Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study with a Web Application for Early Intervention

KRISHNA MRIDHA[1], MEMBER, IEEE, SANDESH GHIMIRE[1], JUNGPIL SHIN[2] (SENIOR MEMBER, IEEE), ANMOL ARAN[1], MD. MEZBAH UDDIN[1], M.F. MRIDHA[3] (SENIOR MEMBER, IEEE)

[1]Department of Computer Engineering, Marwadi University, India
[2]Department of Computer Science and Engineering, University of Aizu, Aizuwakamatsu, Japan
[3]Department of Computer Science, American International University-Bangladesh, Dhaka 1229, Bangladesh

Corresponding author: JUNGPIL SHIN (e-mail: jpshin@u-aizu.ac.jp)

ABSTRACT Stroke is a dangerous medical disorder that occurs when blood flow to the brain is disrupted, resulting in neurological impairment. It is a big worldwide threat with serious health and economic implications. To solve this, researchers are developing automated stroke prediction algorithms, which would allow for early intervention and perhaps save lives. The number of people at risk for stroke is growing as the population ages, making precise and effective prediction systems increasingly critical. The goal of this study was study aimed to achieve three objectives: (i) to create a trustworthy machine learning model to predict stroke disease, (ii) to address the severe class imbalance issue that results from the stroke patients' class being substantially smaller than the healthy class, and (iii) by using Mutual Information Score, Chi-Square Score, and ANOVA test, find the import feature(iv) to interpret the model output to gain a better comprehension of the decision-making process (v) balancing the dataset from the ratio of 19: 1 for No Stroke: Stroke to equal ratio using SMOTE Analysis (vi) Propose an End-to-End smart healthcare system through an android application. In a comparison examination with six well-known classifiers, the effectiveness of the proposed ML technique was explored in terms of metrics relating to both generalization capability and prediction accuracy. To give insight into the black-box machine learning models, we also studied two kinds of explainable techniques, namely SHAP and LIME, in this study. SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are well-established and reliable approaches for explaining model decision-making, particularly in the medical industry. The findings of the experiment revealed that more complicated models outperformed simpler ones, with the top model obtaining almost 91% accuracy and the other models achieving 83-91% accuracy. The proposed framework, which includes global and local explainable methodologies, can aid in standardizing complicated models and gaining insight into their decision-making, which can enhance stroke care and treatment.

INDEX TERMS Stroke Prediction, Data Leakage, Explainable Machine Learning, ANOVA test, SHAPE, LIME

I. INTRODUCTION

The incidence of stroke has been increasing globally, and it is now considered one of the leading causes of death and disability. Early intervention is crucial in preventing long-term disability and mortality associated with stroke. Traditional methods of predicting stroke risk, however, are often time-consuming and prone to errors.

Recently, machine learning algorithms have shown great promise in accurately predicting stroke risk based on various clinical risk factors. By leveraging these algorithms, clinicians can identify high-risk patients and intervene early, potentially reducing the number of stroke-related complications and improving patient outcomes.

Additionally, there is a growing need for transparency and explainability in machine learning models in healthcare. The use of an interpretable machine learning model can provide clinicians with valuable insights into the factors that contribute to a patient's stroke risk, thereby aiding in treatment decisions.

The World Stroke Organisation estimates that 13 million people worldwide experience a stroke each year, leading to 5.5 million fatalities [1]. Stroke affects all aspects of a patient's life, including their family, social environment, and work, and is one of the top causes of mortality and disability in the world [1, 2]. A common misconception is that certain groups of people, such as the elderly or those with underlying illnesses, are the only ones who are affected by stroke. In reality, anybody can be impacted, regardless of age, gender, or physical health [1, 2]. A stroke is a rapid, serious disruption in blood flow to the brain that deprives brain cells of oxygen. It comes in ischemic and hemorrhagic varieties. Moderate to severe strokes can cause permanent or temporary damage, depending on their severity. Hemorrhagic strokes are uncommon; however, they are brought on by the rupture of a blood vessel in the brain. The most common type of stroke happens when an artery is blocked or narrows, preventing blood flow to the brain [3, 4]. Age over 55, prior stroke or TIA, arrhythmia, high blood pressure, carotid stenosis from atherosclerosis, smoking, high blood cholesterol, diabetes, obesity, inactivity, estrogen therapy, blood clotting disorders, cocaine or amphetamine use, and heart issues like infarction and cardiac arrest are all risk factors for stroke [5-7]. Strokes can occur suddenly, and their symptoms might vary and be unanticipated. The main symptoms of a stroke include paralysis on one side of the body, numbness in the face, arms, or legs, difficulty speaking or walking, dizziness, blurred vision, headache, vomiting, drooping mouth, and, in severe cases, loss of consciousness and coma. These sensations may come on suddenly or gradually, and in certain rare cases, they may cause you to become aware [8–10].

Stroke can impact both men and women, lowering their quality of life and putting a load on public health resources. The scientific community prioritizes building models for predicting strokes to avoid them, and AI plays a critical role in this endeavor because it is extensively employed for disease prevention. Several research has been carried out to construct models for stroke diagnosis [11-13], predict treatment results and patient responses, and design individualized rehabilitation techniques [14-16]. Arslan et al. [17], for example, suggested a data mining system to predict ischemic strokes utilizing data from 80 ischemic stroke patients and 112 healthy persons, with the Support Vector Machine (SVM) classifier achieving the greatest accuracy of 97.89% and AUC of 97.83%. The study also looked at how different factors affected identifying the key risk factors for ischemic stroke.

The motivation for conducting this research is as follows:

- The incidence of stroke is increasing globally, and early intervention is crucial in preventing long-term disability and mortality associated with stroke.
- Traditional methods of predicting stroke risk are often time-consuming and prone to errors, which can result in delayed intervention and worsened patient outcomes.
- Machine learning algorithms have shown great promise in accurately predicting stroke risk based on various clinical risk factors, which can enable early identification of high-risk patients and timely intervention.
- The authors seek to explore the use of these algorithms for stroke prediction while also providing an explainable model and web application for clinicians to use in early intervention.
- There is a growing demand for transparent and interpretable machine learning models in healthcare, and the authors aim to provide a solution to this demand by providing a model that is both accurate and transparent.
- Overall, the motivation of this work is to enhance stroke prediction and early intervention, ultimately reducing the burden of stroke-related disability and mortality.

Some novelties brought in this article are as follows:

- The study aims to create a trustworthy machine learning model to predict stroke disease, which is a crucial step toward enabling early intervention and improving patient outcomes.
- The study addresses the severe class imbalance issue that results from the stroke patients' class being substantially smaller than the healthy class, which is a common challenge in developing effective prediction models.
- By using Mutual Information Score, Chi-Square Score, and ANOVA test, the study identifies important features that contribute to stroke risk, which can aid in understanding the decision-making process of the model.
- The study proposes an End-to-End smart healthcare system through an android application, which is a unique contribution to the field of stroke prediction.
- The study compares the proposed machine learning technique with six well-known classifiers and demonstrates its effectiveness in terms of both generalization capability and prediction accuracy.
- The study employs two kinds of explainable techniques, namely SHAP and LIME, to gain insight into the decision-making process of the model, particularly in the medical industry.
- The findings of the experiment reveal that more complicated models outperform simpler ones,

which is an important insight for developing accurate stroke prediction models.

- The proposed framework, which includes global and local explainable methodologies, can aid in standardizing complicated models and enhancing stroke care and treatment, which is a novel contribution to the field.

In this article, we provide a robust model with improved accuracy when XAI approaches are used in skin cancer diagnosis, and we make the following major contributions:

- Using XAI techniques like SHAP and LIME to explain the network's predictions can improve the transparency and precision of a deep-learning model for skin lesions. This can strengthen the model's transparency and general safety, which will boost confidence in the diagnostic system.
- Balance Dataset: The dataset is Unbalanced with a bias towards No Stroke in a ratio of 19: 1 for No Stroke: Stroke. We balance the dataset using SMOTE Analysis
- Feature Selection: By using Mutual Information Score, Chi-Square Score, and ANOVA test, find the Important Feature.
- Implement a Web-Based real-time application and Propose an End-to-End smart healthcare system through an android application.

