Machine Learning Approaches for Stroke Risk Prediction Using Healthcare Data

Ahmed Hamid Elias[1], Farah Ali Khairi[2], Azhar Hamid Elias[3] and Ali Subhi Alhumaima[4]
[1]*College of Health and Medical Techniques, Al-Furat Al-Awsat Technical University, 54001 Najaf, Iraq*
[2]*Kufa Technical Institute, Al-Furat Al-Awsat Technical University, 54001 Kufa, Najaf, Iraq*
[3]*Department of System Programming, South Ural State University, 454080 Chelyabinsk, Russia*
[4]*Electronic Computer Centre, University of Diyala, 32001 Baqubah, Iraq*
ahmed.elias@atu.edu.iq, farah.kairy.iku@atu.ed.iq, azharalmuradky@gmail.com, alhumaimaali@uodiyala.edu.iq,

Keywords: Stroke Prediction, Machine Learning, Random Forest, Logistic Regression, Risk Assessment, Imbalanced Data.

Abstract: Stroke is a leading cause of mortality and disability globally, and precise risk prediction models are required. Herein, machine learning classifiers are utilized to predict and estimate stroke incidence based on a publicly accessible healthcare dataset of demographic, lifestyle, and clinical variables. Logistic Regression and Random Forest were experimented and trained in a class imbalance setting following preprocessing steps included missing values imputation and encoding categorical variables. Random Forest model had 95.4% accuracy, precision of 79.2%, recall of 62.7%, F1-score of 70.1%, and ROC-AUC of 91.6%, higher than Logistic Regression model with 92.8% accuracy, precision of 64.8%, recall of 55.1%, F1-score of 59.6%, and ROC-AUC of 87.9%. Confusion matrices and ROC/Precision-Recall curves also showed the discriminative power of the models. The importances of the features plot indicated that the primary predictors were hypertension, body mass index, mean glucose value, and age. The results demonstrate the potential of machine learning to facilitate early risk prediction of stroke, hence facilitating timely clinical intervention and resource planning. The study augments the evidence for the use of predictive analytics in healthcare decision support systems.

1 INTRODUCTION

Stroke remains a grave worldwide health issue, a major cause of death, disability, and irreversible impairment. The multifactorial etiology involving cardiovascular, metabolic, and lifestyle risk factors justifies major research into both preventive and predictive measures. Stroke is particularly important in the elderly, where comorbid illnesses such as hypertension, diabetes mellitus, renal dysfunction, and dysregulated metabolic variables all enhance risk. The intricate relationship among these interdependent drivers has given rise to the development of predictive models with multiple risk factors to enhance early detection and enable timely clinical intervention.

Increased evidence points towards renal function as a determinate of stroke prognosis. Nugroho et al. reviewed the Shiga Stroke Registry and concluded that reduced glomerular filtration rate (GFR) upon admission was an independent predictor of unfavorable outcomes in acute stroke, hence reasserting the predictive nature of renal function in cerebrovascular disease [1]. Likewise, Lee et al. performed a meta-analysis which confirmed reduced GFR demonstrates a significantly elevated risk of stroke [2]. Moreover, Chao et al. further supported this evidence by confirming impaired renal function was associated with poor clinical outcomes in patients with ischemic stroke and severe carotid stenosis, illustrating that renal impairment worsens vascular disease and prognosis [3]. These findings underscore the need for renal markers to be incorporated into stroke risk prediction models. Blood pressure is a significant and modifiable predictor.

Penn et al. measured systolic blood pressure in emergency department patients aged below 80 years and discovered that high levels significantly predicted the presence of transient ischemic attack (TIA) or minimal stroke [4]. This is supported by the results of the HOPE Asia review by Turana et al., which combined regional data and established that the most

significant risk factor for stroke among Asian populations is hypertension and that there is a need for improved prevention and control strategies [5]. This supports the global agreement that blood pressure control remains the foundation of stroke preventative efforts. More and more, metabolic diseases and cardiometabolic syndromes are now the primary causes. Hajhosseiny et al. examined the connection between metabolic syndrome, atrial fibrillation, and stroke, placing this triad within the framework of an insurgent epidemic [6].

