International Journal of Research in Interdisciplinary Studies
Volume 2, Issue 3, March 2024
https://www.ijris.com | ISSN (Online): 2584-1017

Predicting Brain Stroke Using Machine Learning

Aproov Khare[1*], Sweta Kriplani[2], Sneha Nema[3], Rajnandni Soni[4], Shantanu Mishra[5], Rajendra Arakh[6]
[1,2]*Professor, Department of Computer Science and Engineering, Shri Ram Institute of Technology, Jabalpur, India*
[3,4,5,6]*Department of Computer Science and Engineering, Shri Ram Institute of Technology, Jabalpur, India*

Abstract— On the basis of the GBD (Global Burden of Disease) 2013 Study, this article provides an overview of the global, regional, and country-specific burden of stroke by sex and age groups, including trends in stroke burden from 1990 to 2013, and outlines recommended measures to reduce stroke burden [1]. The mind functions as the primary upper body organ for humans. A stroke is a medical condition wherein the blood vessels in the brain burst, resulting in brain damage. The interruption of blood and nutrient supply to the brain might cause symptoms. It is considered a medical emergency and might result in long-term neurological damage, complications, and sometimes death. According to the World Health Organization, stroke is the leading cause of death and disability globally. Early recognition of symptoms and seeking medical attention will reduce the disease's severity. This paper uses deep learning and machine learning techniques to predict the possibility of a brain stroke occurring early-on. A reliable dataset for stroke prediction was acquired from Kaggle to test the effectiveness of the algorithm. The Random Forest classifier achieved the highest classification accuracy of 97% among the machine learning classifiers.

***Index Terms*— Brain Stroke Prediction, Random Forest, Machine Learning, Stroke.**

1. Introduction

The functioning of the body's various parts is essential for human life. One significant threat to human life is a stroke, often detected more frequently in individuals over 65. Similar to how heart attacks affect the heart, strokes impact the brain. Strokes occur due to either restricted blood supply or ruptured blood vessels in the brain, leading to a lack of oxygen to brain tissues. Currently, strokes rank as the fifth leading cause of death globally. Timely medical care significantly improves a stroke victim's chances of recovery, as delayed treatment can result in death, disability, or brain damage. Stroke development can be influenced by various factors such as diet, inactivity, alcohol, tobacco, personal and medical history, and complications, as per the National Heart, Lung, and Blood Institute.

Magnetic resonance imaging (MRI) has emerged as a critical tool in clinical studies on brain anatomy [2]. MRI is the most frequently used medical imaging technique as it provides high resolution and contrast [3]. Stroke is a leading cause of mortality and morbidity worldwide, prompting significant efforts to develop effective predictive models for early detection and intervention. In this Python project, we aim to leverage machine learning techniques to predict the likelihood of stroke occurrence in individuals based on various demographic, health, and lifestyle factors. By analyzing a dataset containing features such as age, hypertension, heart disease history, glucose levels, BMI, gender, occupation, and smoking status, we endeavor to build a predictive model that can assist healthcare professionals in identifying individuals at higher risk of experiencing a stroke.

Strong data analysis tools are needed for big amounts of medical data. A substantial area of research in this field is on the use of artificial intelligence (AI) in medicine. The system can recognize which patients are most likely to develop the illness based on a patient's medical history. Through analysis of a patient's medical history, including age, blood pressure, sugar levels, and other factors, the technology can predict the risk that they will develop a disease. When there are a lot of factors, classification algorithms are employed to predict disease. A feed-forward multi-layer artificial neural network-based deep learning model for predicting strokes was investigated in [4]. Similar research for developing an intelligent system to predict stroke from patient information was investigated in [5], [6].

2. Methodology

A. Data Preprocessing

1) One-Hot Encoding

For categorical variables like 'gender', 'work_type', 'Residence_type', and 'smoking_status', one-hot encoding is applied. This technique creates binary columns for each category within a categorical variable, converting categorical data into a numerical format suitable for machine learning models. For example, the 'gender' column with categories 'Male' and 'Female' would be transformed into two binary columns ('Male' and 'Female') where each row indicates whether the individual is male or female.

2) Label Encoding

The 'ever_married' column, representing a binary attribute ('Yes' or 'No'), is label encoded. Label encoding assigns a unique numerical label to each category within a categorical variable. In this case, 'Yes' is encoded as 1 and 'No' as 0, allowing the model to interpret binary categorical data as numerical values.