A. Machine Learning in Stroke Prediction

Machine learning algorithms are trained on data on patients and their medical histories, as well as information about their risk factors and results, in the context of stroke prediction. The objective is to create models that can properly forecast a patient's chance of having a stroke, and then utilize that knowledge to identify individuals who are at high risk and take preventative measures. ML algorithms can examine vast volumes of data and uncover patterns and correlations that would be impossible to notice by hand. These models' results can be utilized to enhance diagnosis, therapy, and patient outcomes.

B. Explainable ML in Stroke Prediction

Explainable Machine Learning (XAI) is an artificial intelligence discipline that focuses on creating algorithms that can offer interpretable and transparent explanations for their predictions. The overall architecture to predict the Stroke using XAI and ML is shown in Figure 1. XAI algorithms attempt to give explanations for the reasons behind a forecast in the context of stroke prediction, allowing medical practitioners to understand the elements that impacted the prediction and make educated decisions. By giving clear and succinct explanations of the models' decision-making process, XAI algorithms can aid to build trust in machine learning models for stroke prediction. They can also assist in identifying and correcting any biases or flaws in the models. This is especially true in the medical industry, where decisions can have serious effects on patients.

Feature significance analysis, decision trees, and attribution approaches are some of the techniques often utilized in XAI for stroke prediction. These strategies can serve to offer a better understanding of the links between various risk factors and the possibility of having a stroke, as well as determine which risk variables are most important in predicting a stroke. This data may be utilized to create more effective preventative and treatment plans.

The accuracy and dependability of AI-assisted diagnostics may be improved by using XAI in stroke prediction, resulting in more confidence in the diagnostic system. To further enhance the performance of the model, XAI frameworks additionally provide an interface for domain specialists to submit comments and justifications. These insights can aid medical practitioners in making better treatment choices.

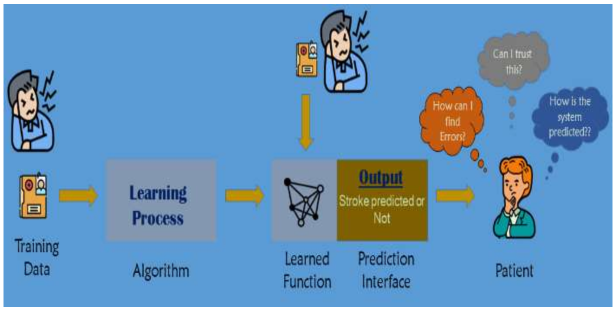


Figure 1: ML workflow with XAI: The model accuracy explains the prediction and gives the answer "why I should believe this output", "How it predicts" and "How can I find an error"

The format of this document is as follows: The history of the subject and a review of similar research are covered in Section II, with a focus on the advantages and disadvantages of contemporary methodologies. While Section IV offers the findings of the experimental investigation, Section III discusses the constructed model. Section V discusses the issues with the study's reliability, and Section VI brings the analysis to a close by detailing potential future research.

II. Related Work

Table 1 consists of the summary of some related papers' works in terms of accuracy, algorithm, dataset, and publication year.

Explainable AI (XAI) and Machine Learning are effective techniques for predicting outcomes based on prior data.

Mohammed Saidul Islam et al. [18] conducted research on the usage of XAI in healthcare technology to swiftly diagnose patients. Machine learning methods have been shown to predict stroke outcomes quickly and accurately. A large quantity of data from patients with and without strokes is required to train these algorithms. The goal of the study was to predict brain stroke using XAI and machine learning models with EEG signal data from stroke and non-stroke patients in a variety of situations. The use of electroencephalography (EEG) to predict acute stroke induced by ischemia episodes is a promising technology. The study focused on ischemic stroke patients and healthy persons in active situations, and it employed three-month-old datasets from 48 patients (45 with ischemic stroke and 75 healthy adults without neurological diseases). The model's adaptive gradient boosting, Xgboost, and LightGBM were utilized, and adaptive gradient boosting achieved 80% accuracy. To describe the model's behavior, Eli5 and LIME (Local Interpretable Model-Agnostic Explanations) were employed. This study and its models should make diagnostic judgments easier, quicker, and clearer.

Machine learning approaches were used by Elias Dritsas et al. [19] to investigate the early identification of stroke. A stroke happens when the blood supply to the brain is suddenly cut off. Early detection of such episodes, which can result in impairment or death, is critical. Several machine learning models and approaches were investigated in this study, and the stacking method showed to be the most successful. Various models, including RF, NBs, LR, KNN, and Stacking, were evaluated on precision, recall, F-measure, and accuracy using datasets containing 3254 individuals aged 18 and up. The stacking technique obtained 80% accuracy, 98.9% AUC, and 97.4% precision and recall.

Stroke is a worldwide concern caused by a disruption or decrease in blood circulation to the brain, resulting in a shortage of oxygen to brain tissue. It causes early death and expensive economic effects, such as a loss of productivity in Europe estimated at EUR 12 billion in 2017 and healthcare costs estimated at EUR 27 billion [20]. Explainable AI and machine learning are effective techniques for predicting strokes. To assess the efficiency of the proposed machine learning technique, a comparative study of six well-known classifiers was performed. The multi-layer perceptron classifier outperformed six other machine learning models in terms of G-Mean and false-negative rate, with an overall false-negative rate of 18.60%. The influence of risk variables on prediction output was investigated using Shapley Additive Explanations, a method for analyzing the contribution of input variables based on coalition game theory. The contribution of each characteristic in predicting the score was calculated by measuring the points gained or lost in the presence or absence of a feature. The MLP classifier was chosen as the best model for this binary problem because it achieved the best balance of G-Mean and false-negative rate, with a G-Mean of 75.83% and the lowest false-negative rate.

Redwanul Islam et al. [21] investigated the application of machine learning approaches to predict stroke in their study. A stroke occurs when blood flow to the brain is restricted or decreased, depriving brain tissue of oxygen and vital nutrients. The efficacy of the DT, SDG, KNN, SVM, and XGBoost classifier to predict stroke risk was tested in the study. Stroke was the top cause of death globally in 2016, accounting for 5.7 million fatalities, or 13% of all deaths. The stroke datasets utilized in the study were gathered from several hospitals in Bangladesh and contained 8600 patients, 2500 of whom had had a stroke.

The annual cost of treating stroke sufferers is projected to be over $26 billion. Darabi et [22] study.'s sought to identify high-risk patients who would benefit from targeted treatments to prevent 30-day readmission after an ischemic stroke. The study analyzed the performance of five machine learning algorithms to develop 15 models for predicting readmission using patient-level data from electronic health records. The dataset includes 3184 ischemic stroke patients, 1,960 of whom were from Geisinger Medical Centre and the remainder from various institutions. The study employed a data-driven feature selection technique as well as an adaptive sampling method.

Youngkeun Choi et al. [23] use machine learning to improve knowledge of factors in stroke modeling and to assess the prediction accuracy of decision trees. To build decision trees, the study used two algorithms: Cart (Classification and Regression Tree) and ID3. The overall accuracy rate is 0.981, implying a 0.019 error rate. 98.17% of patients projected not to have a stroke were properly predicted, whereas 16.67% of those anticipated to have a stroke were correctly recognized.

Tahia Tazin et al. [24] constructed four machine-learning models to predict stroke using physiological signs. Logistic Regression, Decision Tree Classification, Random Forest Classification, and Voting Classifier were among the models used. The SMOTE technique was used for data preparation to balance the skewed dataset. Random Forest exhibited the best accuracy, around 96%, which was greater than in earlier experiments.

Harshitha K V et al. [25] tested five different machine learning methods to predict the likelihood of stroke. Random Forest, Logistic Regression, K Nearest Neighbor, Decision Tree, and Support Vector Machine were the models used. Random Forest had the greatest accuracy of 95.5% and was chosen as the top model owing to its high accuracy and few false negatives.

