Singaporean Journal of Scientific Research(SJSR)
An International Journal (AIJ)
Vol.16.No.1 2024,Pp.48-55
ISSN: 1205-2421
available at :www.sjsronline.com
Paper Received : 04-09-2024 Paper Accepted: 05-12-2024
Paper Reviewed by: 1.Prof. Cheng Yu 2. Dr.Yarab Baig
Editor : Dr. Chen Du Fensidal

Enhancing Stroke Risk Prediction with AI: Utilizing the 2022BRFSS Dataset and Random Forest-Based Imputation

Revathi [1], S. Sathya [2]
Research Scholar [1], Assistant Professor [2],
Vels Institute of Science, Technology and Advanced Studies, Pallavaram,
Chennai, Tamil Nadu 600117, India nithrubanrevathi@gmail.com[1],
ssathya.scs@velsuniv.ac.in [2]

Abstract

Improving dataset reliability, model performance, and interoperability is crucial for advancing stroke risk prediction using AI in healthcare. This study leverages the CDC's newly released 2022 Behavioral Risk Factor Surveillance System (BRFSS) dataset to develop AI-based stroke risk prediction models. A novel Random Forest (RF)-based imputation technique addresses challenges related to missing data, enhancing dataset dependability. Six different AI models— Decision Tree (DT), Random Forest (RF), Gaussian Naive Bayes (GNB), Reboots, Adobos, and Convolutional Neural Networks (CNN)—are meticulously evaluated to identify the most promising approaches. Additionally, the study integrates high-performing models to further optimize predictive accuracy. These collaborative efforts highlight the transformative potential of AI in stroke risk assessment and its impact on improving healthcare outcomes.

Keywords:
Stroke Risk Prediction, Behavioral Risk Factor Surveillance System (BRFSS), Random Forest-Based Imputation, Hybrid Learning, Explainable AI (XAI), Residual Networks, AI in Healthcare

1. INTRODUCTION

Stroke is a leading cause of death and disability worldwide, placing a significant burden on healthcare systems [1]. Artificial Intelligence (AI) has the potential to revolutionize stroke prevention strategies by enhancing the accuracy of risk assessments, optimizing healthcare resource allocation, and improving patient outcomes. Machine learning and neural networks have been extensively utilized in stroke risk prediction research, incorporating various risk indicators, including genetic markers, to improve predictive accuracy [2][3][4]. However, challenges such as model generalization, data quality, and interpretability remain significant barriers to clinical adoption.

Current literature suggests that leveraging new datasets can unlock novel opportunities for AI in stroke risk prediction [5][6]. However, common issues like missing data, class imbalance, and the need for model comparison

complicate the development of effective predictive models. Addressing these challenges requires innovative preprocessing techniques and Explainable AI (XAI) approaches to enhance the interpretability of AI models, thus making them more clinicallyrelevant [7][8].

In this study, we utilize the 2022 BRFSS dataset from the Centers for Disease Control and Prevention (CDC) to explore AI-based stroke risk prediction. Comprehensive preprocessing strategies are applied to handle missing data and address class imbalance. Our approach includes Exploratory Data Analysis (EDA) to uncover key patterns and statistics in the dataset. We also implement a Machine Learning-based imputation technique using clean data insights to predict missing values for critical features.

Moreover, we evaluate the performance of various machine learning and deep learning models, such as Decision Trees (DT), Random Forests (RF), Gaussian Naive Bayes (GNB), Reboots, Adobos, and Convolutional Neural Networks (CNN). Model fusion techniques are also explored to enhance stroke risk prediction accuracy. To ensure clinical applicability, Explainable AI (XAI) methodologies are employed to identify and interpret the most influential input parameters, ultimately contributing to a robust and interpretable diagnostic framework.