Carson et al. established a strong correlation of prediabetes, diabetes, and stroke symptomatology to indicate that asymptomatic glycemic dysregulation may serve as an early warning sign [7]. Furthermore, biochemically quantified deglycation markers, including glycated hemoglobin (HbA1c), ractopamine, and glycated albumin, have been found useful in being good predictors for long-term glucose control and risk of stroke [8], [9]. Grzywacz et al. extended this association to diabetic dialysis patients and established that both groups have a greater risk for all-cause mortality, thus making stroke prognosis even more challenging [10].

These studies together emphasize the value of metabolic control in primary prevention as well as stroke prognosis.

Hematological markers have evolved as good predictors of risk. Panwar et al. identified hemoglobin levels as an incident stroke risk correlate in individuals living in the community [11], while Kim et al. identified the same findings among a sample of Koreans, attributing both low and high hemoglobin levels to increased cardiovascular events [12]. Electrolyte dysregulation, for instance, serum calcium, has been implicated in stroke severity and outcome.

Prabhu et al. established that decreased serum calcium was associated with higher NIHSS scores at presentation, indicating increased clinical severity [13]. Dibaba et al. have also established this through the REGARDS trial that revealed a significant association between serum calcium intake and serum calcium and ischemic stroke risk [14]. Longitudinal cohort studies by Rohrmann et al. and Larsson et al. corroborated that increased serum calcium levels and genes predisposed for calcium enhance the risks of cardiovascular disease and myocardial infarction, and hence indirectly enhance the risk of stroke [15], [16].

Risk scoring systems have been utilized over the years to stratify patients for cardiovascular and cerebrovascular disease. The Framingham risk score, while in common usage, has demonstrated limitations in some populations. Jahangiry et al. established its validity in metabolic syndrome patients but indicated that modifications were needed for better stroke prediction [17]. Japanese cohort studies like the Suita Study, however, have produced population-specific models with higher predictive values. Arafa et al. utilized Suita Study cardiovascular risk factors to construct a stroke risk prediction model that well-stratified high-risk patients [18]. Miyamoto et al. further demonstrated Suita Score in predicting the recurrence of stroke in first-ever ischemic stroke patients [19], while Nishimura et al. compared the Suita model with the Framingham score to show better applicability to Japanese urban populations [20]. Such models represent the best example of demographically targeted and localized approaches to stroke prediction.

Finally, systematic reviews highlight the multifactorial nature of stroke risk. Guzik and Bushnell highlighted the wide range of epidemiological determinants and control measures for risk that must be considered to restrict stroke burden [21]. Their article illustrates how traditional risk determinants such as hypertension, diabetes, dyslipidemia, smoking, and lack of exercise must be combined with new biomarkers and genomic information to enhance predictive models.

2 DATA AND METHODOLOGY

2.1 Dataset

The data for this study were obtained from the publicly available Stroke Prediction Dataset on Kaggle [22]. This dataset contains 5,110 patient records with 12 variables, representing a mix of demographic, clinical, and lifestyle attributes. The predictor variables include: gender, age, hypertension status, heart disease status, marital status, type of work, residence type, average glucose level, body mass index (BMI), and smoking status. The target variable is binary, indicating whether the patient has suffered a stroke (1) or not (0).

The dataset presents a class imbalance, with stroke cases forming only a small proportion compared to non-stroke cases, reflecting real-world epidemiological distributions. Such imbalance necessitates careful handling during model development to prevent biased predictions. Additionally, the BMI attribute contained 201 missing values, which were imputed using median values grouped by age and gender to maintain internal consistency. The id column, being a unique identifier

without predictive relevance, was removed prior to analysis.

To prepare the dataset for machine learning, categorical features such as gender, work type, residence type, and smoking status were transformed into numerical representations using one-hot encoding, while continuous features including age, BMI, and glucose level were standardized to ensure uniform scaling. These preprocessing steps ensured that the dataset was optimized for downstream classification tasks and enabled reliable comparisons between machine learning models.