The research was undertaken by Soumyabrata Dev et al. [26] to find essential indicators for stroke prediction using statistical methodologies. The performance of neural networks, decision trees, and random forests was evaluated using three scenarios: original features, PCA-transformed data from the first two main components, and PCA-transformed data from the first eight components. The most essential markers for stroke diagnosis were discovered, and among the approaches tested, a perceptron neural network with four critical characteristics had the highest accuracy rate and the lowest miss rate.

Hager Saleh et al. [27] used the Healthcare Dataset Stroke to assess the efficacy of distributed machine-learning algorithms in predicting stroke. The stroke prediction model was built using Apache Spark, a prominent big data platform, in conjunction with the MLlib package. Decision Tree, Support Vector Machine, Random Forest Classifier, and Logistic Regression were the four classification techniques employed. To improve the findings, cross-validation, and hyperparameter tweaking were applied. The Random Forest Classifier beat the other models, obtaining 90% accuracy across all performance measures such as Accuracy, Precision, Recall, and F1-measure.

In the paper [33], the authors discuss the challenges and potential biases of deep learning algorithms in medical image analysis. They propose strategies for improving the explainability and trustworthiness of these algorithms, including visualization techniques, feature attribution methods, and interpretable models. The article provides valuable insights into the importance of ensuring that deep learning algorithms in medical image analysis are transparent and can be understood by medical professionals and patients alike.

Table 1: Summary of some related papers' works in terms of accuracy, algorithm, dataset, and publication year

Ref	Year	DL Algorithm	Dataset	Performance
[18]	2022	Eli5, LIME, AGB	Kaggle	Accuracy: 80%
[19]	2022	NB, LR, K-NN, SGD, DT, MLP, RF, Stacking	Kaggle	Accuracy: 80%
[20]	2022	LR, RF, KNN, SVM, MLP	Cerebral Stroke Prediction-Imbalanced Dataset (Kaggle)	false-negative rate (18.60%) , Accuracy (73.52%)
[21]	2021	AGB, LR, RF, KNN, SVM, MLP	Hospitals in Bangladesh	Accuracy: 98%
[22]	2021	LR, RF, KNN, SVM, MLP	Geisinger Health Open-Source Data Set	Accuracy: 95%
[23]	2020	DT and ID3.	HealthCare Dataset	Accuracy: 98%
[24]	2021	LR, RF, SVM	The open-access Stroke Prediction dataset	Accuracy: Random forest: 96%
[25]	2021	Rf, LR, and DT	HealthCare Dataset	Accuracy: KNN: 95%
[26]	2022	NN, DT, and RF	EHRs by McKinsey & Company	Accuracy: NN: 77%
[27]	2019	RF, DT, and RF	Healthcare Dataset Stroke	RF:90%, DT:79%, SVM: 77%, LR:77%.

The research gap in the given texts is the absence of explainability and interpretability of the machine learning models used for predicting stroke. While the studies demonstrate the effectiveness of machine learning and XAI techniques in predicting stroke, they do not provide clear explanations of how the models arrive at their predictions. Additionally, the studies do not evaluate the trustworthiness of the models, which is essential in healthcare applications where incorrect predictions can have serious consequences. Future studies should focus on developing more interpretable machine learning models that can be trusted and provide clear explanations of their predictions. Another area for future research is the development of more comprehensive datasets that include a wide range of patient populations to improve the generalizability of stroke prediction models. Finally, further research can investigate the integration of clinical knowledge into machine learning models to improve their accuracy and interpretability.

III. Methodologies

Due to its efficiency in analyzing massive volumes of medical data, including photos of skin lesions, machine learning is being utilized more and more in medical diagnostics, including the categorization of skin cancer. The main objectives of employing machine learning models in the context of stroke prediction are to increase diagnostic precision and classification efficiency. Various machine learning models are often used to create an automated stroke prediction system, which is then assessed using metrics like accuracy, recall, and F1 score to find the best model for the job.

This study's method for categorizing stroke predictions automatically entails making a "Yes" or "No" prediction. A five-step approach is used to create the model, as illustrated in Figure 2: Getting a dataset of electronic health records is step one. Steps two and three involve pre-processing the dataset by rescaling and normalizing the data, step four involves extracting features, step five involves building a classifier algorithm using the extracted feature vectors, and step six involves using the SHAP and LIME methods to shed light on the model's decision-making process. This improved strategy attempts to increase the precision of stroke prediction and assist medical practitioners in making more knowledgeable treatment decisions.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3278273

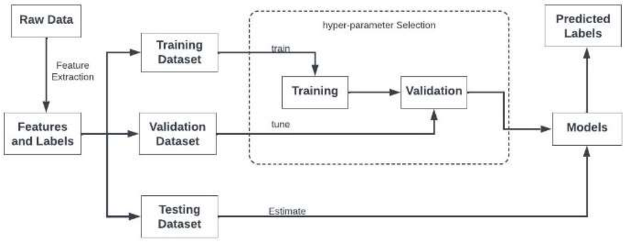


Figure 2: The framework of Proposed ML

A. Dataset Description

Obtaining a well-prepared dataset is critical for the efficient use of deep learning models in a variety of applications. Nonetheless, high-quality datasets are not always easy to come by. To create predictions, machine learning algorithms rely on the characteristics and patterns in the dataset. As a result, having a clean and well-prepared dataset is crucial for optimizing performance with deep learning models. The stroke prediction dataset utilized in the study has 5110 rows and 12 columns and was collected from Kaggle, a popular scientific community website. The dataset was unbalanced, with only 249 rows having a stroke value of one and 4861 rows having a stroke value of zero. To increase accuracy, the data was preprocessed and balanced using the SMOTE method.

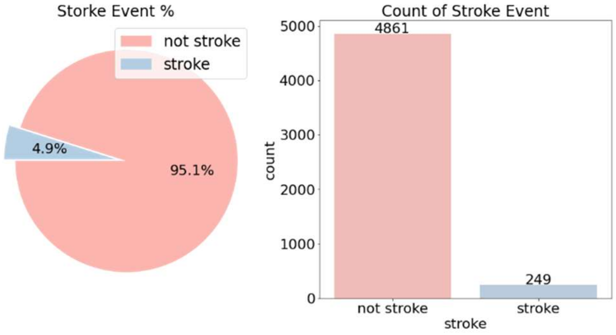


Figure 3: Target Samples Distribution from Original Dataset

Table 2: Dataset Description

Features	Description	Variable Type
Gender	Male", "Female" or "Other"	Categorical Features
Ever Married	"No" or "Yes"	
Work Type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"	
Residence Type	"Rural" or "Urban"	
Smoking Status	"Formerly smoked", "never smoked", "smokes" or "Unknown"*	
Hypertension	no hypertension = 0, hypertension = 1	Binary Numerical Features
Heart Disease	No heart diseases = 0, heart disease = 1	
Stroke	Healthy = 0, Stroke = 1	
Age	age of the patient	Continuous Numerical Features
Average Glucose Level	the average glucose level in the blood	
BMI	body mass index	

B. Data Preprocessing

Before developing a model, data pre-processing is required to remove noise and outliers in the dataset that might jeopardize the model's training. This step fixes any flaws that may prevent the model from working properly. Following the acquisition of the appropriate dataset, the data must be cleaned and structured in preparation for model building. The dataset utilized includes twelve characteristics, with the "id" column deleted because it has no bearing on model creation. The dataset is next examined for missing values and, if required, filled. In this situation, the mean of the column data was used to fill in the missing values in the "BMI" column. Label encoding turns the string literals in the dataset into integer values that the computer can understand. Strings must be translated to integers since the computer is typically educated on numbers. The obtained dataset has five columns of string data. During Label encoding, all strings are encoded, and the entire dataset is converted into a collection of integers. The dataset utilized for stroke prediction is highly skewed. The dataset contains 5110 rows, with 249 suggesting the likelihood of a stroke and 4861 proving the absence of a stroke. While utilizing such data to train a machine-learning model may result in accuracy, other metrics of accuracy, such as precision and recall, are insufficient. If such uneven data is not handled correctly, the results will be erroneous, and the prediction will be unsuccessful. As a result, to develop an efficient model, this uneven data must first be addressed.