2. LITERATURE REVIEW

Stroke risk prediction, along with the analysis of health behaviors using survey data, plays a vital role in improving healthcare outcomes. It enables early detection of individuals at high risk of stroke, facilitating targeted preventive interventions and personalized care strategies. Additionally, analyzing health-related behaviors provides insights that healthcare practitioners can leverage to tailor therapies and promote healthier lifestyles. Recognizing prevalent health issues among specific demographic groups aids in better resource allocation and healthcare policy formulation. Several studies have utilized the Behavioral Risk Factor Surveillance System (BRFSS) dataset to explore cardiovascular and stroke-related health risks:

Connie et al. [10] conducted an in-depth analysis using BRFSS data to investigate health disparities, incorporating COVID-19 variables into heart disease and stroke research.Ryan [11] utilized BRFSS data and government records to examine the relationship between law enforcement encounters and cardiovascular health, focusing on hypertension, diabetes, heart attacks, and strokes. Yashvanth et al. [12] applied machine learning techniques to BRFSS data for predicting conditions like diabetes, stroke, and hypertension, emphasizing the importance of data quality and optimal model selection. Chuan et al. [13] evaluated stroke risk models across diverse populations, highlighting the need to improve modeling approaches and include risk factors to address racial disparities in stroke predictions. Debora et al. [15] used logistic regression to analyze the prevalence of stroke risk factors in rural versus urban areas, exploring the impact of neighborhood socioeconomic status.

Marufuzzaman et al. [17] leveraged the Florida BRFSS dataset to estimate stroke prevalence and predictors among individuals with prediabetes and diabetes. These studies highlight the use of artificial intelligence (AI) and machine learning for cardiovascular and stroke prediction. However, critical research gaps remain, including handling data imbalances, improving methods for missing data imputation, incorporating Explainable AI (XAI) for deeper model insights, and conducting comprehensive model comparisons. Addressing these gaps is essential for enhancing the precision and reliability of stroke risk prediction models, ultimately contributing to more effective prevention and treatment strategies.

3. DATASET

This study utilizes the Behavioral Risk Factor Surveillance System (BRFSS) 2022 dataset, a comprehensive collection of health-related information gathered from surveys conducted across the United States. The dataset is particularly suited for analyzing stroke risk due to its extensive range of demographic, lifestyle, and health-related variables [11][12][14][16].

The primary objective of this research is to develop an AI model to predict stroke risk. The target variable for prediction is whether participants have ever been diagnosed with a stroke. Notably, there is a significant class imbalance, with a disproportionately larger number of records for healthy individuals compared to stroke cases.

Figure 1 illustrates the distribution of the target class, categorized by gender, age groups, and race/ethnicity. Initial findings suggest a higher prevalence of stroke among women and older adults. Additionally, data indicates that White Americans exhibit the highest stroke rates. This demographic analysis provides essential insights into the distribution of stroke incidents across various subgroups.

4. METHODOLOGY

Predicting stroke risk using survey data is crucial for the early identification of individuals at high risk, enabling timely interventions and preventive strategies [2][4][5][6]. Stroke is a severe condition that can lead to long-term disability and increased healthcare costs; however, many stroke risk factors are modifiable if detected early. Leveraging medical survey data allows for targeted strategies in education, awareness, and efficient healthcare resource allocation, ultimately reducing stroke rates and enhancing public health. This section outlines the comprehensive pipeline for AI-based stroke risk prediction using the BRFSS 2022 dataset, as depicted in **Figure 2.**

