Machine Learning Based Early Detection for Brain Stroke

Abdullah Khan, Hasnain Abbas Sherazi, Nimra Saleem, and Waqas Tariq Toor

2024 Horizons of Information Technology and Engineering (HITE) | 979-8-3315-1605-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/HITE63532.2024.10777174

***Abstract*—Brain attack or stroke is one of the major causes of illness and death on a global level; it is important to detect it at an early stage to deal with it on time and save lives. Our study shows how machine learning can be used in the prediction of brain strokes by using a dataset of some common clinical features. Our model predicts stroke with approximately 80% accuracy by using traditional logistics regression. Analysis of features shows that age, gender, different diseases, and smoking status are significant in the prediction of brain stroke. Our outcomes reveal how machine learning is helping in medicine to find patients with high risk and comfort early interventions. This approach can assist clinicians in making better decisions and improving patient treatment.**

***Index Terms*—Brain Stroke, Machine Learning, Prediction, Data-Driven Approach, Clinical Decision Support.**

I. Introduction

The project focuses on predicting the chances that the patient will have a stroke or not by using clinical data like age, gender, different diseases, and smoking status. The data used for the project, named 'stroke prediction dataset', was taken from Kaggle. It is a classification problem. Where '1' indicates high chances of stroke and '0' indicates no chance of stroke. Supervised learning methods were applied when training the model.

According to the World Health Organization (WHO), the reason for stroke is the breakup of continuous blood flow to a specific part of the brain. Which results in the depletion of oxygen and nutrients in the cells and finally leads to their death. It is considered an alarming situation in medicine that requires immediate attention. On the other hand, if we detect it at an early stage and make a proper treatment, further damage to that affected part of the brain can be minimized. WHO reports state that at the global level, at least 15 million people deal with stroke every year, and the number of effected individuals who die due to this condition is at the rate of one in every 4–5 minutes.

Two main types of strokes are ischemic and hemorrhagic strokes. Ischemic strokes occurs when flow of blood is blocked due to the formation of clots. On the otherhand, hemorrhagic strokes result from the rupture of weak blood vessels, causing bleeding in the brain. Prevention of strokes includes in embracing a healthy lifestyle, which demands the avoiding of detrimental habits like smoking and excessive intake of alcohol, managing body mass index (BMI), and ensuring optimal glucose levels. Predicting stroke occurences is crucial for implementing timely inventions in order to pervent time lasting damage or fatalities. This study incorporates hypertension, BMI levels, heart disease and average optimal glucose levels as predictive features for brain strokes. Additionally, machine learning shows a considerable promise in enhancing the decision making within this predictive framework. [1]. Govindarajan et al [2] In their research, they implemented different machine learning methods to differentiate stroke disorders. They used a combination of machine learning classifiers and text mining techniques and collected the data from 507 different patients. They used different machine learning models, like artificial neural networks (ANN). The stochastic gradient descent (DDG) algorithm has the highest accuracy of 95%.

Earlier literature has examined various aspects associated with stroke prediction. Jeena et al. [3] conducted a study mainly concentrating on various risk factors to calculate the likelihood of stroke occurence. They applied a regression based technique to clarify the relation among each factor and it's effect on the stroke incidence. Adam et al. [4] organize a research employing both the decision tree method and k-nearest neighbour algorithm to predict stroke occurence. Their findings implies that the decision tree method was favoured by medical professionals for this motive.

Various studies have utilized different datasets and methodologies to predict strokes in individuals. Singh and Choudhary [5] utilized the Cardiovascular Health Study (CHS) dataset for their predictive model. Emon et al. [1] utilized learning-based classification algorithms, such as XGBoost, Random Forest, Naive Bayes, Logistic Regression, and Decision Tree, on a dataset sourced from Kaggle which has limitations. We acknowledge potential biases and gaps in representation. To better serve diverse populations, future research must prioritize inclusive, high-quality data. Kansadub et al. [6] explored stroke probability through decision trees, neural networks, and Naive Bayes analysis, while assessing precision and the Area Under the Curve (AUC) in their investigation. Tazin et al. [7] suggested early-stage stroke prediction employing Logistic Regression (LR), Decision Tree (DT) Classification, Random Forest (RF) Classification, and a Voting Classifier, with Random Forest demonstrating superior per-

A. Khan, H. A. Sherazi, N. Saleem are students at the Department of Electrical Engineering, University of Engineering and Technology Lahore, Narowal Campus, Pakistan (e-mail: {ak4169732, Syedhasnainsherazi86, ns7058043}@gmail.com).