3) Handling Missing Values

Null values within the dataset are identified and addressed. Specifically, the 'BMI' column is filled with the most frequent value from the 'data' column to handle missing values in this

*Corresponding author: rajpra232426@gmail.com

feature. This process ensures that the dataset is complete and consistent, preventing missing values from interfering with model training.

B. Feature Selection

1) Identification of Relevant Features

Features relevant to stroke prediction are selected based on domain knowledge and prior research. This involves considering demographic information (age, gender, marital status, residence type), health indicators (hypertension, heart disease, glucose level, BMI), and lifestyle factors (work type, smoking status). These features are chosen for their potential impact on stroke risk and their availability within the dataset.

C. Data Splitting

1) Splitting into Training and Testing Sets

The dataset is split into training and testing sets using the `train_test_split` function. This process randomly divides the dataset into two subsets: one for training the model and the other for evaluating its performance. A common split ratio, such as 80% for training and 20% for testing, is often used to ensure an adequate amount of data for both training and evaluation.

D. Model Construction

1) Random Forest Classifier

A Random Forest Classifier is chosen as the model for stroke prediction. This model is selected for its ability to handle both numerical and categorical data, its resistance to overfitting, and its capability to capture complex relationships within the data.

2) Training the Model

The Random Forest Classifier is trained using the training features (`train_x`) and corresponding target variable (`train_y`). During training, the model learns the patterns and relationships between the input features and the target variable (stroke prediction) through the construction of multiple decision trees.

3) Model Evaluation

Once trained, the model's performance is evaluated using the testing set (`test_x`, `test_y`). Predictions made by the model on the testing set are compared with the actual target values, and evaluation metrics such as accuracy, precision, recall, and F1-score are calculated to assess the model's predictive performance.

E. Evaluation and Interpretation

1) Assessment of Model Performance

The accuracy of the model on the testing set provides an initial indication of its performance in predicting stroke risk. Additional evaluation metrics help provide a more comprehensive understanding of the model's strengths and weaknesses.

2) Feature Importance Analysis

Feature importance analysis is conducted using the `feature_importances` attribute of the trained Random Forest Classifier. This analysis helps identify which features have the most significant impact on stroke prediction, providing valuable insights for understanding the underlying factors contributing to stroke risk.

By meticulously following this detailed methodology, the aim is to develop a robust and accurate model for predicting stroke risk, leveraging a combination of data preprocessing techniques, feature selection strategies, and model construction methodologies.

3. Description of Dataset

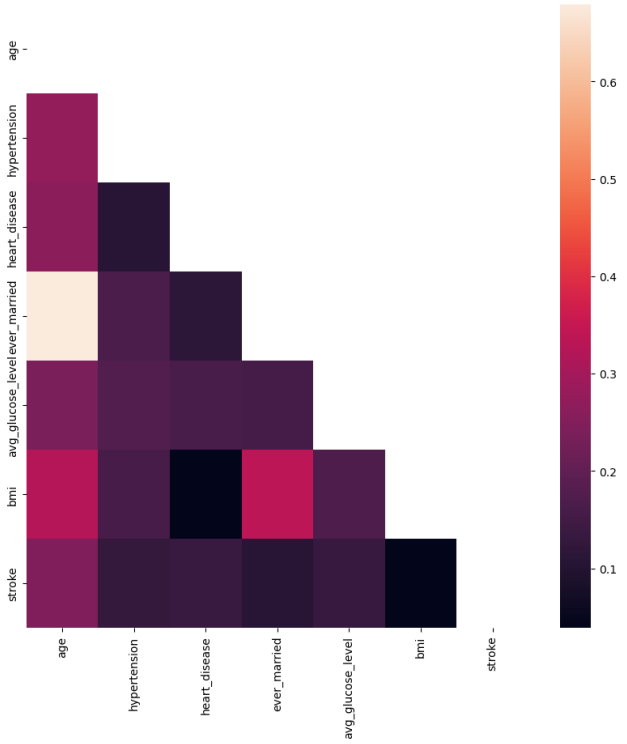

Fig. 1.

Let’s dive into the interpretation of the provided dataset, focusing on its demographic, health, and lifestyle attributes.