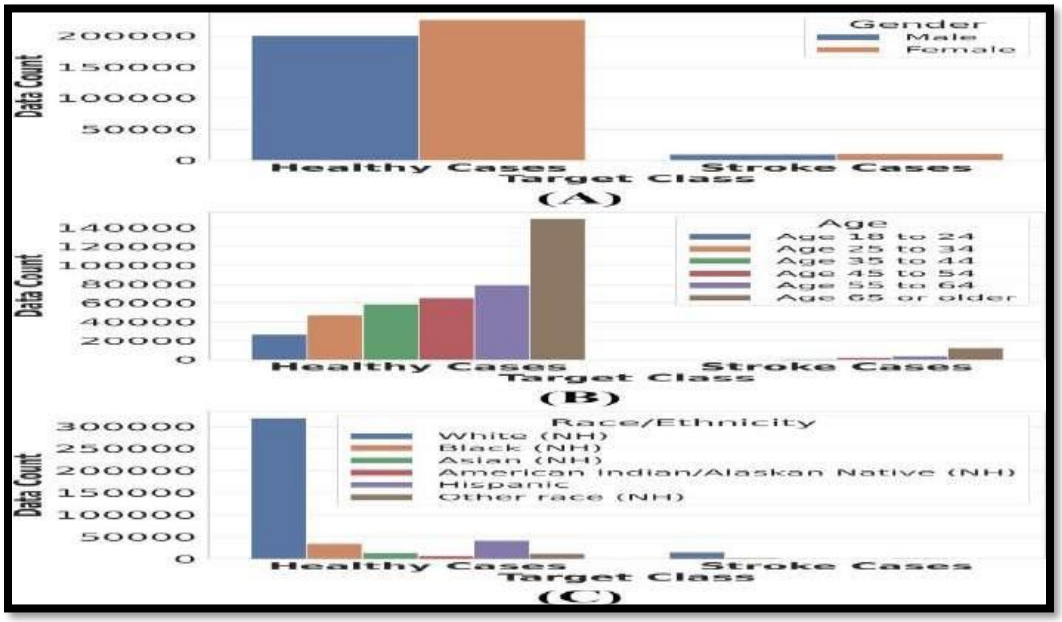

Figure 1: Representation of the target class based on various input factors such as gender, agegroup, and race or ethnicity.

5. Demographic Analysis in Stroke Risk Prediction

(A) Gender-Based Health Status Classification: This part of the research examines the distribution of participants by gender, differentiating between those who are healthy and those diagnosed with a stroke. Understanding gender-specific patterns is crucial for targeted stroke prevention strategies, as studies indicate potential differences in stroke incidence between males and females.

(B) Age Group and Health Status Analysis: The study further classifies participants by age, providing insights into how stroke risk varies across different age brackets. Age is a well- documented risk factor for stroke, and this analysis helps identify vulnerable age groups, enabling healthcare providers to focus on early interventions for older populations.

(C) Racial and Ethnic Disparities in Stroke: This section investigates stroke prevalence among different racial and ethnic groups, distinguishing between healthy individuals and those with a history of stroke. By analyzing disparities in stroke cases across demographics, the study highlights key risk factors linked to race and ethnicity, aiding in the development of culturally sensitive healthcare policies.

6. Feature Selection for Stroke Risk Prediction

Feature selection plays a crucial role in improving the predictive accuracy and efficiency of machine learning models. By identifying the most important variables and removing redundant or irrelevant ones, feature selection enhances the model's performance, reduces computational complexity, and ultimately leads to more accurate stroke risk assessments from survey data [16,17]. In this study, we carefully selected features that are most relevant to stroke risk prediction from the BRFSS dataset, which contains a vast array of variables.

Selected Input Features

The dataset used in this study includes data across multiple domains, and approximately 300 input characteristics are available for analysis. However, for effective stroke risk prediction, 40 critical features were selected from seven distinct data domains, as shown in **Table I**:

TABLE I: Selected Input Features Across Data Domains

Data Domain	Feature Description
Social and DemographicFactors	Marital status, Residential status, Military record, Ethnicity (Race),State, Spoken languages, Gender, Children count, Age group
Socio-economicStatus	Literacy status, Employment status, Earning level (Income), Mobileusage
Medical History	Skin cancer history, Other cancer history, Chronic bronchitis history, Depressive disorder history, Renal disease history, Diabetes history, Arthritis history, Myocardial infarction history, Angina history, Asthmahistory, Self-rated health status, BMI level
Disability Status	Difficulty in seeing, Difficulty in dressing or bathing, Difficulty inwalking, Difficulty in doing errands alone
HealthcareServices	Health insurance, Doctor affordability, Routine checkup, Availability ofpersonal health care assistant
Personal Health Behavior	Alcohol consumption, Tobacco consumption, Smoking level, Exerciselevel
VaccinationHistory	HIV status, Vaccination records

These selected features were chosen based on their relevance to stroke risk prediction, ensuring that the models built from this data are both robust and applicable. The selection of these featuresallows for more focused analysis of factors that directly contribute to stroke risk, enabling better-targeted interventions.

7. Explainable AI for Feature Relevance: SHAP

To better understand the contribution of each feature in stroke risk prediction, we used **SHapley Additive exPlanations (SHAP)** [22, 23], a popular technique for interpretability in machine learning models. SHAP values help in identifying the significance of each feature by quantifying its impact on the model's output. In this study, the SHAP summary plot is utilized to highlight the 24 most significant variables, providing a graphical representation of feature importance. This approach allows us to gain insights into which factors are most influential in predicting stroke risk,ensuring that the models are transparent and interpretable.