W. T. Toor is Associate Professor at the Department of Electrical Engineering, University of Engineering and Technology Lahore, Narowal Campus, Pakistan 51600 (e-mail: drwaqas.toor@uet.edu.pk).

formance.Chetan Sharma et al. [8] utilized the supervised algorithm Random Forest on an openly accessible dataset to forecast stroke incidence. Additionally, they investigated a feed-forward multi-layer artificial neural network-based deep learning model for stroke prediction.. Similar research aimed at developing intelligent systems for stroke prediction using patient records was conducted by others (Authors et al., Year). Hung et al. [9] evaluated machine learning and deep learning models for constructing stroke prediction models based on electronic medical claims databases. Fang et al. [10] applied current Deep Learning (DL) approaches, such as CNN, LSTM, and ResNet, and compared them with traditional machine learning algorithms for clinical prediction. Mahesh et al. [7] employed various DL algorithms, including CNN, DenseNet, and VGG16, to automatically predict brain strokes.

II. Methodology/Model

"This study employed a machine learning approach to predict brain stroke using a dataset comprising clinical and demographic features of patients. The methodology consisted of data preprocessing, feature engineering, model selection, and evaluation. First, the dataset was cleaned and preprocessed by handling missing values and normalizing the data. Then, relevant features were extracted and engineered to enhance the model's predictive power. Following that, various machine learning algorithms, including logistic regression, decision trees, random forest, and support vector machines, were utilized and compared to identify the best-performing model. Different matrices, like the F1 score, recall, precision, and accuracy, were used to determine the impact of the model. The model with the best performance was selected, and to check its validity for the prediction of stroke holdout method, it was used. By using machine learning techniques, this study focuses on the development of a model for stroke prediction to assist clinicians in the detection of high-risk patients for early treatment and prevention of brain stroke.

A. Initial Setup

Importing Libraries: The first cell imports required libraries like NumPy for linear algebra operations, Pandas for data manipulation and analysis, Seaborn and Matplotlib for plots and visualization of data, and different Scikit Learn tools for machine learning purpose.

Reading the Data: The dataset is loaded using Pandas, and basic information about the dataset is displayed using df.info().

B. Data Cleaning and Exploration

- Cleaning: The 'id' column is dropped from the dataset, indicating an initial step in data cleaning to remove unnecessary information.
- Basic Data Inspection: The notebook displays the first few rows of the dataset using df.head(10), providing a glimpse into the actual data entries.
- Dataframe Shape and Statistics: The shape of the dataframe is printed, and basic statistical measures are obtained using df.describe(). These steps help in understanding the distribution of data and its various statistical properties.
- Null Value Check: The notebook includes a check for null values in the dataset using df.isnull().sum(). Managing null values is a crucial stage in prepping the dataset for machine learning models.
- The next steps would likely include more detailed data visualization, feature engineering, checking correlations, data splitting, and applying machine learning models.
- Handling Missing Values.
- Filling Null Values: The notebook addresses missing values in the 'bmi' column by filling them with the mean value of the 'bmi'. This is a common approach to handle missing data without dropping entire rows or columns.