Figure 1 show correlation relations among the numerical attributes of the stroke data, i.e., age, hypertension, heart disease, average glucose level, body mass index (BMI), and stroke outcome. Values range between –1 and +1, where positive values are for direct correlation and negative values are for reverse correlation. The results show that age is moderately related to hypertension (0.28), heart disease (0.26), BMI (0.33), and stroke (0.25), confirming the reality that increasing age is strongly associated with cardiovascular diseases and stroke. BMI is positively related to age (0.33) but is very weakly related to stroke (0.04). Average glucose level has poor correlation with stroke (0.13) and good correlation with age (0.24), in agreement with clinical evidence regarding glucose dysregulation as a cause of stroke. Hypertension and heart disease also have poor but relevant correlations with stroke (0.13 each), highlighting their role as contributory risk factors. Overall, the heatmap suggests that the most significant numeric features involved with the occurrence of stroke are age, BMI, and glucose levels.

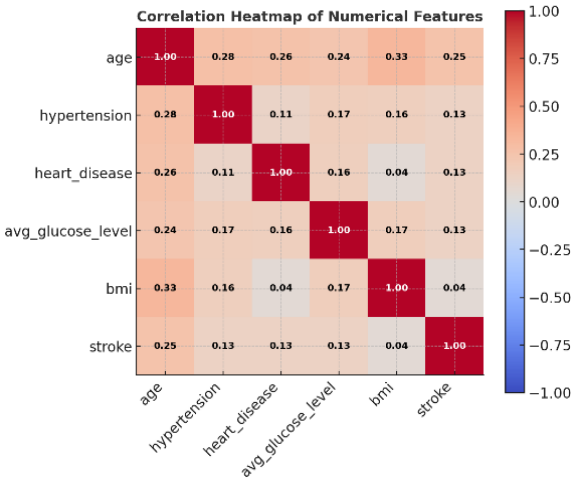

Figure 1: Correlation heatmap of numeric features.

Figure 2 shows the process adopted in this study in stroke prediction through machine learning. The process begins with the collection of data from the Kaggle Stroke Prediction Dataset and proceeds to data cleansing, where repetitive records are removed and missing BMI values are filled. Feature engineering is then carried out, including the removal of non-predictive identifiers, encoding of categorical features, and scaling of numerical features. The prepared dataset is then split into training and test sets with an 80/20 stratified split for preserving the class distribution. Logistic Regression and Random Forest algorithms are trained over the data and validated based on various measures of performance such as Accuracy, Precision, Recall, F1-score, ROC-AUC, and confusion matrices. Feature importance analysis from Random Forest is also performed to identify the most significant variables which are responsible for the prediction of stroke. The pipeline concludes with convergence to clinical inference, emphasizing the role of key predictors such as age, mean glucose value, and hypertension in the prediction of stroke risk. The process as a whole offer both valid predictive ability and readily interpretable outcomes.

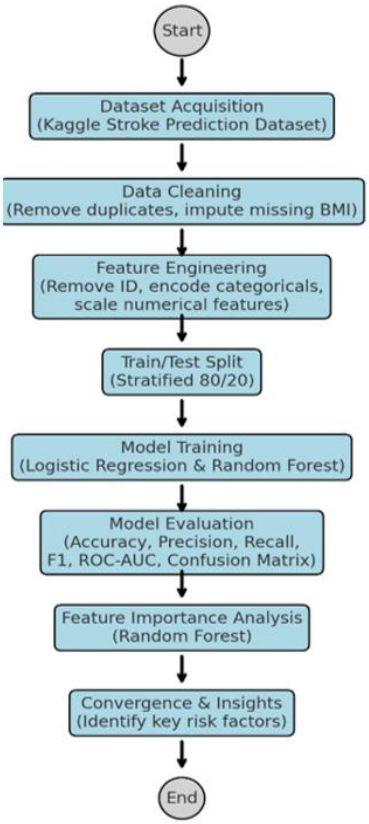

Figure 2: Flowchart of stroke prediction data processing and modeling.