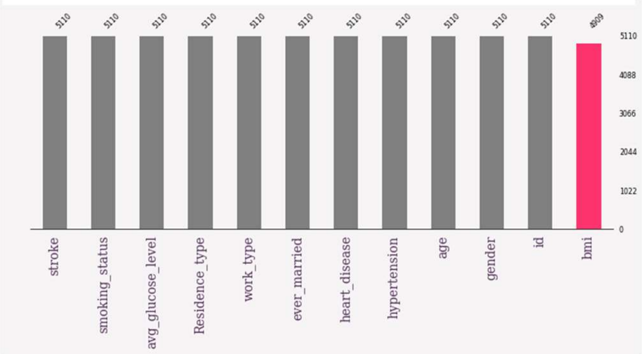


Figure 4: Null Values Visualization

C. Numerous Features Visualization

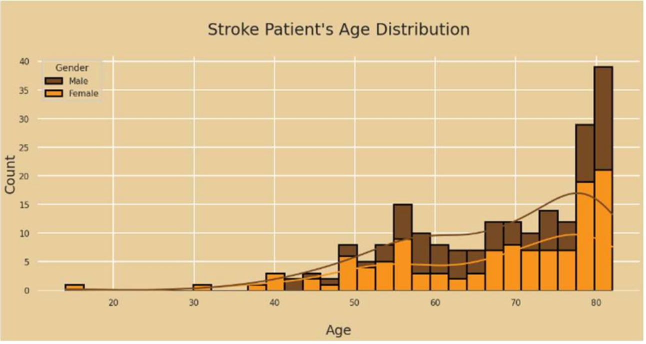


Figure 5: Age Distribution of Stroke patients

From Figure 5., We can see the stroke patient's age distribution is left-skewed. Most of the patients fall between 60 years to 82 years. Also, there are some young and children female stroke patients too.

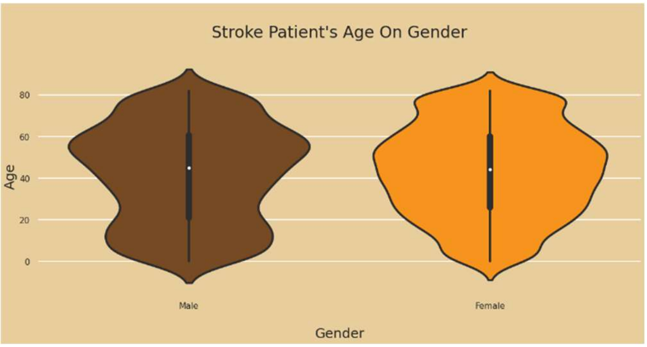


Figure 6: Stroke Patient's Age on Gender

And also Figure 6, provide evidence of the gender where most of the Male patients fall between 55 years to 82 years. Most of the Female patients fall between 48 years to 82 years.

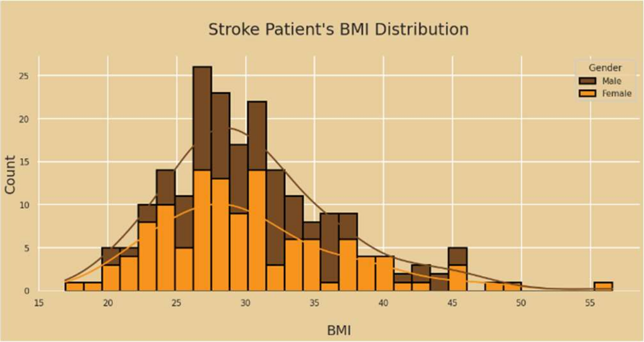


Figure 7: Stroke Patient's BMI Distribution

Figure 7 describes the stroke patient's BMI distribution as right-skewed. Most of the patient's BMI falls between 25 to 35. Also, there are some high BMI values too.

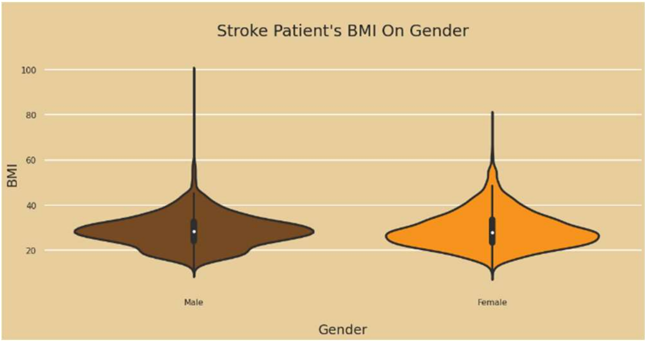


Figure 8: Stroke Patient's BMI on Gender

From Figure 8, we can see that most of the Male patient's BMI falls between 25 to 35 whereas most of the Female patient's BMI falls between 23 to 31.

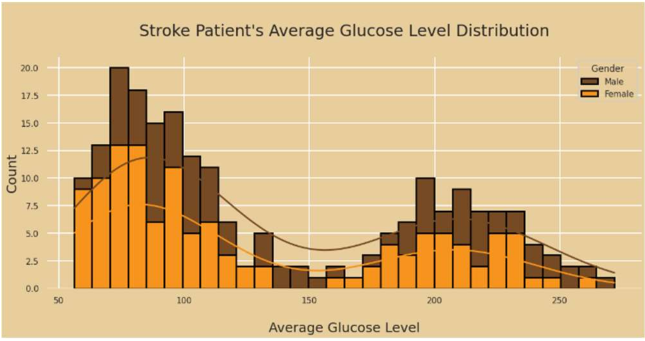


Figure 9: Stroke Patient's Average Glucose Level Distribution

Also, Figure 9 depicts the average glucose level where we can see most of the patient's average glucose levels fall between 60 to 120. In addition, there are some high average glucose levels too.

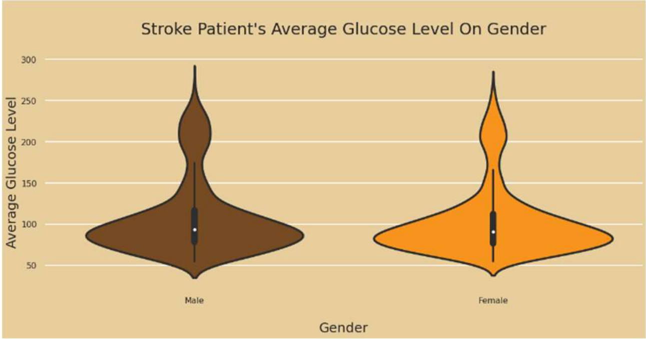


Figure 10: Stroke Patient's Average Glucose Level on Gender

From Figure 10, it is concluded that most of the Male patient's average glucose levels fall between 70 to 120 whereas most of the Female patient's average glucose levels fall between 55 to 115.

D. Binary Numerical Features Visualization

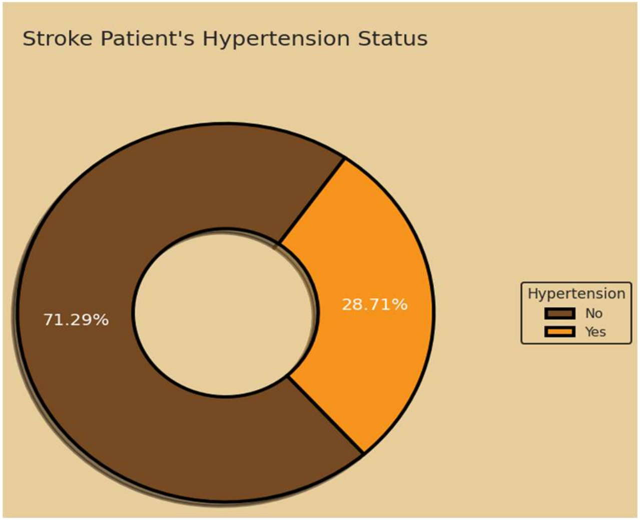


Figure 11: Stroke Patient's Hypertension Status

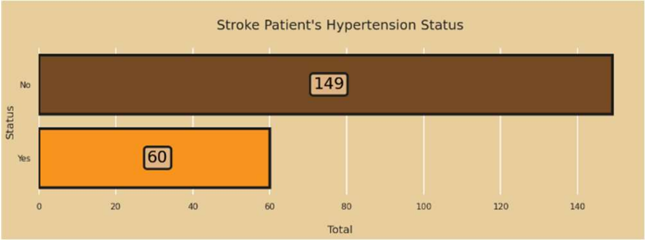


Figure 12: Stroke Patient's Hypertension Status

From Figures 11 and 12, we can visualize that most stroke patients do not have hypertension. Only 28.71% of patients have hypertension.