C. Visualization

- Class Distribution Plot: A count plot is used to visualize the distribution of the target variable 'stroke'. This is essential to understand the balance or imbalance in the dataset with respect to the target classes shown in Fig. 1. Here The graph displays the count of individuals with

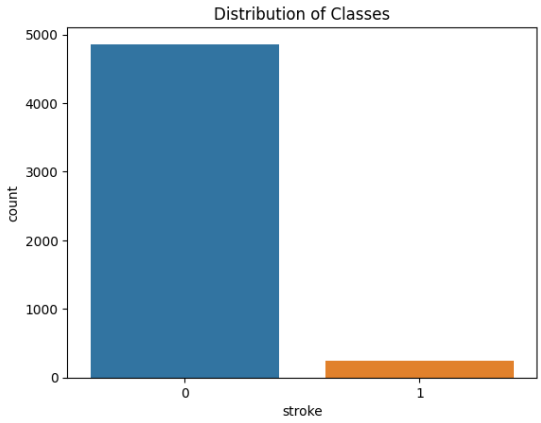

Fig. 1. Class Distribution

and without a stroke. The x-axis represents stroke status (0 for no stroke, 1 for stroke), while the y-axis shows the count. Over 5000 individuals are in the "with no stroke" category, whereas fewer than 1000 individuals fall into the "with stroke" category. The legend indicates that orange corresponds to "with no stroke," and green corresponds to "with stroke."
- Boxplots for Numerical Features: Boxplots are created for the numerical features, excluding certain categorical features. This visualization is crucial for identifying outliers and understanding the spread of each numerical feature shown in Fig. 2. The graph displays six numerical features. These features include age, hypertension, heart disease, average-glucose-level, BMI, and stroke. Each boxplot provides information about the distribution of

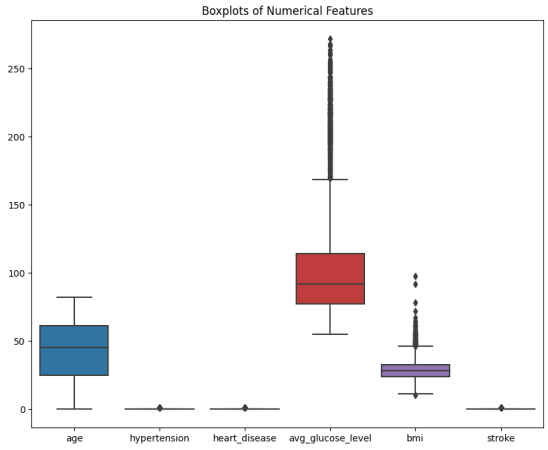

Fig. 2. Boxplots

data. The median age is around 50, while hypertension, heart disease, and stroke appear to be binary categorical data. The average glucose level has a median around 100, with an outlier extending up to 250. BMI's median is approximately 30, with several outliers above it.

- We used different methods to handle imbalance in data including smote oversampling, ensemble and underfitting but underfitting gave us the optimum results.

- Correlation Heatmap: The notebook includes a heatmap to display the correlation between different numerical features. This is a valuable step in understanding how different features relate to each other, which can inform feature selection and engineering shown in Fig. 3. The

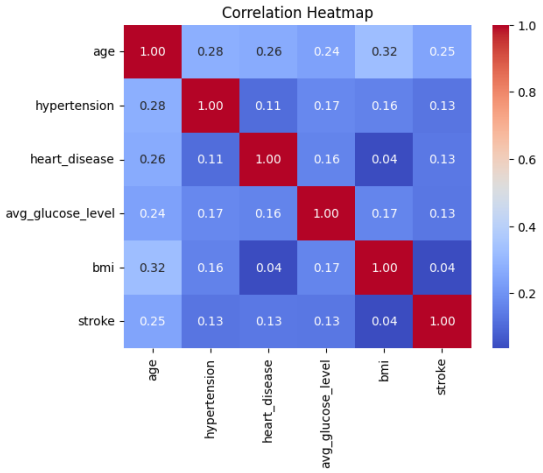

Fig. 3. Corelation Heatmap

correlation heatmap shows how different features like age, hypertension, heart disease, average glucose level, body mass index (BMI), and stroke relate to each other. Each cell in the heatmap represents how much two variables are correlated. Darker shades of red denote stronger positive correlations, whereas darker shades of blue signify stronger negative correlations. Notable findings include a positive correlation between age and BMI (approximately 0.32) and between average glucose level and BMI (about 0.17).