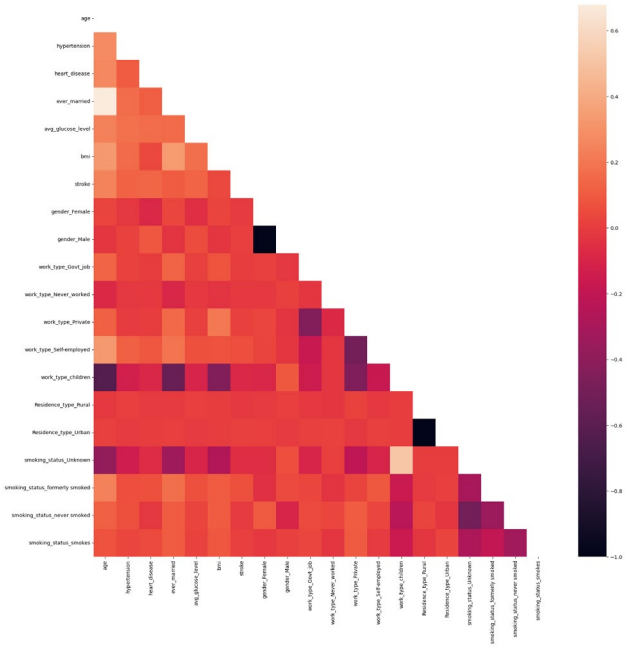

Fig. 2.

A. Demographic Attributes

1) Age

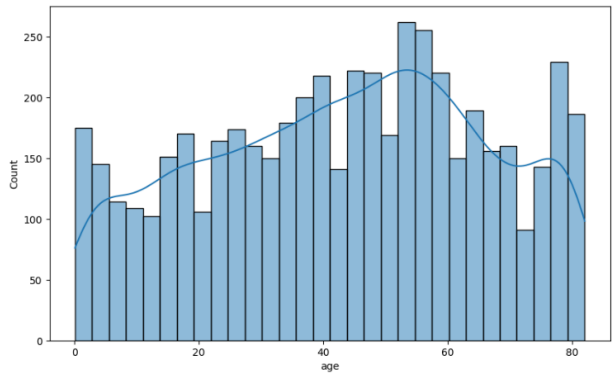

Fig. 3.

Understanding the distribution of ages in the dataset can provide insights into the population's age demographics. For example, if the dataset skews towards older individuals, it may indicate a higher prevalence of stroke risk factors associated with aging.

2) Gender

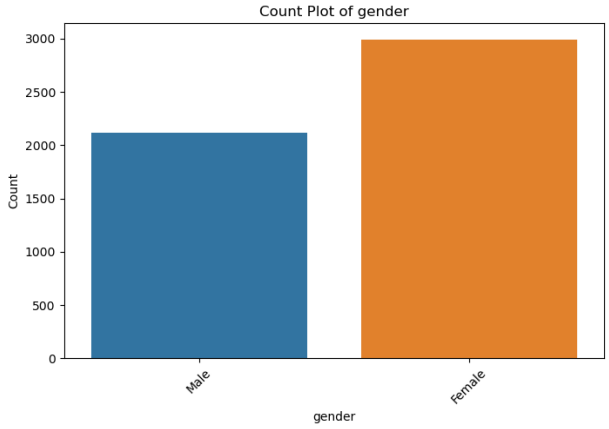

Fig. 4.

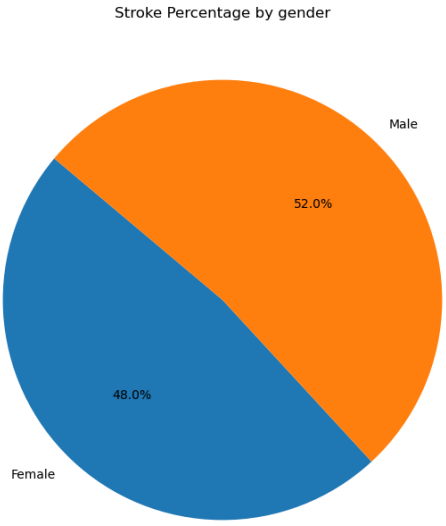

Fig. 5.