The SHAP analysis enhances the explainability of the AI models used in this study, promoting trust and allowing healthcare professionals to better understand the reasoning behind stroke risk predictions. By visualizing feature importance, this approach also aids in identifying key risk factors that require further investigation or intervention in real-world healthcare settings.

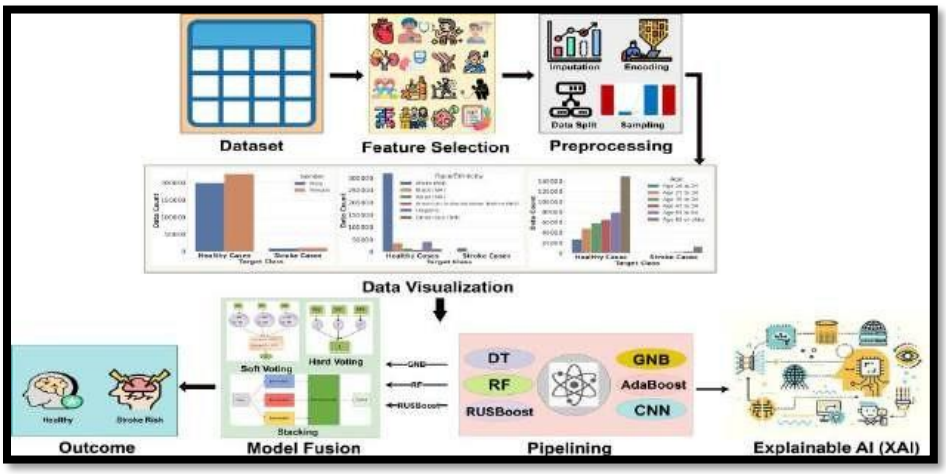

Figure 2: depicts the whole pathway for predicting stroke risk from medicalsurvey data.

This section includes all of the steps required to conduct extensive studies and comparisons of the various methodologies for stroke risk prediction.

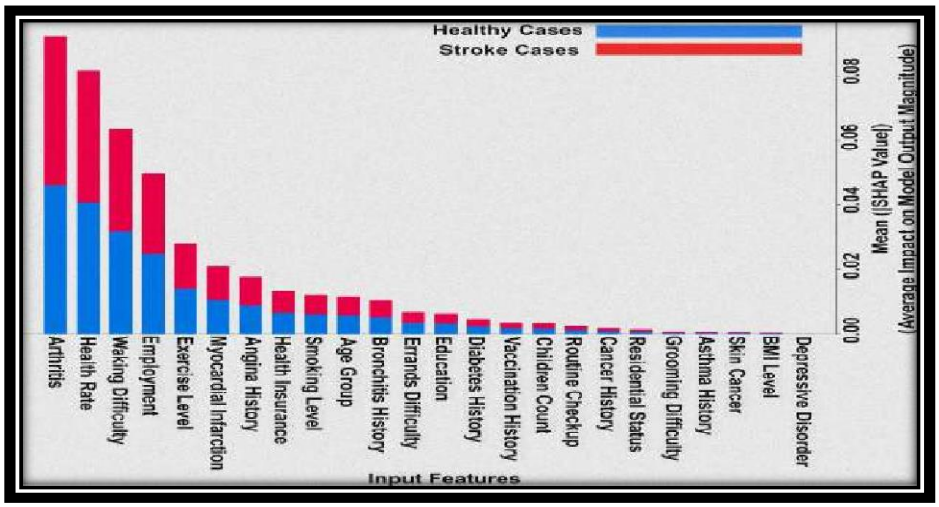

Figure 3: illustrates the value of deep insightful features using the SHapley AdditiveExplanations (SHAP).

8. Preprocessing

Survey data collected via phone interviews often suffer from significant missing values [11] [14]. Proper handling of this missing data is critical for reliable analysis and accurate stroke risk prediction. Imputation techniques, specifically Random Forest-based imputation, are employed in this study to address missing values in the dataset. The target variable is subsequently divided into different classes for stroke risk prediction—healthy participants, stroke patients, those who refused to provide responses, and individuals unaware of their health status. The dataset is further split intotraining and testing sets to facilitate model evaluation.