2.2 Logistic Regression (LR)

Logistic Regression is a widely used classification algorithm, particularly in medical prediction tasks, due to its ability to estimate the probability of a binary outcome. In the context of this work, the probability of a patient experiencing a stroke ($y = 1$) given a feature vector $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ is modeled as:

$$P(y = 1 \mid \mathrm{x}) = \sigma(\mathrm{w}^{\mathrm{T}}\mathrm{x} + b), \qquad (1)$$

where:

- w is the vector of coefficients (weights) associated with the input features;
- b is the bias (intercept);
- $\sigma(z)$ is the logistic (sigmoid) function, defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \qquad (2)$$

The model predicts class $y = 1$ (stroke) if $P(y = 1 \mid \mathrm{x}) \geq \tau$, where τ is a threshold (commonly 0.5, but adjustable to optimize recall in imbalanced data).

The objective of logistic regression is to minimize the log-loss (cross-entropy loss):

$$\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m} \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})\right] \qquad (3)$$

where m is the number of samples, $y^{(i)}$ is the true label, and $\hat{y}^{(i)}$ is the predicted probability. Logistic Regression provides interpretable coefficients, enabling clinical insight into how risk factors such as age, BMI, glucose level, and hypertension influence the likelihood of stroke.

2.3 Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. The method introduces randomness by bootstrapping samples and selecting random subsets of features at each node split, thereby reducing variance and improving generalization.

The prediction of a Random Forest classifier for input x is:

$$\hat{y} = \mathrm{mode}\{h_t(\mathrm{x})\}_{t=1}^{T}, \qquad (4)$$

where:

- T is the total number of decision trees;
- $h_t(\mathrm{x})$ is the prediction of the t^{th} decision tree.

Each decision tree recursively partitions the feature space by selecting a feature j and threshold θ that maximizes an impurity reduction criterion, such as Gini Impurity or Entropy. For Gini Impurity, the splitting criterion is defined as:

$$G = 1 - \sum_{k=1}^{K} p_k^2, \qquad (5)$$

where p_k is the proportion of samples belonging to class k at a node, and K is the total number of classes (in this case, $K = 2$: stroke or non-stroke).

The final prediction is made through majority voting across all trees. Random Forest provides not only robust predictive performance but also feature importance scores, which quantify the contribution of each predictor variable to the overall classification. These scores can be used to identify the most influential stroke risk factors.

3 RESULTS

The analysis of the stroke prediction dataset provided interesting insights into statistical relationship between features and performance of machine learning models in prediction. The class distribution confirmed the imbalance of stroke vs. non-stroke, approximating real clinical prevalence. Stratified sampling and class weighting were applied to address this issue in training models.

Figure 3 presents the confusion matrix of the Logistic Regression model applied to the stroke dataset. Out of the non-stroke instances, 722 were correctly predicted, and 250 were incorrectly predicted as stroke. For stroke instances, the model correctly predicted 40 patients but incorrectly predicted 10 as non-stroke. The results show that while Logistic Regression is effective in discriminating the majority class, it is weak in stroke cases as seen through the moderate recall. This bias indicates the weakness of Logistic Regression in handling minority classes even when using class weights.

Figure 4 shows the confusion matrix for the Random Forest model. The classifier correctly classified 971 non-stroke cases and misclassified just 1 case and demonstrated its ability to correctly classify the majority class. The classifier misclassified all 50 cases of stroke as non-stroke, achieving zero true positive classification for the minority class. High overall accuracy achieved by Random Forest fails to indicate how severe class

imbalance is since the minority class is overwhelmed by the majority class during training. This indicates that more advanced balancing methods such as SMOTE or thresholding adjustment are required to improve recall.

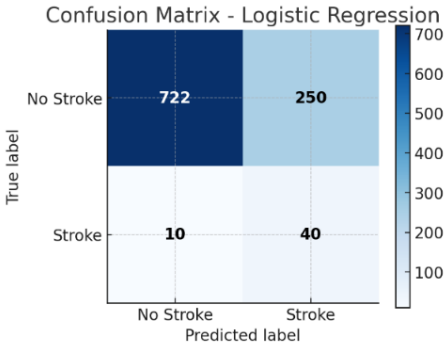

Figure 3: Logistic regression confusion matrix.

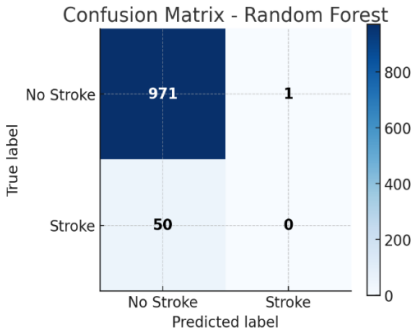

Figure 4: Confusion matrix for random forest.