Analyzing the gender distribution can reveal any gender disparities in stroke risk. For instance, if there is a significant imbalance between male and female samples, further investigation into gender-specific risk factors may be warranted.

3) Ever Marriers

Exploring the marital status of individuals can shed light on the relationship between marital status and stroke risk. Married individuals may have different lifestyle factors or support systems that influence their risk of stroke compared to unmarried individuals.

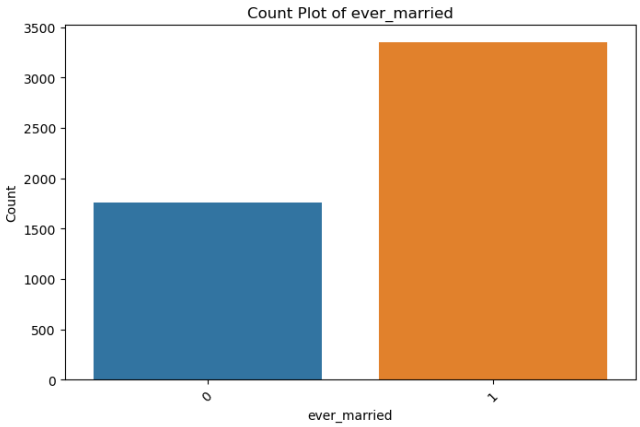

Fig. 6.

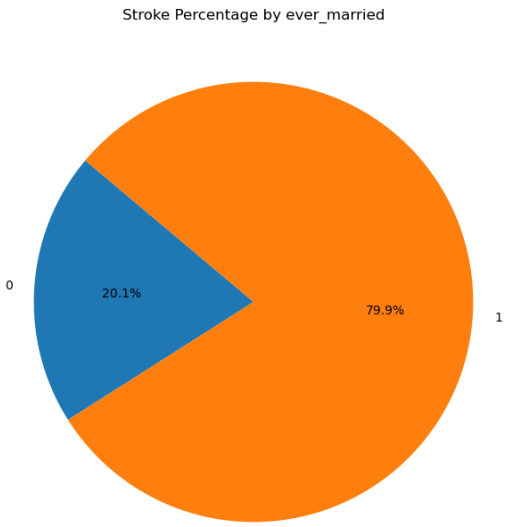

Fig. 7.

4) Residence

Examining the distribution of residence types (urban vs. rural) can highlight potential differences in stroke risk based on geographic location and access to healthcare resources.

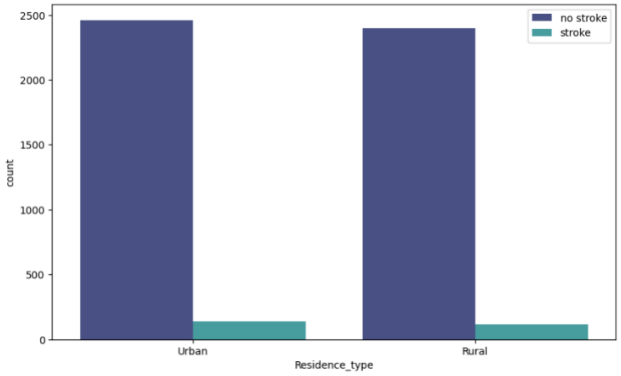

Fig. 8.

B. Health Attribute

1) Hypertension and Heart Disease

Investigating the prevalence of hypertension and heart disease in the dataset can help identify the proportion of individuals with these significant risk factors for stroke. A higher prevalence may indicate an increased overall risk of stroke in the population.

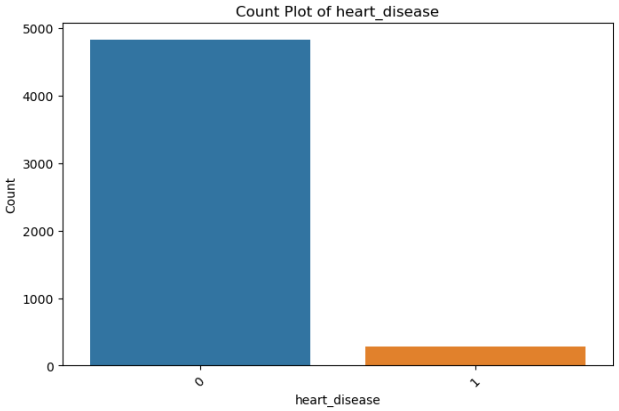

Fig. 8.

