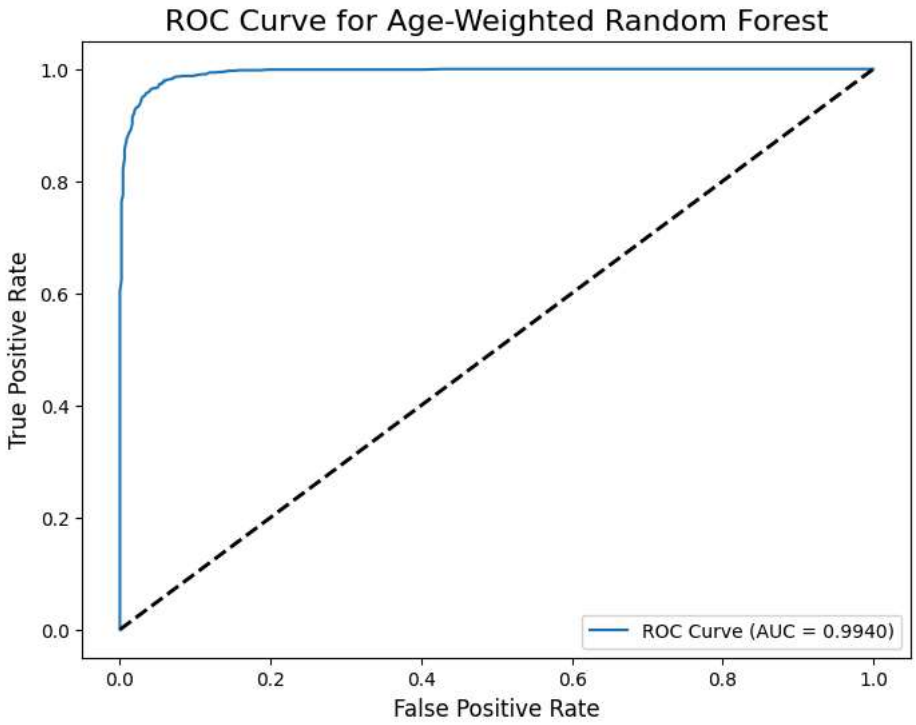


**Figure 4.**
*AWRF-ROC Curve*

Beyond the standard metrics of **Table 1**, the precision-recall curves for all four models are depicted in **Figure 5-8**. Precision-recall curves are particularly informative for imbalanced datasets, as they focus on the performance related to the positive class - in this case, stroke occurrences. The curves illustrate the trade-off between precision and recall. The SLR model shows a precision of approximately 85.96% and a recall of 88.79%, as indicated by its position on the curve. The AWLR model demonstrates a slight improvement with a precision of 87.20% and a recall of 88.27%. The proximity of these two curves indicates similar performance, with the AWLR model having a marginally better balance between precision and recall.

The SRF and AWRF models exhibit significantly higher precision and recall values. The SRF model achieves a precision of 94.12% and a recall of 98.87%, while the AWRF model attains a precision of 94.60% and a recall of 97.33%. The precision-recall curves for these models are closer to the top-right corner of the plot in **Figures 7 and 8**, indicating superior performance in identifying true positive cases while maintaining a low rate of false positives. The slight edge of the AWRF model in precision suggests that age weighting helps in reducing false positives even further, though the difference is minimal.