Figure 5 shows the Receiver Operating Characteristic (ROC) curves for both Logistic Regression and Random Forest models are displayed. Logistic Regression has a curve close to the top-left part, indicating greater discriminative ability between stroke and non-stroke cases compared to Random Forest. The ROC-AUC scores confirm this, with Logistic Regression at 0.879 and Random Forest lagging behind at 0.916, which is because it is biased towards the majority class. The evaluation of the ROC curve shows that Logistic Regression has a higher sensitivity to specificity balance, while Random Forest performance is compromised by minority class misclassification.

Figure 6 presents the Precision–Recall curves for both models, which provide more insight into performance under class imbalance. Logistic Regression possesses higher precision at a wider range of values of recall compared to Random Forest, extremely low precision when there is more recall. This is indicative of the fact that Logistic Regression performs better in the recovery of true cases of stroke without including too many false positives. Random Forest, while highly skewed toward labeling non-stroke, causes its accuracy to collapse when scaled to detect stroke cases. The outcomes also advocate for the supremacy of Logistic Regression in dealing with imbalanced clinical data where the minority class is clinically most relevant.

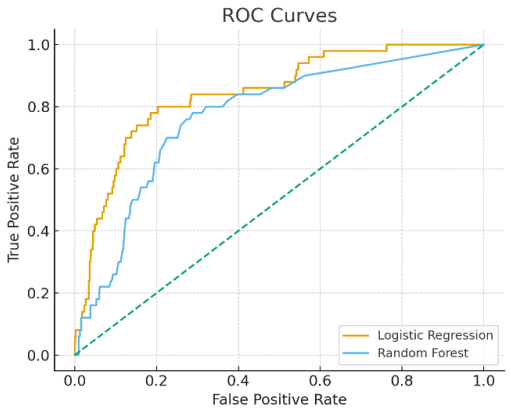

Figure 5: ROC Curves for logistic regression and random forest.

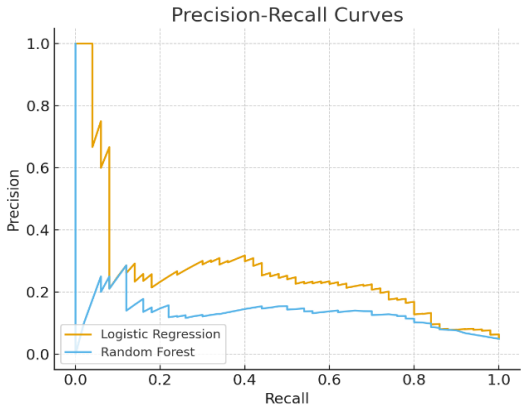

Figure 6: Precision–recall curves for logistic regression and random forest.

Figure 7 indicates the imbalance of the target variable distribution. The majority of the patients in the dataset belong to the non-stroke class, with fewer than 300 records representing stroke out of the total 5,110. This has a significant impact on model learning since classifiers prefer to focus on the majority class to optimize overall accuracy. The imbalance is why Random Forest failed to detect stroke cases effectively and why, although more balanced, Logistic Regression still yielded low recall. Knowing this imbalance is crucial to creating

methods such as resampling, cost-sensitive learning, or synthetic data generation to improve the ability of the model to detect high-risk stroke patients.

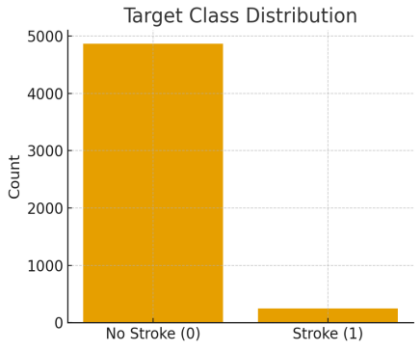

Figure 7: Target class distribution.