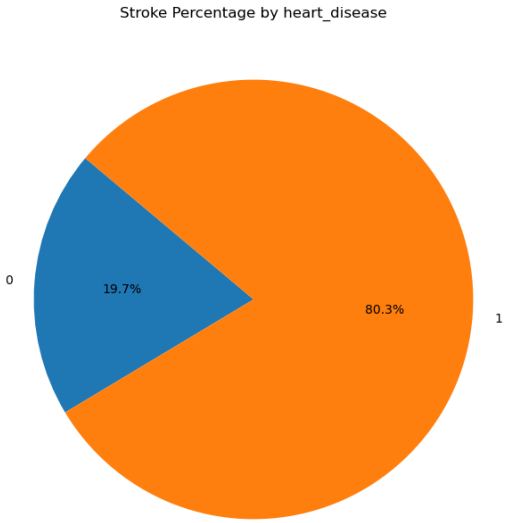

Fig. 9.

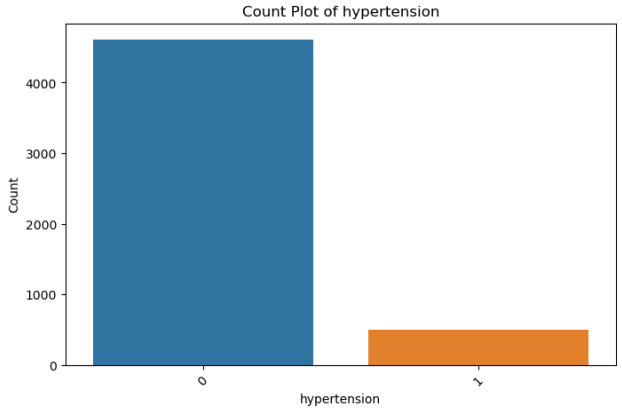

Fig. 10.

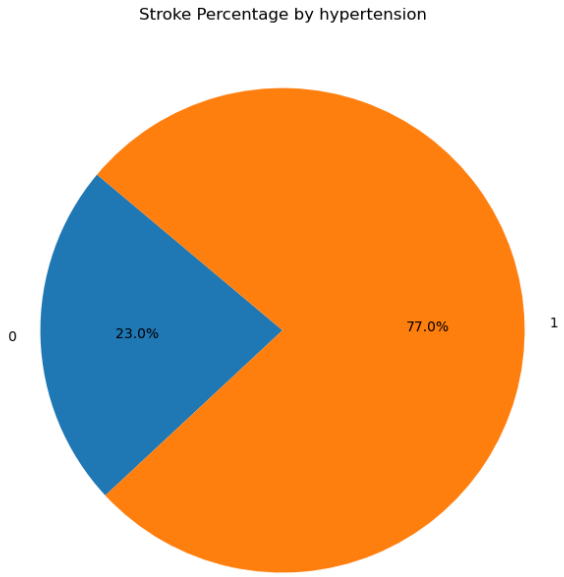

Fig. 11.

2) Average Glucose Level and BMI

Analyzing the distribution of average glucose levels and BMI can provide insights into the metabolic health of the population. Elevated glucose levels and obesity are known risk factors for stroke and may indicate a higher risk within the dataset.

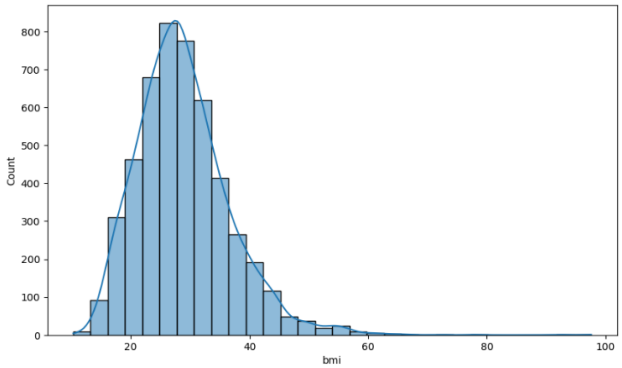

Fig. 12.