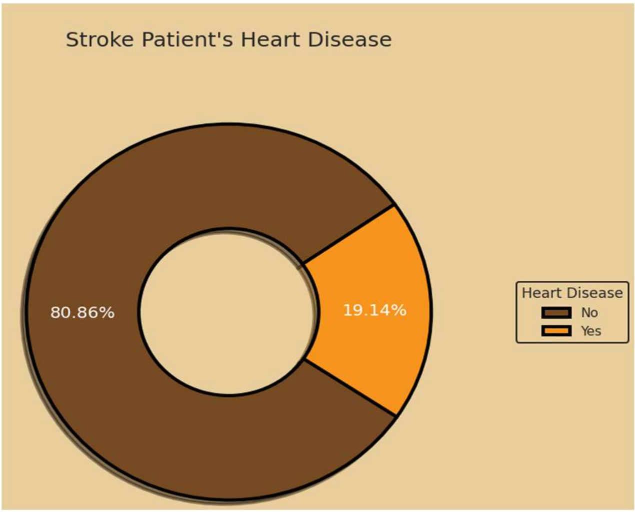


Figure 13: Stroke Patient's Heart Disease

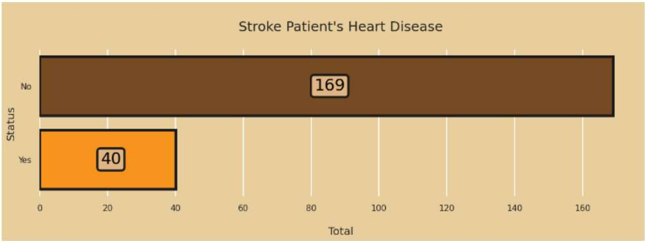


Figure 14: Stroke Patient's Heart Disease

Figures 13 and 14 visualizations of the Stroke Patient's Heart Disease number. We can see that most stroke patients do not have heart disease. Only 19.14% of patients have heart disease.

E. Categorical Features Visualization

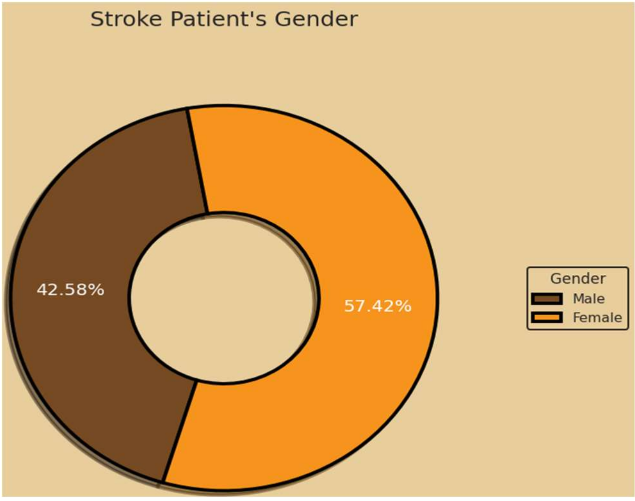


Figure 15: Stroke Patient's Gender

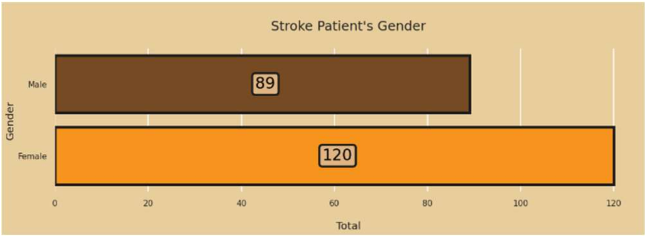


Figure 16: Stroke Patient's Gender

Figures 15 and 16 describe that most of the stroke patients are Female with a ratio of 57.42% followed by Males with a ratio of 42.58%.

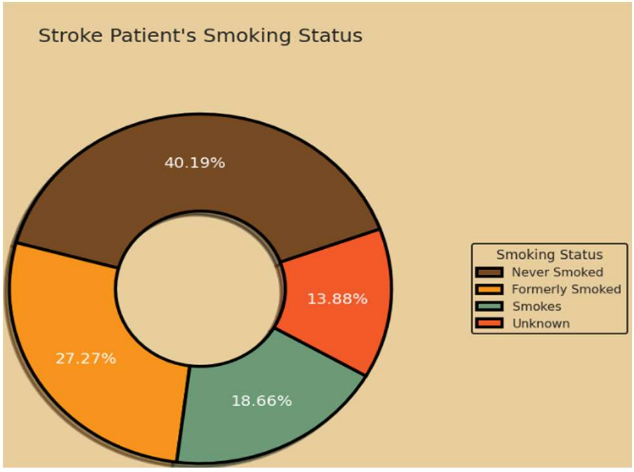


Figure 17: Stroke Patient's Smoking Status

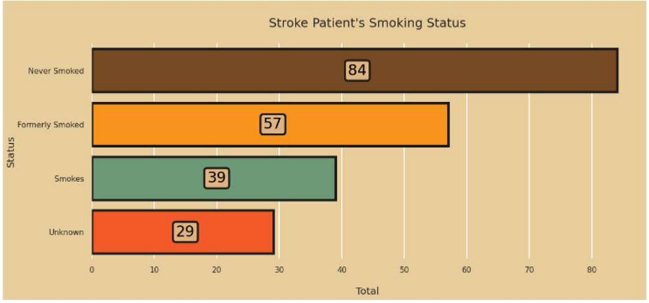


Figure 18: Stroke Patients Smoking Status

Figures 17 and 18 describe Smoking status where four types of status are available to predict stroke. In this dataset, most of the stroke patients have Never Smoked with a ratio of 40.19%. Some of the stroke patients have Smoked Previously with a ratio of 27.27%. For some patients, the smoking status is unknown.

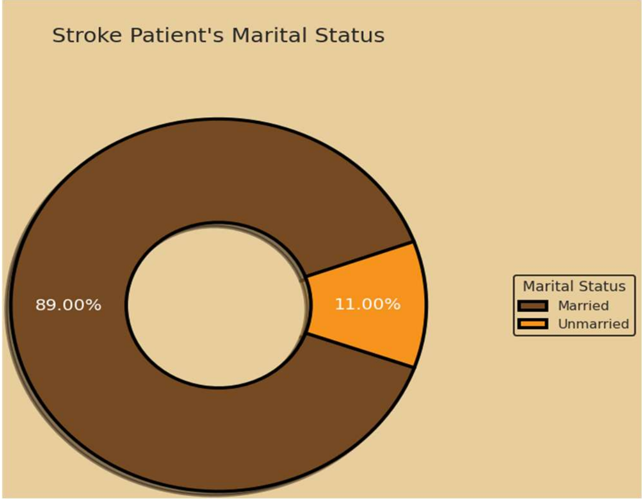


Figure 19: Stroke Patient's Marital Status

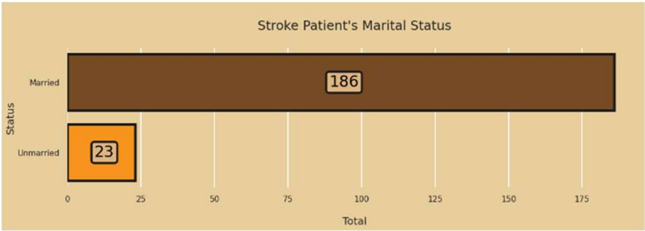


Figure 20: Stroke Patient's Status.