To overcome challenges such as class imbalance, with healthy cases significantly outnumbering stroke cases, the Synthetic Minority Over-Sampling Technique (SMOTE) is applied to balance the dataset. This ensures that the predictive models receive adequate representation of both classes during training. Once preprocessing steps are

completed, the data is ready for model input.

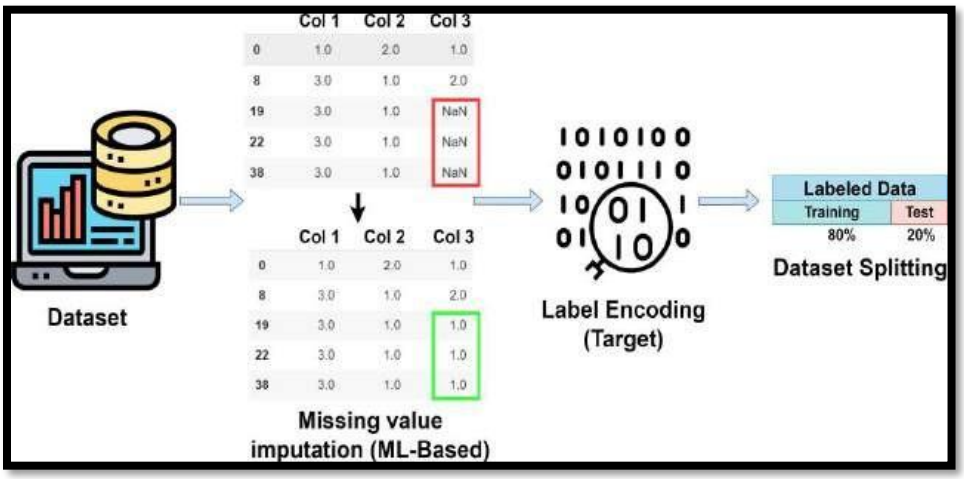

Figure 4: The preprocessing pipeline involves several stages

1. **Handling Missing Data**: Random Forest-based imputation is used to predict and fill in missingvalues for selected features.
2. **Class Imbalance**: SMOTE is applied to balance the target classes (healthy vs. stroke patients) inthe training dataset.
3. **Data Splitting**: The dataset is divided into training and test sets for model validation andevaluation.

II. Modeling Approach and Performance Metrics

Various machine learning and deep learning models are utilized for stroke risk prediction, including traditional methods such as Decision Trees (DT), ensemble techniques like Random Forest (RF) and Adaptive Boosting (AdaBoost), and more advanced approaches such as the 1D Convolutional Neural Network (CNN) with residual networks. Ensemble methods such as soft and hard voting, as well as stacking with Logistic Regression (LR) as a meta-classifier, are also explored.

The table below provides a summary of model performance, showing accuracy, precision, recall, and F1 score for each technique. The models are ranked based on their F1 score, which provides a balance between precision and recall:

Table II: Model Performance Evaluation

Modeling Approach	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score
Stacking (GNB, RF, RusBoost), LR-based	84.25%	84.64%	0.94	0.85	0.88
Random Under-sampling Boosting	83.90%	83.90%	0.94	0.84	0.88
Hard Voting (GNB, RF, RusBoost)	81.98%	79.33%	0.94	0.79	0.85
Random Forest	80.30%	79.29%	0.94	0.79	0.85
Soft Voting (GNB, RF, RusBoost)	74.11%	65.33%	0.94	0.65	0.75
Gaussian Naive Bayes	95.68%	94.50%	0.94	0.61	0.72
1D-CNN (Residual Network-based)	94.19%	91.85%	0.54	0.57	0.56
Decision Tree	99.99%	91.38%	0.54	0.55	0.54
Adaptive Boosting (AdaBoost)	71.73%	61.12%	0.57	0.53	0.54

Experiment and Results Analysis

In this section, the results of the various stroke risk prediction models are compared to assess their predictive performance. Traditional machine learning models such as **Decision Trees** (DT), **Random Forest** (RF), and **Gaussian Naive Bayes** (GNB) are evaluated alongside more sophisticated ensemble methods and deep learning models.