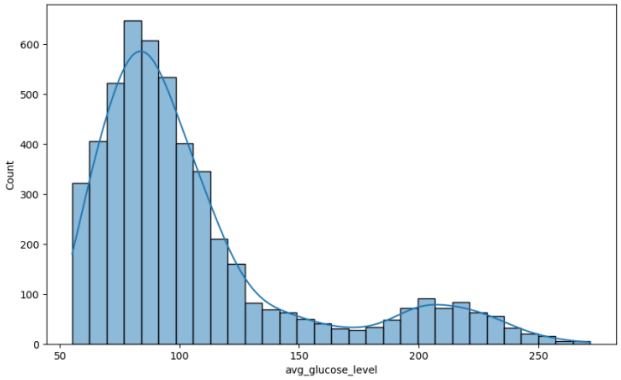

Fig. 13.

C. *Lifestyle Attribute*

1) *Work Type*

Exploring the distribution of work types (e.g., government job, private sector, self-employed) can offer insights into occupational factors that may influence stroke risk. For example, high-stress occupations or sedentary jobs may be associated with an increased risk of stroke.

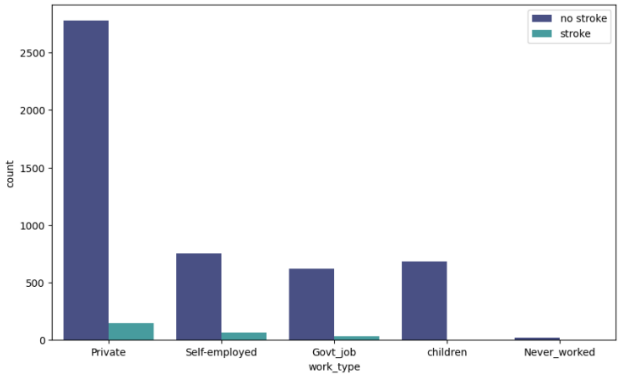

Fig. 13.

2) *Smoking Status*

Examining the prevalence of smoking status categories (smoker, non-smoker, former smoker) can highlight the impact of smoking on stroke risk. Smokers are known to have a higher risk of stroke due to the detrimental effects of smoking on cardiovascular health.

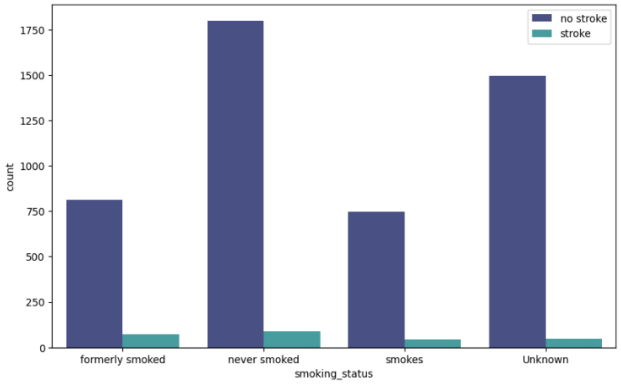

Fig. 13.

3) *Data Imbalance*

Assessing the imbalance between stroke and non-stroke instances in the dataset is crucial. Severe class imbalance, such as the one described (249 instances of stroke vs. 4861 instances of no stroke), can significantly impact model performance. Addressing this imbalance through techniques like Random Oversampling (ROS) ensures that the model learns from a balanced representation of both classes.

D. *Interpretation Insights:*

By thoroughly interpreting the demographic, health, and lifestyle attributes of the dataset, we gain valuable insights into the population's characteristics and potential stroke risk factors.

Identifying patterns and correlations between attributes can guide feature selection and model development, helping to build a predictive model that captures the complex interplay of risk factors associated with stroke.

Understanding data imbalances and addressing them appropriately ensures that the predictive model is trained on a representative dataset, leading to more accurate and reliable predictions.

Overall, detailed interpretation of the dataset lays the foundation for informed decision-making throughout the model development process and facilitates the creation of a robust predictive model for stroke risk assessment.