The Matarial status affects Stroke Patients. Most of the stroke patients are Married with a ratio of 89.00% followed by Unmarried with a ratio of 11.00%

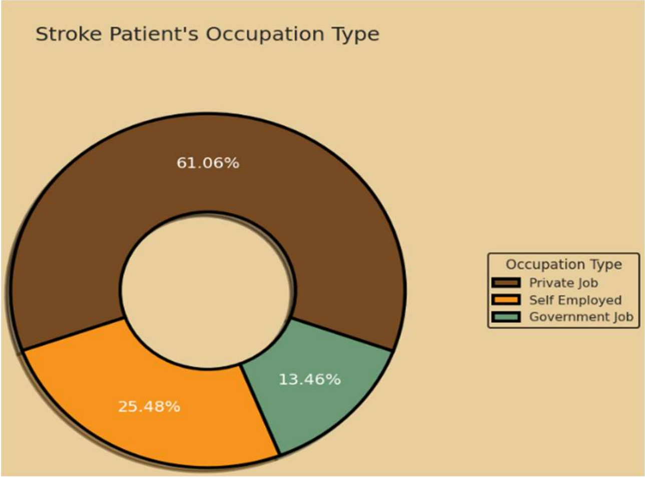


Figure 21: Occupation Type of Stroke Patient's

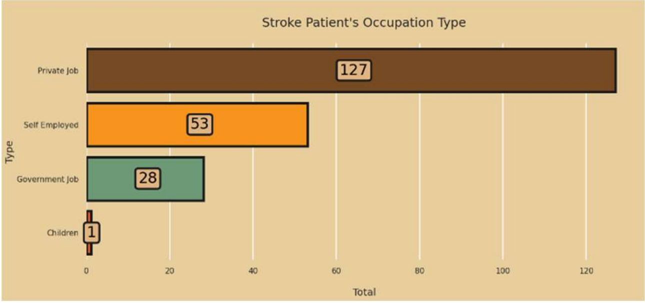


Figure 22: Stroke Patient's Occupation Type

Figures 21 and 22 describe the Occupation type where most of the stroke patients have experienced Private Jobs with a ratio of 61.06%. Some of the stroke patients have experienced Self Employment with a ratio of 25.48%. Some of the stroke patients have experience in Government Jobs with a ratio of 13.46%. Only 1 patient is children that's why it was not included in the donut chart.

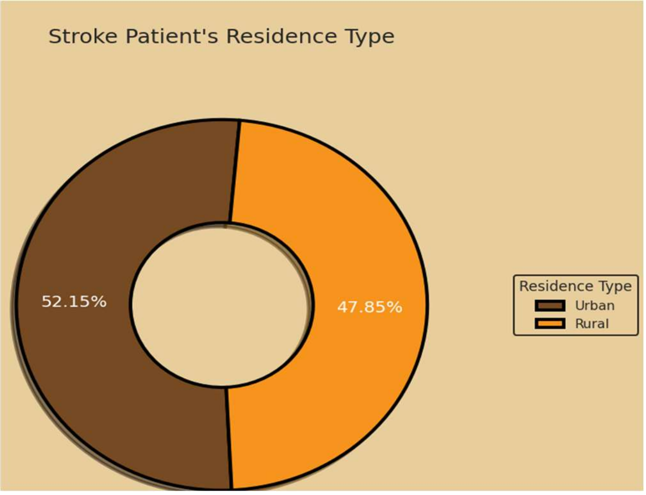


Figure 23: Stroke Patient's Residence Type

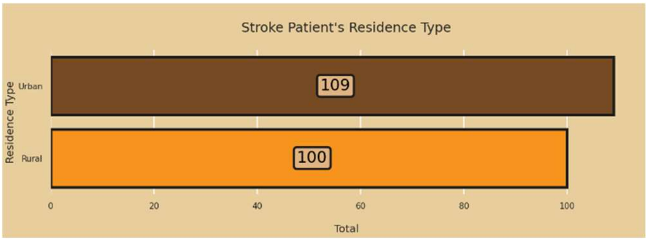


Figure 24: Stroke Patient's Residence Type

Figures 23 and 24 Stroke Patient Residence type where 52.15% of patients live in Urban areas and 47.85% of patients live in Rural areas.

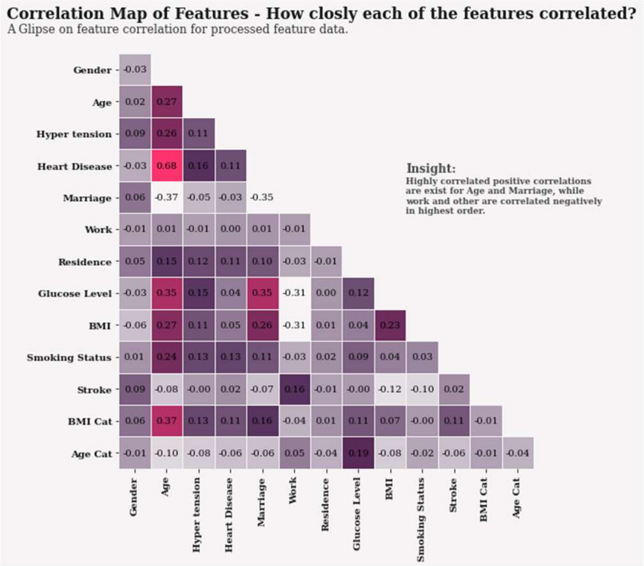


Figure 25: Correlation Map of Features: How closely are each of the features correlated?

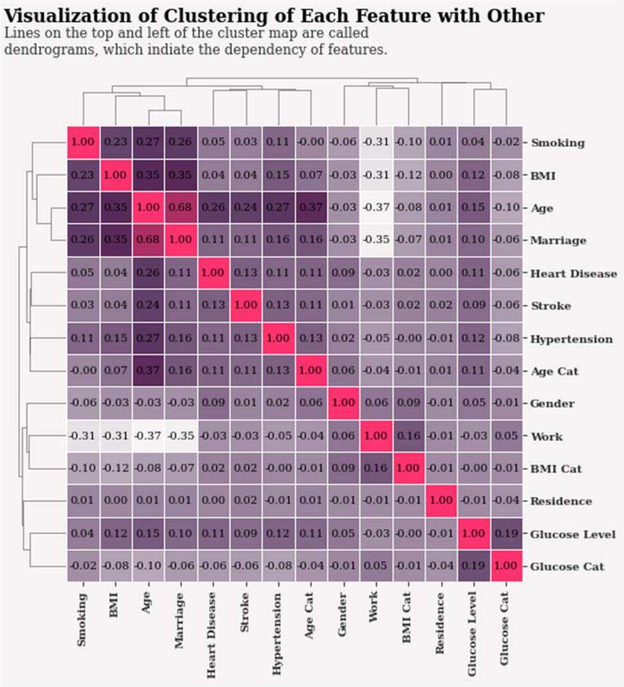


Figure 26: Visualization of Clustering of Each Feature with Other.

Figures 25 and 26 reveal that the target feature and other characteristics have a poor association. The target characteristic has a little positive connection with age, hypertension, heart disease, and average glucose level. Stroke, Hypertension, Heart Disease, Average Glucose Level, and BMI all have a minor positive connection with age. Marital Status, Occupation Type, and BMI all have a slight positive connection with smoking status. Age and Occupation Type have a medium positive association, whereas Age and Marital Status have a medium negative correlation.

F. Feature Engineering

Categorical Features (Order):

- Gender: male > female
- Hypertension: hypertension > no hypertension
- Heart disease: heart disease > no heart disease
- Ever married: married > no married
- Working type: Private > Self-employed > Govt job > children. the stress from work can lead to stroke.
- Residence type: Urban > Rural. mortality due to stroke is higher in rural areas than in urban areas due to poor medical treatment.
- Smoking status: never smoked > formerly smoked > smokes. smoking increases the risk of stroke.

Discrete Features (Range):

- Age (55 – 80): The chance of having a stroke doubles every 10 years after age 55.
- Avg glucose level (80 – 200): High blood glucose is found in stroke cases. A value of 126+ has been observed a lot.
- BMI (20 – 40): High BMI values increase the chances of ischemic stroke.