The performance metrics, including precision, recall, F1 score, and accuracy, provide a comprehensive understanding of the models' strengths and weaknesses. It is essential to focus not only on overall accuracy but also on the ability of the models to correctly identify stroke patients (positive class) while minimizing false negatives, which is crucial for stroke risk prediction.

Area Under the Receiver Operating Characteristic (AUC) Curve

The **AUC** curve is also computed to assess the models' ability to distinguish between healthy and stroke patients across various discrimination thresholds. **Fig. 5** presents the AUC curves for the top-performing models, highlighting the superior discriminative capabilities of ensemble models such as stacking and random under-sampling boosting, as compared to classical classifiers like **Gaussian Naive Bayes** and **Random Forest**.

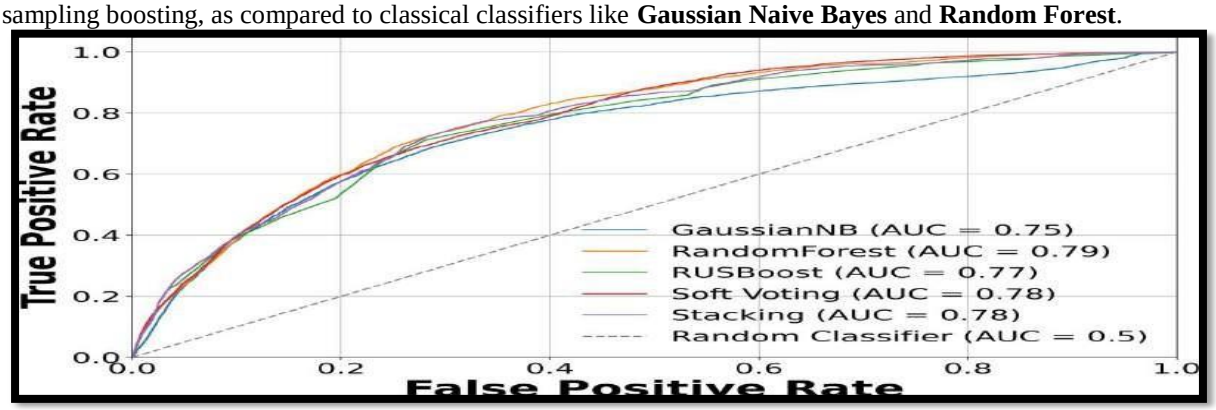

Figure 5: depicts the area under the receiver operating characteristic (AUC) curve for theindicated models.

Overall, the ensemble methods—particularly **stacking** and **random under-sampling boosting**— demonstrate the most promising performance, with higher F1 scores and better ability to discriminate between stroke cases and healthy individuals. These findings underline the potential of ensemble and hybrid models in improving stroke risk prediction, making them a strong candidate for further clinical application.

9. Conclusion

This study, using the CDC's 2022 BRFSS dataset, significantly advances stroke risk prediction. By combining machine learning and ensemble methods, it improves prediction accuracy and provides insights into early stroke detection. The fusion models, which combine multiple classifiers, outperformed traditional models like Random Forest and Gaussian Naive Bayes, highlighting their better predictive ability. The careful selection of important features from demographic, health, and lifestyle data was crucial for improving model performance. Additionally, Explainable AI (XAI) methods helped identify key factors contributing to stroke risk, allowing for better understanding and decision-making. Ultimately, the study emphasizes the value of early detection and treatment in preventing strokes, which can help reduce healthcare costs and improve patient outcomes. This research paves the way for more effective, data-driven approaches to stroke prevention and public health strategies.

References

[1] Tore´n, K., Neitzel, R.L., Eriksson, H.P., et al., "Occupational exposure to noise and dust in Swedish soft paper mills and mortality from ischemic heart disease and ischemic stroke: a cohort study," *Int Arch Occup Environ Health*, 96, 965–972 (2023).

[2] Sirsat, M. S., Ferme, E., Caˆmara, J., "Machine Learning for Brain Stroke: A Review," *Journal of Stroke and Cerebrovascular Diseases*, Volume 29, Issue 10, 2020, 105162, ISSN 1052-3057.

[3] Chandrabhatla, A. S., Kuo, E. A., Sokolowski, J. D., et al., "Artificial Intelligence and Machine Learning in the Diagnosis and Management of Stroke: A Narrative Review of United States Food and Drug Administration-Approved Technologies," *Journal of Clinical Medicine*, 10.3390/jcm12113755, 12, 11, (3755), (2023).