- Feature Engineering and Model Building.

- The next steps likely involve feature engineering, such as encoding categorical variables and scaling numerical values. This is followed by splitting the data, where data is partitioned into training and testing sets, followed by the application of various machine learning models to identify the most precise one.

- Advanced Data Visualization.

- Boxplots by Stroke Status: Custom boxplots are created for various features (age, hypertension, BMI, avg-glucose-level) against the stroke status. This step is crucial for visualizing the distribution of these features in patients with and without stroke, highlighting potential risk factors shown in Fig. 4.

  Each boxplot provides information about the distribution of data. The median age is around 50, while hypertension, heart disease, and stroke appear to be binary categorical data. The average glucose level has a median around 100, with an outlier extending up to 250. BMI's median is approximately 30, with several outliers above it.
- Handling outliers in Fig. 5.

D. Feature Engineering

- Encoding Categorical Variables: The notebook demonstrates the use of Label Encoding to convert categorical variables into a format readable by machines. This phase is indispensable for prepping the dataset for machine learning algorithms since the majority of algorithms mandate numerical input.
- Scaling Numerical Features: The MinMaxScaler from Scikit-Learn is employed for the scaling of numerical characteristics. Scaling is critical in ensuring that features with larger ranges do not dominate those with smaller ranges in the model training process.

E. Checking for Relation

The project explores the relationship between the 'stroke' target variable and various features like 'age', 'avg-glucose-level', and 'bmi'. This is done by computing descriptive statistics for these features grouped by stroke status. Understanding these relationships is crucial for feature selection and gaining insights into the factors that may influence the likelihood of a stroke Shown in Fig. 6,7,8,9,10,11 and 12

F. Data Consistency

Dataframe Inspection: The notebook includes repeated inspections of the dataframe (dfcopy.head() and df.head()), likely to ensure the consistency and correctness of the data after each transformation step.