The dataset is Unbalanced with a bias towards No Stroke in a ratio of 19: 1 for No Stroke: Stroke. We will first balance the dataset using SMOTE Analysis!

By using SMOTE to generate synthetic data points, we can ensure that the new data is representative of the original data, while also addressing any imbalances or biases that may be present. This can lead to more accurate and reliable results, which can ultimately help to improve our understanding of the phenomenon being studied.

To cope with unbalanced data, there are 2 options:

- Under-sampling: Trim down the majority of samples of the target variable.
- Oversampling: Increase the minority samples of the target variable to the majority samples.

For best performances, the combination of under-sampling and oversampling is recommended.

- First, we will undersample the majority samples and it is followed by oversampling minority samples.
- For data balancing, we will use learn.
- PIP statement: pip install imbalanced-learn

The calculation for Data Balancing:

- Sampling Strategy: It is a ratio that is the common parameter for oversampling and under-sampling.
- Sampling Strategy: (Samples of Minority Class) / (Samples of Majority Class)
- In this case,
  - Majority Class: No Stroke: 4861 samples
  - Minority Class: Stroke: 249 samples

Under-sampling: Decrease the majority class

- Sampling Strategy = 0.1
- 0.1 = (49) / Majority Class Samples
- After under-sampling,
- Majority Class: No Stroke: 2490 samples
- Minority Class: Stroke: 249 samples

Oversampling: Increase the minority class samples

- Sampling Strategy = 1
- 1 = (Minority Class Samples) / 2490
- After oversampling,

- Majority Class: No Stroke: 2490 samples
- Minority Class: Stroke: 2490 samples

Final Class Samples:

- Majority Class: No Stroke: 2490 samples
- Minority Class: Stroke: 2490 samples
- Here, we balance the data by reducing the majority group samples & then increasing the minority group to the majority group.
- In the case of imbalanced datasets, we duplicate the data to account for potential bias in the predictions.
- Because of the duplication process, we are modeling with synthetic data to verify that the forecasts are not skewed towards the majority target class value.
- As a result, using accuracy to evaluate models will be deceptive. Instead, for model evaluation, we will use the confusion matrix, ROC-AUC graph, and ROC-AUC score.

Data Leakage

Data Leakage is the problem when information outside the training data is used for model creation. It is one of the most ignored problems. To create robust models, solving data leakage is a must! The creation of overly optimistic models which are practically useless & cannot be used in production has become common. Model performance degrades when Data Leakage is not dealt with & the model is sent online. It is a difficult concept to understand because it seems quite trivial. The typical approach used is transforming/modifying the entire dataset by filling NAN values with mean, median & mode, standardization, normalization, etc. When we execute the above process to make the dataset ready for modelling, we use the values from the entire dataset & thus indirectly provide information from the to-be test data i.e. outside of the training data. Thus, to avoid Data Leakage, it is advised to use train-test-split before any transformations. Execute the transformations according to the training data for the training as well as test data.

Firstly, identified data leakage in our machine learning project by carefully reviewing our data pre-processing and feature engineering steps. Specifically, we noticed that some of the information from the target variable was inadvertently included in our input features. To address this issue, we removed these features and re-ran our experiments to ensure that our models were not relying on this leakage.
To convert the data leakage into a percentage, we first calculated the proportion of our input features that were

affected by the leakage. We then calculated the percentage by multiplying this proportion by 100. This allowed us to quantify the extent of the data leakage and communicate it clearly in our analysis. We also included a discussion of the potential impact of the leakage on our results and any steps we took to mitigate its effects.

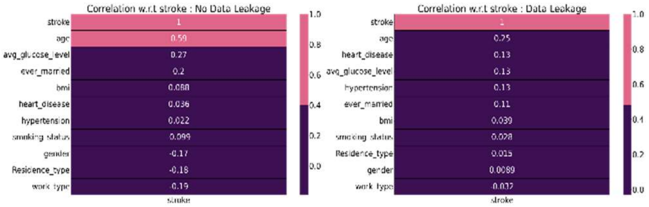


Figure 27: Correlation w.r.t stroke: No Data Leakage vs Data Leakage.

We can see from Figure 27 the difference in values between Data Leakage & No Data Leakage. In the case of No Data Leakage, age displays a strong positive correlation with stroke. avg_glucose_level & ever_married display some kind of positive correlation. Opposite to positive correlation, gender, Residence type & work type has a negative correlation with stroke. In the case of Data Leakage, none of the features display an extremely positive or negative correlation with stroke. age, heart_disease, avg_glucose_level, hypertension & ever_married display some kind of positive correlation. Overall, all the features have a value very close to 0, displaying a neutral correlation with stroke.

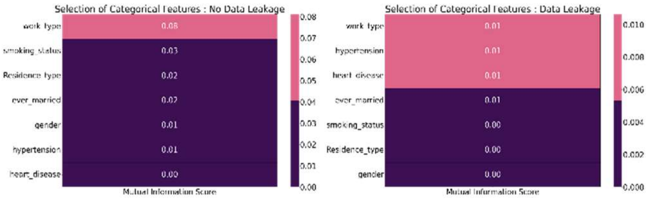


Figure 28: Mutual Information Score to describe the selection of Categorical features: No Data Leakage vs Data Leakage.

The mutual Information Score shown in Figure 28 said that the strokes with categorical features display very low scores irrespective of Data Leakage or No Data Leakage. According to the above scores, none of the features should be selected for modeling.

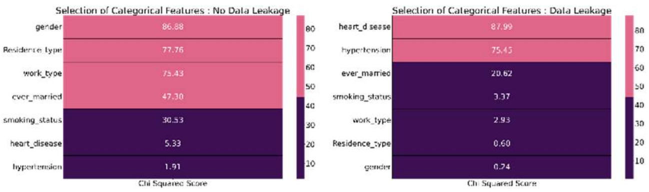


Figure 29: Chi-Square Score to describe the Selection of Categorical Features: No Data Leakage vs Data Leakage.

For No Data Leakage, we should reject the features that have low values. We will reject features with scores less than 20. Hence, we will not use smoking status, heart disease & hypertension. This does contradict the Domain Information. For Data Leakage, heart disease & hypertension need to be selected for modeling, and reject the other features due to the low Chi Squared Score.

There is a lot of importance in preventing data leakage and performing k-fold cross-validation on the dataset. To prevent data leakage, we took care to ensure that no information from the test set was used during the training process. This was accomplished by using only the training set to build and tune our models, and by not looking at the test set until the final evaluation stage.

In terms of k-fold cross-validation, we used this technique to evaluate the performance of our models and to check for any potential data leakage issues. The overall dataset was divided into k parts, and for each fold, we used one part for cross-validation and the remaining parts for training. This allowed us to train and evaluate our models on different subsets of the data, which can help to reduce the risk of overfitting and improve the generalization performance of the models.

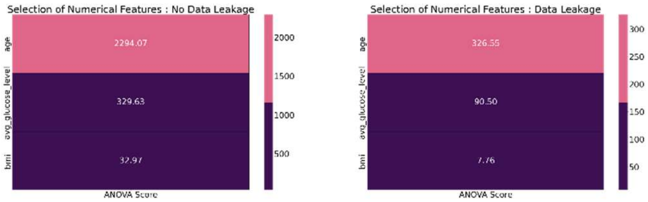


Figure 30: ANOVA Score to describe the selection of Numerical Features: No Data Leakage vs Data Leakage.

From the above ANOVA Scores shown in Figure 30, we ignore the features with values less than 20. Hence, we reject BMI for modeling irrespective of Data Leakage or No Data Leakage. We ready the datasets for data scaling by dropping the features based on the above statistical tests. We will ignore the Domain Information!

Data Scaling:

The machine learning model does not understand the units of the values of the features. It treats the input just as a simple number but does not understand the true meaning of that value. Thus, it becomes necessary to scale the data.

- We have 2 options for data scaling:

  1) **Normalization**
  2) **Standardization**.

As most of the algorithms assume the data to be normally (Gaussian) distributed, Normalization is done for features whose data does not display normal distribution, and standardization is carried out for features that are normally distributed but the range of values is huge or small as compared to other features.