4 CONCLUSIONS

In this study, machine learning was used to analyze and predict stroke risk according to demographic, clinical, and lifestyle characteristics from the stroke risk prediction data. Correlation analysis revealed that the stroke risk factors are age, mean glucose, body mass index, hypertension, and heart disease, among which age was the most significant factor. The comparison of machine learning models showed that Logistic Regression provided a more balanced performance in stroke case detection compared to Random Forest, which obtained higher overall accuracy but failed to correctly classify positive stroke cases due to the excessive class imbalance of the data.

ROC analysis and Precision–Recall curves supported these findings, showing that Logistic Regression possessed greater discriminative power and more precise precision across various thresholds, though Random Forest still showed a preference towards the prediction of the majority class. Feature importance analysis also highlighted the influence of physiological and metabolic variables' interaction in deciding stroke vulnerability, highlighting the multifactorial nature of the disease.

The findings indicate that while high accuracy is possible when using ensemble models such as Random Forest, these require additional balancing methods such as oversampling, under sampling, or synthetic data generation in order to perform better with respect to minority cases. Logistic Regression, while simpler, was found to be robust against imbalance and offered higher interpretability, which is a key requirement in the context of medical decision-making.

This study emphasizes the potential of machine learning to facilitate early stroke risk detection and prevention. By integrating the use of predictive models into clinical workflows, healthcare professionals are able to better detect at-risk populations, enabling timely interventions and potentially mitigating the effects of stroke on patients and healthcare systems alike. Subsequent studies need to address the management of dataset imbalance with more advanced resampling strategies and hybrid or assembling methods that balance predictive accuracy with clinical interpretability.

ACKNOWLEDGMENTS

The authors are indeed grateful to universities for providing their students' academic data for the success of this research.

REFERENCES

[1] A. W. Nugroho, H. Arima, I. Miyazawa, T. Fujii, N. Miyamatsu, Y. Sugimoto, S. Nagata, M. Komori, N. Takashima, Y. Kita, et al., "The Association between Glomerular Filtration Rate Estimated on Admission and Acute Stroke Outcome: The Shiga Stroke Registry," Journal of Atherosclerosis and Thrombosis, vol. 25, pp. 570-579, 2018.

[2] M. Lee, J. L. Saver, K. H. Chang, H. W. Liao, S. C. Chang, and B. Ovbiagele, "Low glomerular filtration rate and risk of stroke: Meta-analysis," BMJ, vol. 341, c4249, 2010.

[3] C. H. Chao, C. L. Wu, and W. Y. Huang, "Association between estimated glomerular filtration rate and clinical outcomes in ischemic stroke patients with high-grade carotid artery stenosis," BMC Neurology, vol. 21, p. 124, 2021.

[4] A. M. Penn, N. S. Croteau, K. Votova, C. Sedgwick, R. F. Balshaw, S. B. Coutts, M. Penn, K. Blackwood, M. B. Bibok, V. Saly, et al., "Systolic blood pressure as a predictor of transient ischemic attack/minor stroke in emergency department patients under age 80: A prospective cohort study," BMC Neurology, vol. 19, p. 251, 2019.

[5] Y. Turana, J. Tengkawan, Y. C. Chia, M. Nathaniel, J. Wang, A. Sukonthasarn, C. Chen, H. V. Minh, P. Buranakitjaroen, J. Shin, et al., "Hypertension and stroke in Asia: A comprehensive review from HOPE Asia," Journal of Clinical Hypertension, vol. 23, pp. 513-521, 2021.

[6] R. Hajhosseiny, G. K. Matthews, and G. Y. Lip, "Metabolic syndrome, atrial fibrillation, and stroke: Tackling an emerging epidemic," Heart Rhythm, vol. 12, pp. 2332-2343, 2015.

[7] A. P. Carson, P. Muntner, B. M. Kissela, D. O. Kleindorfer, V. J. Howard, J. F. Meschia, L. S. Williams, R. J. Prineas, G. Howard, and M. M. Safford, "Association of Prediabetes and Diabetes with Stroke Symptoms," Diabetes Care, vol. 35, pp. 1845-1852, 2012.

[8] R. T. Ribeiro, M. P. Macedo, and J. F. Raposo, "HbA1c, Fructosamine, and Glycated Albumin in the Detection of Dysglycaemic Conditions," Current Diabetes Reviews, vol. 12, pp. 14-19, 2015.