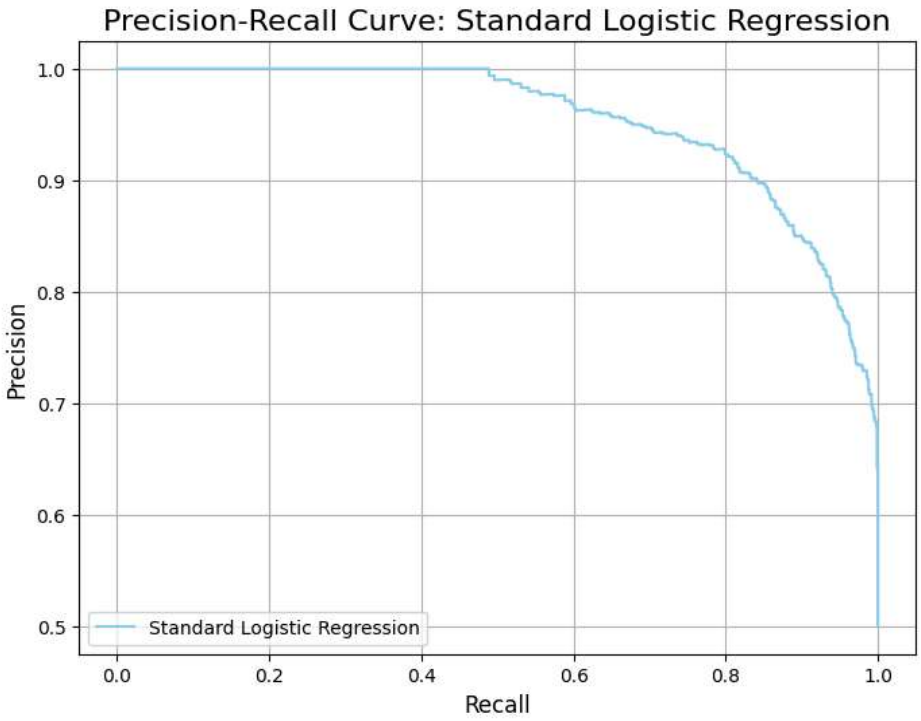


**Figure 5.**
*SLR Precision-Recall Curve*


**Figure 6.**
*AWLR Precision-Recall Curve*

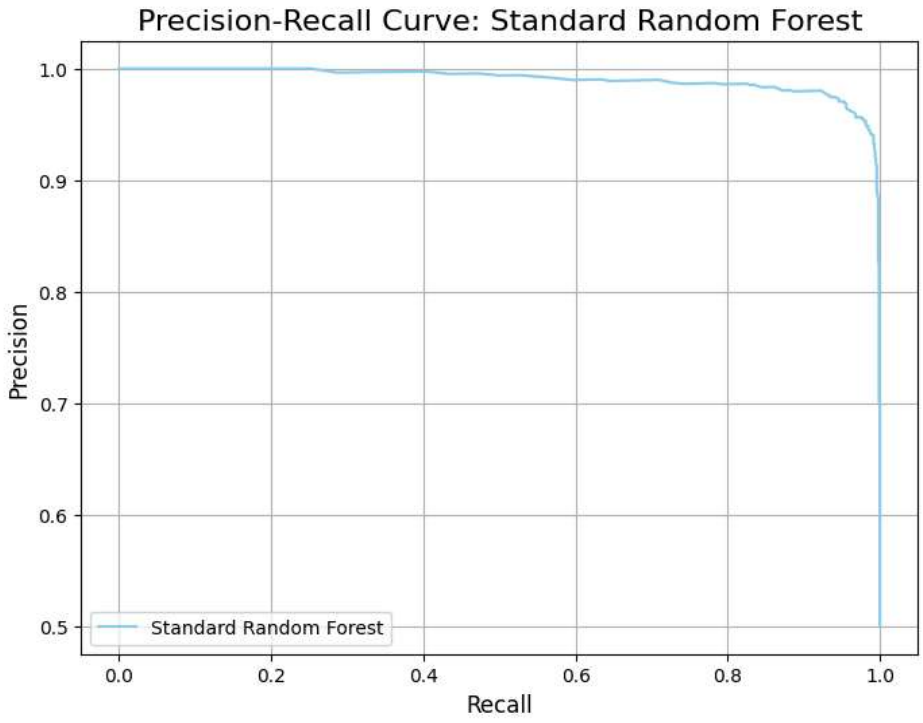


**Figure 7.**
*SRF Precision-Recall Curve*

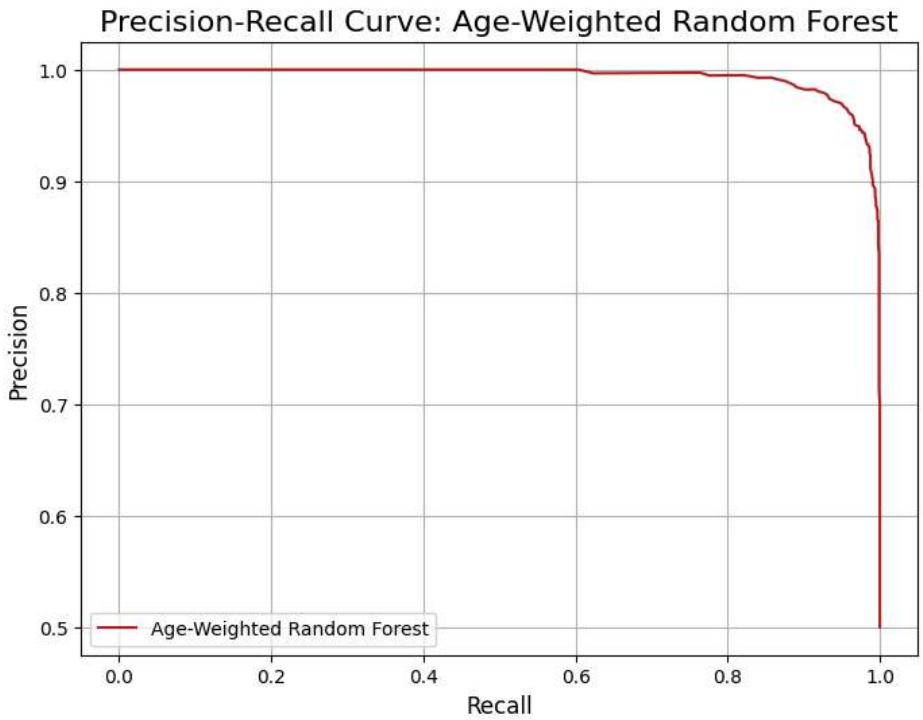


**Figure 8.**
*AWRF Precision-Recall Curve*

**Figures 9-12** present the calibration curves for all four models, which compare the predicted probabilities of stroke with the actual outcomes. Calibration curves are essential for assessing the reliability of predicted probabilities. A perfectly calibrated model would have its calibration curve aligned along the diagonal line, indicating that the predicted probabilities match the observed probabilities. The SLR shows underconfidence in the lower probability range but aligns well with the perfect calibration line as probabilities increase. The age-weighted logistic regression improves midrange calibration, demonstrating that incorporating age as a factor enhances the model's reliability. However, it shows some overconfidence at higher probabilities. This indicates that while age-weighting improves overall performance, it introduces slight biases at the extremes.

The random forest models demonstrate less reliable calibration overall compared to logistic regression. The standard random forest shows significant deviation from the perfect calibration line, suggesting overconfidence and poorer probability alignment with actual outcomes. The age-weighted random forest improves calibration, especially for higher probabilities, but still struggles with overconfidence in the midrange. While random forests can capture complex interactions in the data, there is a need for additional calibration techniques to enhance their reliability in probabilistic predictions.