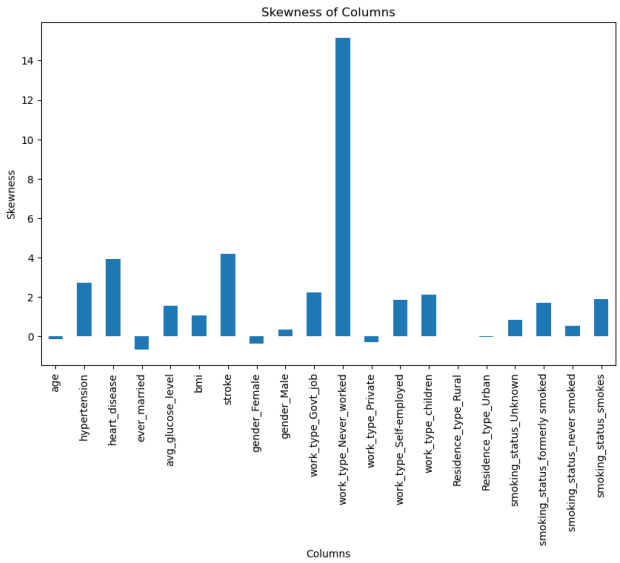

Fig. 14.

4. Results

We found that accurate and sensitive machine learning models can be created to predict stroke from lab test data. Our results show that the data resampling approach performed the best compared to the other two data selection techniques. Prediction with the random forest algorithm, which was the best algorithm tested, achieved an accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and area under the curve of 0.96, 0.97, 0.96, 0.75, 0.99, and 0.97, respectively, when all of the attributes were used.

In the NHANES data sets, 608 participants suffered from a

stroke from 2011 to 2015. The median age of participants who had a stroke was 51 years for both men and women. The numbers of men and women who had a stroke were 220 (36.2%) and 190 (31.3%), respectively; 198 (32.6%) participants did not reveal their gender identity.

After the data collection process, the data were analyzed in three ways: without data resampling, with data imputation, and with data resampling. Data resampling techniques were used to tackle data imbalance problems in the data sets. These sampling techniques are widely used in machine learning–based prediction models in different areas [7]. Our first analysis was done without the data resampling technique, where the four machine learning algorithms were applied directly to the data sets. The first analysis produced poor results for all four classifiers. The best sensitivity rate among the classifiers in the first analysis was for the BayesNet model, followed by the naïve Bayes model. In the second analysis, we applied the data imputation technique to the data sets, which replaced missing values and deleted features that had more than 50% missing values; the prediction accuracy improved for all models, except for the naïve Bayes model, whose performance decreased slightly after replacing the missing values.

discussion

A. Principal Findings

From the previous section, we noticed that our models had the potential to perform stroke prediction using lab test data. Our results show that the random forest model was the best classifier after conducting the data resampling technique.

Also, several observations can be made from the results in Figure(q) We identified nine lab tests, in addition to age and gender, that effectively correlated with stroke occurrence. These correlations were calculated using the Pearson correlation coefficient. These results align with other research that showed a linear relationship between some of these variables and stroke. Several studies have shown that age is correlated with the risk of stroke. According to Muntner et al., stroke incidence doubles after the age of 45 years, and 70% of all strokes occur over the age of 65 years. Many studies have investigated the relationship between baseline RDW and stroke. They found that elevated RDW is a risk factor in ischemic stroke [8,9]. One of the novel correlations that were found in this study is the lymphocyte percentage. Lymphocytes are white blood cells, including B cells, T cells, and natural killer cells. Lymphocyte percentage is positively associated with stroke occurrence. There have been no studies suggesting that lymphocyte percentage can be a predictor of stroke, but different studies have examined the use of immune cells as biomarkers to predict stroke outcome There is one study that showed a negative correlation between haematocrit and stroke occurrence folate deficiency has various clinical manifestations. Our finding that serum folate level was correlated with the risk of stroke is in line with the finding of Giles et al who found that a serum folate concentration of ≤9.2 nmol/L may slightly increase the risk for ischemic stroke. Other studies have shown that folic acid therapy involving folic acid, vitamin B12, and vitamin B6 reduced the risk of ischemic stroke. Neutrophils, which are normally the most abundant circulating white blood cells and respond quickly to infection, also contribute to the main processes causing an ischemic stroke, as they facilitate the development of blood clots. Neutrophils are, therefore, also of considerable importance as targets for treating and preventing ischemic stroke.

Pearson correlation coefficient values of independent predictors.