[4] Arafa, A., Kokubo, Y., Sheerah, H. A., et al., "Developing a Stroke Risk Prediction Model Using Cardiovascular Risk Factors: The Suita Study," *Cerebrovasc Dis*, 51 (3): 323–330 (2022).

[5] Biswas, N., Mohi Uddin, K. M., Rikta, S. T., et al., "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthcare Analytics*, Volume 2, 2022, 100116, ISSN 2772-4425.

[6] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI," *Information Fusion*, Volume 58, 2020, Pages 82-115, ISSN 1566-2535.

[7] Nazar, M., Alam, M. M., Yafi, E., et al., "A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques," *IEEE Access*, vol. 9, pp. 153316-153348, 2021.

[8] Tsao, C. W., Aday, A. W., Almarzooq, Z. I., et al., "Heart Disease and Stroke Statistics—2023 Update: A Report from the American Heart Association," *Circulation*, 147, 8, e93-621 (2023).

[9] Talbert, R.D., "Lethal Police Encounters and Cardiovascular Health among Black Americans," *J. Racial and Ethnic Health Disparities*, 10, 1756–1767 (2023).

[10] Yashvanth, R., Rehan, M., Kodipalli, A., et al., "Diabetes, Hypertension, and Stroke Prediction Using Computational Algorithms," *2023 World Conference on Communication & Computing (WCONF)*, RAIPUR, India, 2023, pp.1-5.

[11] Hong, C., Pencina, M. J., Wojdyla, D. M., et al., "Predictive Accuracy of Stroke Risk Prediction Models Across Black and White Race, Sex, and Age Groups," *JAMA*, 329(4):306–317(2023).

[12] Das, M. C., et al., "A comparative study of machine learning approaches for heart stroke prediction," *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*, Istanbul, Turkiye, 2023, pp. 1-6.

[13] Mukaz, K., Dawson, D., Howard, V. J., et al., "Rural/urban differences in the prevalence of stroke risk factors: a cross-sectional analysis from the REGARDS study," *J Rural Health*, 38:668–673 (2022).

[14] Tsao, C. W., Aday, A. W., Almarzooq, Z. I., et al., "Heart Disease and Stroke Statistics— 2022 Update: A Report from the American Heart Association," *Circulation*, 145, 8, e153-e639 (2022).

[15] Khan, M. M., Roberson, S., Reid, K., et al., "Prevalence and predictors of stroke among individuals with prediabetes and diabetes in Florida," *BMC Public Health*, 22, 243 (2022).

[16] Tran, P. M., Tran, L. T., Zhu, C., et al., "Rural Residence and Antihypertensive Medication Use in US Stroke Survivors," *Journal of the American Heart Association*, 11, 15, e026678 (2022).

[17] "Building risk prediction models for daily use of marijuana using machine learning techniques," *Drug and Alcohol Dependence*, 225, 108789, 2021, 0376-8716.

[18] Banerjee, D., Singh, J., "Prediction of Stroke Risk Factors for Better Pre-emptive Healthcare: A Public-Survey-Based Approach," in *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing, vol 1199. Springer, Singapore, 2021.

[19] Daneshvar, N. H., Masoudi-Sobhanzadeh, Y., & Omidi, Y., "A voting-based machine learning approach for classifying biological and clinical datasets," *BMC Bioinformatics*, 24, 140 (2023).

[20] Mohapatra, S., Maneesha, S., Mohanty, S., et al., "A stacking classifiers model for detecting heart irregularities and predicting Cardiovascular Disease," *Healthcare Analytics*, Volume 3, 2023,100133, ISSN 2772-4425.

[21] Sun, J., Sun, C.-K., Tang, Y.-X., et al., "Application of SHAP for Explainable Machine Learning on Age-Based Subgrouping Mammography Questionnaire Data for Positive Mammography Prediction and Risk Factor Identification," *Healthcare*, 2023, 11, 2000.

[22] Kessler, R., Philipp, J., Wilfer, J., et al., "Predictive Attributes for Developing Long COVID— A Study Using Machine Learning and Real-World Data from Primary Care Physicians in Germany," *J. Clin. Med.*, 12, 3511 (2023).