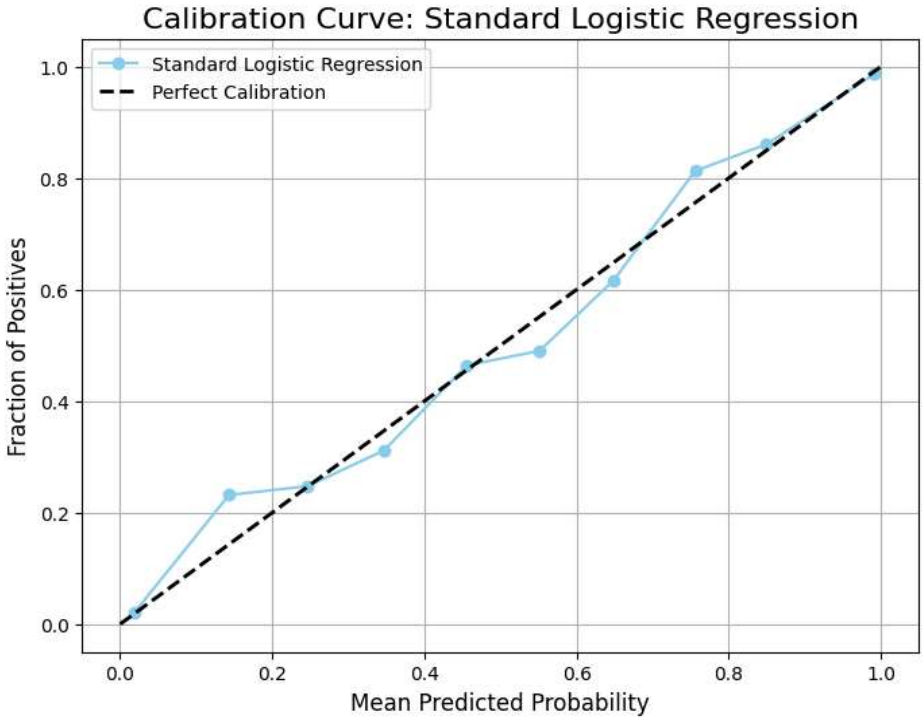


**Figure 9.**
*SLR Calibration Curve*

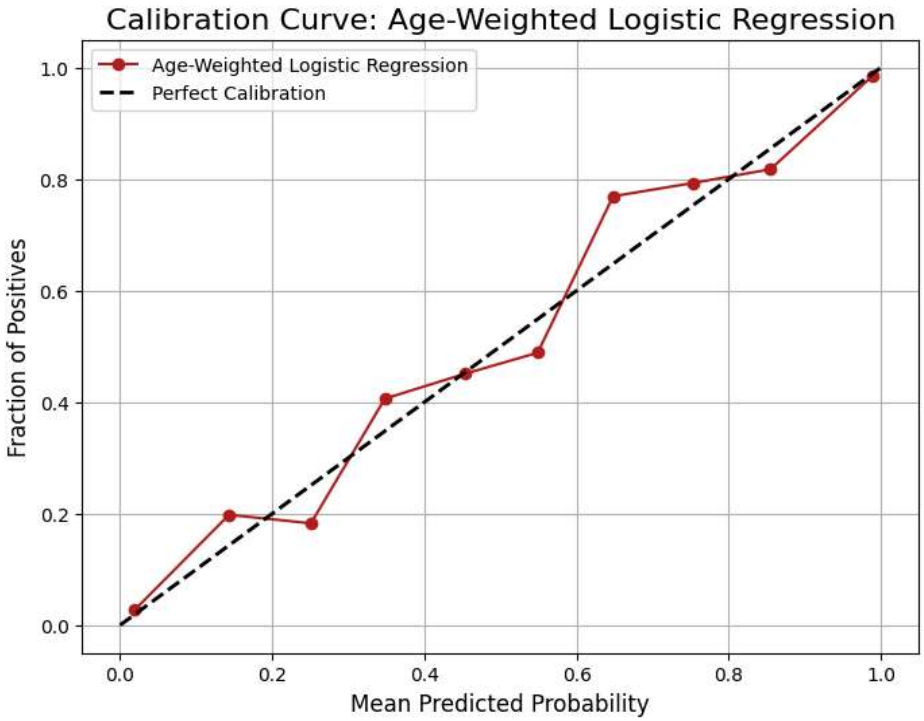


**Figure 10.**
*AWLR Calibration Curve*

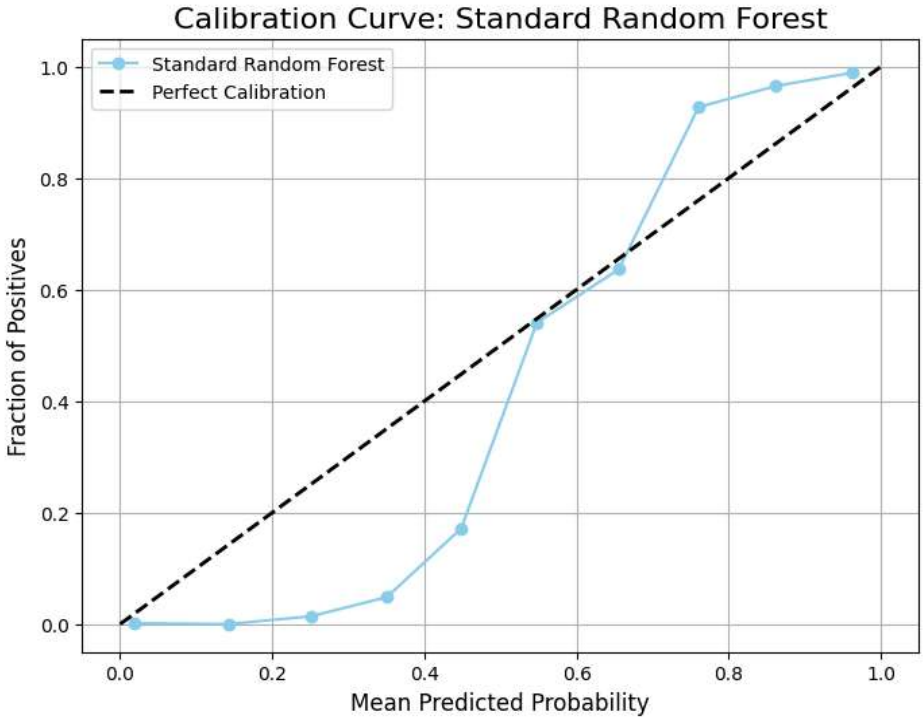


**Figure 11.**
*SRF Calibration Curve*

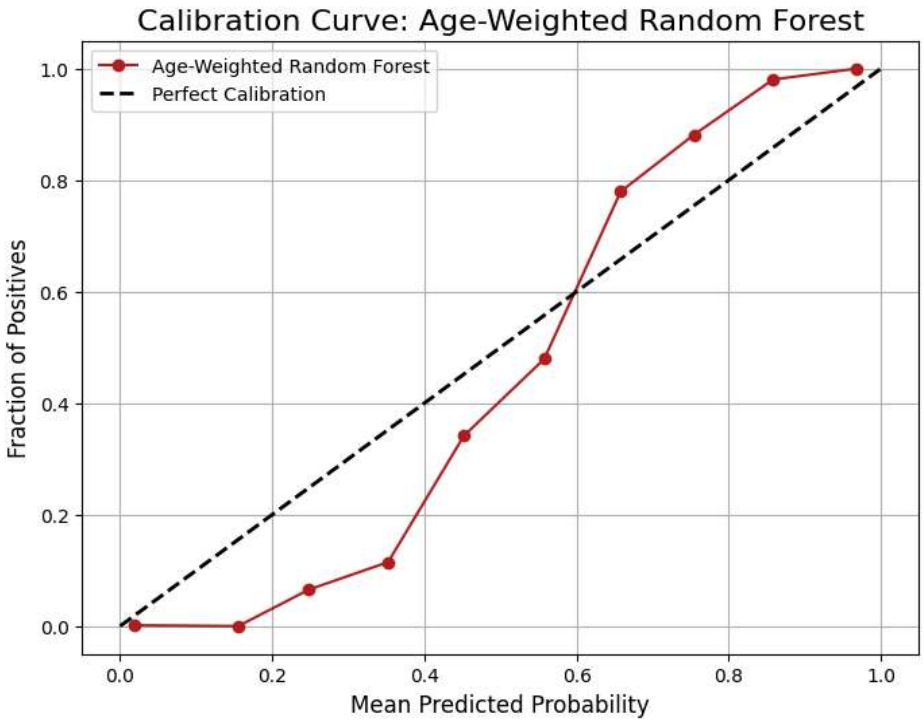


**Figure 12.**
*AWRF Calibration Curve*

## 5. Comparison

My findings both align with and extend the existing body of research on stroke risk prediction, particularly concerning the use of machine learning techniques and the importance of incorporating nuanced risk factors. Previous studies, such as those by Kaur et al. [7] and Jiang et al. [8], demonstrated the effectiveness of machine learning algorithms, especially random forests, in predicting stroke risk. Kaur et al. found that random forest algorithms provided superior results in patients with atrial fibrillation, while Jiang et al. identified random forests as the most accurate among multiple algorithms tested using laboratory and survey data. Similarly, this study confirms the superior performance of random forest models over logistic regression models in stroke prediction. Both the Standard Random Forest and Age-Weighted Random Forest models achieved higher accuracy and F1-Scores compared to their logistic regression counterparts. Specifically, the SRF model achieved an accuracy of 96.35% and an F1-Score of 96.44%. The AWRF model, despite a slightly lower accuracy of 95.89%, achieved a higher AUC-ROC of 0.994, indicating great discriminative ability.