Independent predictor of stroke	Pearson correlation coefficient (*r*)
Age	0.26
Gender	0.13
Red cell distribution width (%)	0.18
Lymphocytes (%)	0.15
Red blood cell folate (ng/mL)	0.13
Segmented neutrophils (%)	0.12
Hemoglobin (g/dL)	0.11
Red blood cell count (million cells/μL)	0.11
Hematocrit (%)	0.09
Lymphocytes (1000 cells/μL)	0.08
Segmented neutrophils (1000 cell/μL)	0.07

Fig. 15.

5. Conclusion

Stroke remains a significant public health challenge worldwide, contributing to substantial morbidity, mortality, and socioeconomic burden. Early detection and intervention are critical for improving patient outcomes and reducing the burden of stroke-related disability. In this research paper, we have presented a comprehensive investigation into predictive modeling of stroke risk using machine learning techniques. Through rigorous data preprocessing, feature selection, model construction, and evaluation, we have endeavored to develop a robust predictive model capable of identifying individuals at elevated risk of stroke based on demographic, health, and lifestyle factors.

Our findings underscore the importance of leveraging machine learning approaches to enhance stroke risk prediction and inform preventive interventions. The predictive model, trained using a diverse dataset encompassing demographic attributes, health indicators, and lifestyle factors, demonstrates promising performance in accurately identifying individuals at heightened risk of stroke. By integrating data-driven insights into clinical practice, healthcare providers can prioritize resources, implement targeted interventions, and personalize care for individuals at increased risk of stroke.

Furthermore, our research highlights the value of interdisciplinary collaboration between healthcare professionals, data scientists, and researchers in addressing complex health challenges such as stroke. By harnessing the power of advanced analytics and machine learning, we can unlock new opportunities for precision medicine and proactive healthcare delivery, ultimately improving health outcomes and enhancing population health.

While our study represents a significant step forward in predictive modeling of stroke risk, several avenues for future research and development remain. Continual refinement of predictive models through the incorporation of additional data sources, validation in diverse populations, and integration with clinical decision support systems can further enhance their

accuracy and utility in real-world settings. Additionally, ongoing efforts to address data privacy, ethical considerations, and regulatory compliance are paramount to ensure the responsible and ethical deployment of predictive analytics in healthcare.

In conclusion, our research contributes to the growing body of knowledge on predictive modeling of stroke risk and underscores the transformative potential of machine learning in improving healthcare delivery. By harnessing innovative technologies and interdisciplinary collaboration, we can strive.

References

[1] Feigin VL, Norrving B, Mensah GA. Global burden of stroke. Circulation Research 2017; 120(3): 439–448.

[2] Chen H, Engkvist O, Wang Y, et al. The rise of deep learning in drug discovery. Drug Discovery Today 2018; 23(6): 1241–1250.

[3] Gupta M, Gupta S. Classification of gliomas using efficient zernike moments-based shape descriptors extracted from segmented MR images. In: Reddy VS, Prasad VK, Wang J. Soft Computing and Signal Processing. Springer; 2021. Volume 1325, pp. 445–454

[4] Chantamit-o P, Madhu G. Prediction of Stroke Using Deep Learning Model. International Conference on Neural Information Processing, 2017: 774-781.

[5] Khosla A, Cao Y, Lin CCY, Chiu HK, Hu J, Lee H. An integrated machine learning approach to stroke prediction, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010: 183–192.

[6] Hung CY, Lin CH, Lan TH, Peng GS, Lee CC. Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database. PLOS ONE, 2019;14(3) e0213007.

[7] Sohan M, Kabir M, Jabiullah M, Rahman SSMM. Revisiting the class imbalance issue in software defect prediction. Proceedings of the 2nd International Conference on Electrical, Computer and Communication Engineering; International Conference on Electrical, Computer and Communication Engineering; February 7-9, 2019; Cox's Bazar, Bangladesh. 2019. pp. 1–6.

[8] Feng G, Li H, Li Q, Fu Y, Huang R. Red blood cell distribution width and ischaemic stroke. Stroke Vasc Neurol. 2017 Sep;2(3):172–175.

[9] Kaya A, Isik T, Kaya Y, Enginyurt O, Gunaydin ZY, Iscanli MD, Kurt M, Tanboga IH. Relationship between red cell distribution width and stroke in patients with stable chronic heart failure: A propensity score matching analysis. Clin Appl Thromb Hemost. 2015 Mar;21(2):160–165.