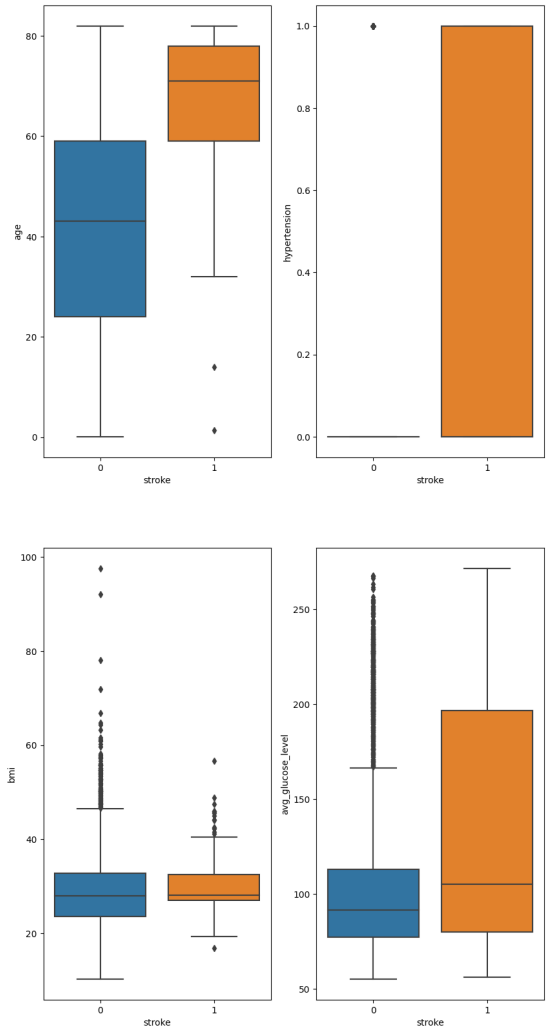

Fig. 4. Custom Boxplots

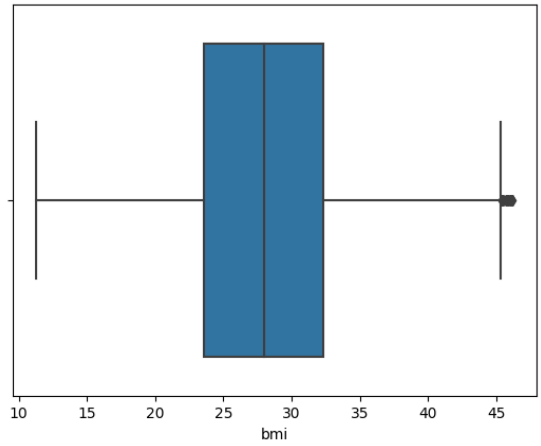

Fig. 5. Outliers

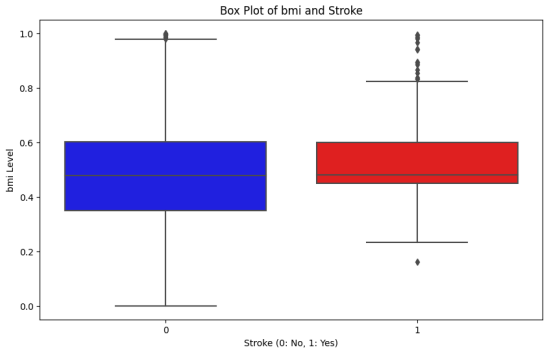

Fig. 6. Boxplot of BMI and Stroke

G. Explore Data Analysis (EDA)

Exploring Unique Classes: There is an exploration of unique classes in the 'gender' column, both in the original and copied data frame. This step might be a part of an exploratory data analysis to understand the diversity and distribution of categorical variables.

H. Model Building and Evaluation

The remaining parts of the notebook are expected to focus on dividing the dataset into training and testing subsets and subsequently employing machine learning models, and evaluating their performance. This is a critical part of any machine learning project as it determines the efficacy of the model in making predictions.

1) Detailed Visualization:

- Gender and Stroke: The notebook presents count plots showing the distribution of stroke among different genders.
- Age Distribution: A detailed box and strip plot for the age distribution, grouped by stroke status, is created.
- Hypertension and Stroke: A count plot is used to visualize the distribution of stroke among patients with and without hypertension.
- Model Training and PredictionModel Application: The notebook has used decision tree, random forest, SVM and logistic regression.
- Overfitting Check: There is a section that checks for overfitting by comparing the model's performance on training and test sets. A good practice to ensure the model generalizes well.

2) Model Evaluation:

- Confusion Matrix: The notebook includes a confusion matrix as a tool for estimating the performance of the model. This is key to understand the quantities of true positives, true negatives, false positives, and false negatives.
- Performance Metrics: Various performance metrics like classification accuracy and error are calculated. These metrics are essential for evaluating the effectiveness of the model.
- Predicted Probabilities: The notebook includes an exploration of predicted probabilities for different thresholds, which is an advanced technique for optimizing the model's decision threshold based on specific requirements

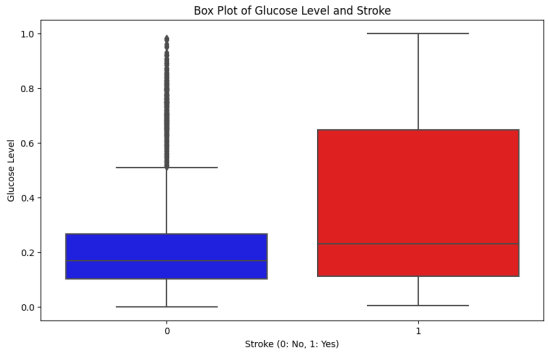

Fig. 7. Boxplot of Gulucose and Stroke

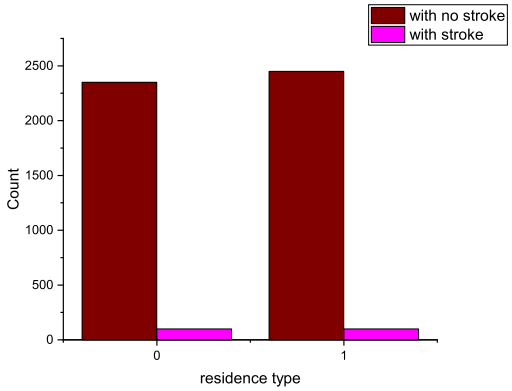

Fig. 8. Distribution of Stroke with respect to Residence type

or constraints.