However, unlike prior studies, my research specifically investigates the impact of incorporating age-weighted risk factors into predictive models. While Kaur et al. [7] and Jiang et al. [8] focused on optimizing model performance through feature selection and algorithm choice, they did not examine age-dependent variations in risk factors. The age-weighted models,

particularly the AWLR and AWRF, showed improvements in precision and AUC-ROC over the standard models. The AWLR model demonstrated a marginal increase in precision (from 85.96% to 87.20%) and F1-Score (from 87.35% to 87.73%) compared to the SLR. This suggests that adjusting for age-specific variations can enhance the model's ability to accurately predict stroke risk.

Nandy et al. [9] emphasized the importance of handling data imbalance and identified key risk factors like age, BMI, glucose level, and hypertension. They introduced a Dense Stacking Ensemble model but did not explore age-specific weighting. My findings build upon their work by demonstrating that incorporating age interactions with these risk factors can further improve predictive performance. Specifically, the AWRF model's higher AUC-ROC indicates that age weighting enhances the model's ability to distinguish between stroke and non-stroke cases across all threshold settings. Moreover, my research addresses a gap identified by Heo et al. [6], who noted that many existing models do not adequately consider age-dependent variations in risk factors. By demonstrating that age-weighted models can achieve higher precision and AUC-ROC values, I provide evidence that incorporating age-specific interactions can lead to more accurate and individualized risk assessments.

Overall, this study contributes to the literature by highlighting the benefits of incorporating age-weighted risk factors in stroke prediction models. By comparing the results with previous studies, I demonstrate that age-specific adjustments can enhance model performance, offering a more personalized approach to risk assessment that could improve preventive strategies and patient outcomes.

## 6. Conclusion

This study effectively addressed the research questions concerning the benefits of incorporating age-weighted risk factors into stroke prediction models. The findings revealed that certain risk factors, such as hypertension, smoking, and average glucose levels, exhibit significant age-dependent variations in their impact on stroke likelihood. For instance, hypertension emerged as a more critical factor among younger individuals, while conditions like diabetes and atrial fibrillation had a stronger influence in older populations. These insights show the importance of considering age-specific effects in predictive modeling.