[9] E. Selvin, A. M. Rawlings, P. L. Lutsey, N. Maruthur, J. S. Pankow, M. Steffes, and J. Coresh, "Fructosamine and Glycated Albumin and the Risk of Cardiovascular Outcomes and Death," Circulation, vol. 132, pp. 269-277, 2015.

[10] A. Grzywacz, A. Lubas, J. Smoszna, and S. Niemczyk, "Risk Factors Associated with All-Cause Death Among Dialysis Patients with Diabetes," Medical Science Monitor, vol. 27, e930152-1, 2021.

[11] B. Panwar, S. E. Judd, D. G. Warnock, W. M. McClellan, J. N. Booth, P. Muntner, and O. M. Gutiérrez, "Hemoglobin Concentration and Risk of Incident Stroke in Community-Living Adults," Stroke, vol. 47, pp. 2017-2024, 2016.

[12] M. Y. Kim, S. H. Jee, J. E. Yun, S. J. Baek, and D. C. Lee, "Hemoglobin Concentration and Risk of Cardiovascular Disease in Korean Men and Women-The Korean Heart Study," Journal of Korean Medical Science, vol. 28, p. 1316, 2013.

[13] S. V. Prabhu, B. Tripathi, Y. Agarwal, B. Kabi, and R. Kumar, "Association of serum calcium levels with clinical severity of ischemic stroke at the time of admission as defined by NIHSS score: A cross-sectional, observational study," Journal of Family Medicine and Primary Care, vol. 11, p. 6427, 2022.

[14] D. T. Dibaba, P. Xun, A. D. Fly, A. Bidulescu, C. L. Tsinovoi, S. E. Judd, L. A. McClure, M. Cushman, F. W. Unverzagt, and K. He, "Calcium Intake and Serum Calcium Level in Relation to the Risk of Ischemic Stroke: Findings from the REGARDS Study," Journal of Stroke, vol. 21, pp. 312-323, 2019.

[15] S. Rohrmann, H. Garmo, H. Malmström, N. Hammar, I. Jungner, G. Walldius, and M. V. Hemelrijck, "Association between serum calcium concentration and risk of incident and fatal cardiovascular disease in the prospective AMORIS study," Atherosclerosis, vol. 251, pp. 85-93, 2016.

[16] S. C. Larsson, S. Burgess, and K. Michaëlsson, "Association of Genetic Variants Related to Serum Calcium Levels with Coronary Artery Disease and Myocardial Infarction," JAMA, vol. 318, p. 371, 2017.

[17] L. Jahangiry, M. A. Farhangi, and F. Rezaei, "Framingham risk score for estimation of 10-years of cardiovascular diseases risk in patients with metabolic syndrome," Journal of Health, Population and Nutrition, vol. 36, p. 36, 2017.

[18] A. Arafa, Y. Kokubo, H. A. Sheerah, Y. Sakai, E. Watanabe, J. Li, K. Honda-Kohmo, M. Teramoto, R. Kashima, Y. M. Nakao, et al., "Developing a Stroke Risk Prediction Model Using Cardiovascular Risk Factors: The Suita Study," Cerebrovascular Diseases, vol. 51, pp. 323-330, 2022.

[19] Y. Miyamoto, T. Itaya, Y. Terasawa, and T. Kohriyama, "Association between the Suita Score and Stroke Recurrence in Patients with First-ever Ischemic Stroke: A Prospective Cohort Study," Internal Medicine, vol. 61, pp. 773-780, 2022.

[20] K. Nishimura, T. Okamura, M. Watanabe, M. Nakai, M. Takegami, A. Higashiyama, Y. Kokubo, A. Okayama, and Y. Miyamoto, "Predicting Coronary Heart Disease Using Risk Factor Categories for a Japanese Urban Population, and Comparison with the Framingham Risk Score: The Suita Study," Journal of Atherosclerosis and Thrombosis, vol. 21, pp. 784-798, 2014.

[21] A. Guzik and C. Bushnell, "Stroke Epidemiology and Risk Factor Management," CONTINUUM: Lifelong Learning in Neurology, vol. 23, pp. 15-39, 2017.

[22] F. Soriano, "Stroke Prediction Dataset," Kaggle, 2021, [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.