III. Results

Precision denotes the proportion of correctly predicted strokes (true positives) relative to all positive predictions (comprising both true positives and false positives). For class 0 (no stroke), the precision is roughly 0.84, or 84%. For class 1 (stroke), the precision is approximately 0.74, or 74%.

Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions to all actual positive instances, encompassing true positives and false negatives. For class 0, the recall is about 0.66, or 66%. For class 1, the recall is approximately 0.88, or 88%.

The F1-score, as it combines precision and recall into a single metric, provides a balanced evaluation of both. For class 0, the F1-score is around 0.74, or 74%. For class 1, the F1-score is approximately 0.81, or 81%. The accuracy metric assesses the overall correctness of predictions across both classes. The model achieves an accuracy of about 0.78, which is 78%.

The macro average computes the average of precision, recall, and F1-score across both classes. The macro average precision is roughly 0.79, equivalent to 79%. The macro average recall stands at approximately 0.77, or 77%. Similarly, the macro average F1-score is approximately 0.77, also 77%.

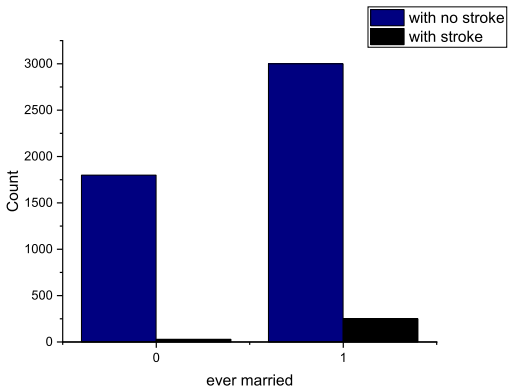

Fig. 9. Distribution of stroke among married and unmarried patients

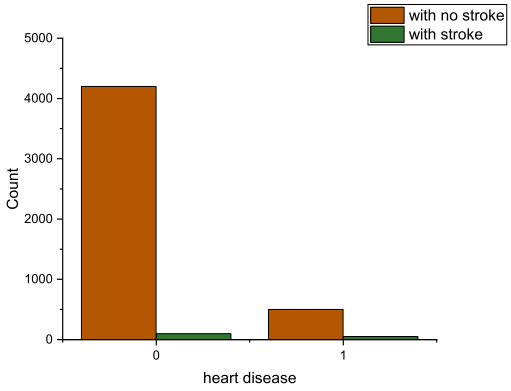

Fig. 10. Distribution of stroke among patient with heart disease and without heart disease

The weighted average considers class imbalance by giving weights to the metrics of each class depending on the number of instances. The weighted average precision, recall, and F1-score are approximately 0.79, which is 79%.

A. Interpretation

- Model performs reasonably well, especially in predicting strokes (class 1) with high recall and F1-score.
- However, there's room for improvement in precision for class 1.
- Consider further tuning your model, exploring feature importance, and potentially addressing class imbalance.

Our model reveals age, hypertension, and heart disease as key predictors. As age increases, risk grows exponentially, and hypertension/heart disease significantly raise risk. These insights help clinicians identify high-risk patients, develop targeted prevention, and improve outcomes. By understanding stroke risk drivers, healthcare professionals can make informed decisions, enhance care, and save lives - paving the way for a healthier future.