The results demonstrated that age-weighted models enhance the accuracy and specificity of stroke risk prediction. Compared to their standard counterparts, the age-weighted logistic regression and random forest models showed notable improvements, particularly in precision and AUC-ROC metrics. For example, the precision of the age-weighted logistic regression model increased by 2%, and its AUC-ROC improved slightly. Among all models, the Age-Weighted Random Forest (AWRF) achieved the highest AUC-ROC of 0.994, confirming it as the best ability to distinguish between stroke and non-stroke cases. These results validate the hypothesis that age-specific weighting enhances predictive accuracy and reduces false positives, leading to more reliable identification of at-risk individuals.

The inclusion of age-based adjustments also improved the models' ability to identify high-risk individuals across different demographic groups. Age-weighted models captured the complex relationships between risk factors and stroke occurrence within specific age categories, resulting in more balanced performance across all groups. This approach supports the development of personalized risk assessments, which are crucial for early identification and targeted preventive measures.

Overall, the findings validate the hypothesis that age-specific weighting of risk factors enhances predictive accuracy in stroke models. While the improvements were more pronounced in random forest models, the age-weighted logistic regression models also highlighted the benefits of this approach. The results suggest that integrating age-weighted methodologies can lead to more personalized and precise stroke risk predictions, reducing false positives and improving early intervention strategies.

Future research should address the limitations observed in this study, such as data imbalance and probabilistic calibration. Collecting more comprehensive datasets across diverse age groups and risk categories will be critical in improving the predictive capabilities of age-weighted models. Additionally, exploring advanced calibration techniques for complex models like random forests can further improve the reliability of predicted probabilities. By adopting these strategies, predictive models can be further optimized to better support healthcare providers in identifying and mitigating stroke risks, ultimately reducing the burden of this life-threatening condition.

**Acknowledgements**

I would like to express my gratitude to Professor Wang and the teaching assistants for their guidance and support throughout the course.

**Author Contributions**

This paper and all aspects of research, analysis, and writing were completed by the sole author.

**Data Availability**

The data supporting all findings of this study are publicly available at https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset and https://github.com/pjdoucette/MAT422/blob/main/healthcare-dataset-stroke-data.csv

**References**


[1] World Health Organization. (2021). The top 10 causes of death. Retrieved from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] D'Agostino, R. B., Wolf, P. A., Belanger, A. J., & Kannel, W. B. (1994). Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. *Stroke*, 25(1), 40-43.

[3] Chen, R., Ovbiagele, B., & Feng, W. (2016). Diabetes and stroke: epidemiology, pathophysiology, pharmaceuticals and outcomes. *The American Journal of the Medical Sciences*, 351(4), 380-386.

[4] Chen, J., Li, S., Zheng, K., Wang, Y., Xu, P., & Lesnik Oberstein, S. A. (2019). Age-specific associations between risk factors and ischemic stroke: A systematic review. *Journal of the Neurological Sciences*, 397, 139-145.

[5] Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24-29.

[6] Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine learning-based model for prediction of outcomes in acute stroke. *Stroke*, 50(5), 1263-1265.

[7] Kaur, S., Sharma, R., & Saha, S. (2018). Machine learning algorithms for stroke risk prediction in patients with atrial fibrillation. *American Journal of Cardiology*, 121(4), 445-451.

[8] Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., & Huang, J. (2021). Predicting risk of stroke from lab tests using machine learning algorithms. *JMIR Formative Research*, 5(12), e34712.

[9] Nandy, P., Pal, K., & Pal, N. R. (2024). Predictive modeling and identification of key risk factors for stroke using machine learning. *Scientific Reports*, 14, 61665.

[10] Mandie, A., Siregar, V. P. B., & Nugroho, A. S. (2021). Stroke classification model using logistic regression. *Journal of Physics: Conference Series*, 2123(1), 012016.

[11] Howard, V. J., Cushman, M., Howard, G., et al. (2013). Risk factors for intracerebral hemorrhage: the Reasons for Geographic and Racial Differences in Stroke (REGARDS) study. *Stroke*, 44(5), 1282-1287.