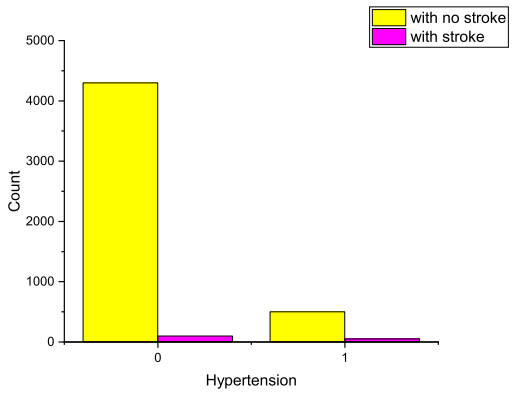

Fig. 11. Distribution of stroke among hypertention patient and non-hyper tension patients

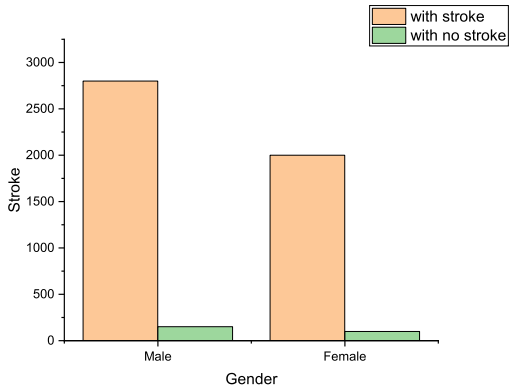

Fig. 12. Distribution of stroke among male and female

IV. Conclusion

This comprehensive analysis of the notebook suggests a thorough approach in predicting strokes using machine learning. The project comprises data preparation, conducting exploratory data analysis, implementing feature engineering techniques, applying models, and assessing their performance. The detailed visualizations provide insightful observations, and the careful consideration of different evaluation metrics ensures a robust assessment of the model's performance.

References

[1] Minhaz Uddin Emon, Maria Sultana Keya, Tamara Islam Meghla, Md Mahfujur Rahman, M Shamim Al Mamun, and M Shamim Kaiser. Performance analysis of machine learning approaches in stroke prediction. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1464–1469. IEEE, 2020.

[2] Priya Govindarajan, Ravichandran Kattur Soundarapandian, Amir H Gandomi, Rizwan Patan, Premaladha Jayaraman, and Ramachandran Manikandan. Classification of stroke disease using machine learning algorithms. *Neural Computing and Applications*, 32(3):817–828, 2020.

[3] RS Jeena, A Sukesh Kumar, and K Mahadevan. Stroke diagnosis from retinal fundus images using multi texture analysis. *Journal of Intelligent & Fuzzy Systems*, 36(3):2025–2032, 2019.

[4] Selma Yahiya Adam, Adil Yousif, and Mohammed Bakri Bashir. Classification of ischemic stroke using machine learning algorithms. *International Journal of Computer Applications*, 149(10):26–31, 2016.

[5] M Sheetal Singh and Prakash Choudhary. Stroke prediction using artificial intelligence. In *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pages 158–161. IEEE, 2017.

[6] T Kansadub. S. ammaboosadee, s. kiattisin, and c. jalayondeja,"stroke risk prediction model based on demographic data,". In *Proceedings of the 2015 8th Biomedical Engineering International Conference (BMEiCON)*, pages 1–3.

[7] Tahia Tazin, Md Nur Alam, Nahian Nakiba Dola, Mohammad Sajibul Bari, Sami Bourouis, and Mohammad Monirujjaman Khan. Stroke disease detection and prediction using robust learning approaches. *Journal of healthcare engineering*, 2021, 2021.

[8] Chetan Sharma, Shamneesh Sharma, Mukesh Kumar, and Ankur Sodhi. Early stroke prediction using machine learning. In *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, pages 890–894. IEEE, 2022.

[9] Chen-Ying Hung, Ching-Heng Lin, Tsuo-Hung Lan, Giia-Sheun Peng, and Chi-Chun Lee. Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database. *PloS one*, 14(3):e0213007, 2019.

[10] Gang Fang, Zhennan Huang, and Zhongrui Wang. Predicting ischemic stroke outcome using deep learning approaches. *Frontiers in genetics*, 12:827522, 2022.