In this article, we use the Standardization technique to scale the input. Standardization is a common preprocessing technique in machine learning that transforms input data to have a mean of zero and a standard deviation of 1. This technique is useful when input features have different scales and ranges, ensuring all features are treated equally during learning. Standardization is a good choice for scaling the Kaggle dataset as it prevents certain features from dominating the learning process due to their larger values and improves the performance of certain algorithms sensitive to feature scale. It also makes the data more interpretable and easier to compare across features, transforming features to a common scale that is easier to interpret and compare. Overall, standardization can improve the accuracy and reliability of machine learning models on the Kaggle dataset.

In addition, scaling data in machine learning projects can have a significant impact on analysis and interpretation. Scaling techniques like standardization can improve the accuracy and reliability of machine learning models, and make the data more interpretable and easier to compare across features. Scaling can influence the performance of machine learning algorithms as many are sensitive to feature scale, and standardization can ensure all features are treated equally during learning. Additionally, scaling makes it easier to identify patterns and relationships between features, which can guide feature selection and engineering efforts, leading to a better understanding of the underlying structure of the data.

G. X-AI: Explainable Artificial Intelligence in Model Explanation

The goal of XAI (Explainable Artificial Intelligence), which is discussed in the article, is to create AI systems that can give concise, intelligible justifications for their predictions and choices. The major objective of XAI is to develop trustworthy and transparent AI systems that enable users to understand their judgments. The article describes many methods used to produce model explanations in XAI, including feature significance determination, influence analysis, and visual explanations. It also emphasizes the value of model explanation in XAI. Additionally, it mentions various conventional approaches to illuminating machine learning models, such as SHAP and LIME.

By utilizing model-agnostic interpretation approaches, the most current developments in machine learning may be used to generate explanations of complex models while preserving high prediction accuracy. Because it isolates the model from explanations, the model-agnostic understanding is far more versatile than the model-specific interpretation method. Local explanation and global explanation are the two categories of model-independent classification strategies. [28]. LIME is the most often used method of local explanation. PDP and SHAP are the two most popular methods that may be globally interpreted.

LIME trains local surrogate models to provide the generalization ability for complex models. By causing the current data to change, LIME first creates a new dataset. After then, LIME trains a clear model, like a decision tree, using the new dataset. The equivalent prediction performance of the interpretable model and the black box model are compared in the last section. LIME is characterized as follows.:

$$\gamma(x) = \underset{\mathrm{g} \in \mathrm{G}}{\arg\min}\, L(f, g, \pi_x) + \Omega(g) \tag{1}$$

Where, L is the loss function used to quantify how close the comprehensible model g is to the projection of the original sophisticated model f . G demonstrates the family of comprehensible models. π_x denotes proximity of the evaluated instances to the instance x. $\Omega(g)$ is the criticality of the model g .

PDP illustrates how a single feature has minimal impact on the outcome anticipated by a sophisticated machine learning system. Whether it is linear, monotonous, or more complex, PDP shows the relationship between the input and the output [29]. The partial dependence function $\hat{f}_{x_x}$ defined as:

$$\hat{f}_{x_s}(x_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{x_x}(x_s, x_c^i) \tag{2}$$

Where partial function is $\hat{f}_{x_x}(x_s)$ which demonstrates the global correlation of an input feature with the projected outcome. S consists of a feature set with just one or two features., x_s represents the set of features to be projected by $\hat{f}_{x_x}(x_s)$, xc denotes the other characteristics used in the machine learning f. x_c^i express the absolute characteristics values from the dataset for the features in which we are not interested, and the number of occurrences of the dataset. is n .

The Shapley values are used by SHAP [29] for the complex model to determine the influence of the features. Shapley values are defined as the weighted average of marginal contributions. It can be identified by how feature value affects projection across all potential relationships. Shapley value is defined as:

$$\phi_j(x) = \sum_{s \subseteq \{x_1, x_2, \ldots . x_m\} \setminus \{x_j\}}^{n} \frac{|s|!(m-|s|-1)!}{m!}(val(s \cup \{x_j\} - val(s))) \tag{3}$$

There $\phi_j(x)$ is the Shapley value x_j , x_j which denotes a feature value, Feature subset of the model is s , m denotes the number of features, val and is the projection for feature values in the set S .

IV. Results

A. Metrics

In our study, we used a range of evaluation metrics [30] to assess the performance of our machine-learning models. Specifically, we utilized metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curve to evaluate the classification performance of our models.

Accuracy was used to measure the overall performance of the model in correctly predicting the class labels. Precision and recall were used to evaluate the model's ability to correctly classify positive and negative samples, respectively. F1-score was used as a harmonic mean of precision and recall to provide a balance between the two metrics. Finally, we used the AUC-ROC curve to evaluate the performance of our models in differentiating between positive and negative samples.

We chose these metrics based on their relevance to our research question and their suitability for evaluating classification performance. Additionally, we compared the performance of different machine learning models using these metrics to select the best-performing model for our analysis.

Overall, the use of these machine learning evaluation metrics allowed us to assess the performance of our models and make informed decisions about their use in our research. [31].

Table 3: Most Common Machine Learning Evaluation Metrics [32]

		Predicted values		
		True	False	
Actual	True	True Positive (TP)	False Negative (FN) Type 1 Error	$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$
	False	False Positive (FP) Type 1 Error	True Negative (TN)	$Specificity=\frac{TN}{TN+FP}$
		$\Pr ecision$ $\frac{TP}{TP+TN}$		$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$ $F1=\frac{2\ x\ \Pr ecision\ \text{x Recall}}{\Pr ecision+\text{Recall}}$

B. Classification Model Results

Table 4: Classification Accuracy for Data Leakage and No Data Leakage

Algorithms	Data Leakage	No Data Leakage
Random Forest	90.36	82.23
Logistic Regression	80.18	74.35
Support Vector Machine	80.18	74.65
K Nearest neighbors	86.74	81.61
Naive Bayes	76.03	71.26
XGB Classifier	89.02	83.43

Looking at the results in Table 4, we can see that Random Forest is the best-performing algorithm in terms of accuracy, with a score of 90.36% without data leakage and 82.23% with data leakage. XGB Classifier is the second-best performing algorithm with accuracy scores of 89.02% and 83.43% without and with data leakage, respectively.

Other algorithms like K Nearest Neighbours, Logistic Regression, and Support Vector Machine also perform well, but their accuracy scores are slightly lower compared to Random Forest and XGB Classifier. Naive Bayes has the lowest accuracy scores among all the algorithms, with 76.03% accuracy without data leakage and 71.26% with data leakage.

Overall, these results suggest that Random Forest and XGB Classifier are strong performers in this dataset, while Naive Bayes is the weakest. However, it is important to consider other factors beyond accuracy, such as interpretability and computational efficiency, when choosing a machine learning algorithm for a particular task.

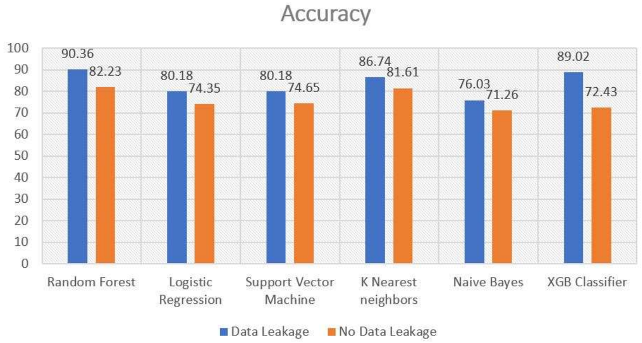


Figure 31: Accuracy Graph for all ML models

Looking at Figure 31, we can see that Random Forest and XGB Classifier consistently perform better than the other algorithms, both with and without data leakage. Naive Bayes has the lowest accuracy scores in both cases. We can also see that the difference in accuracy scores with and without data leakage is generally small, suggesting that data leakage may not be a major issue for this particular dataset and set of algorithms.

Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. Each decision tree in the Random Forest is constructed based on a subset of the data and a random subset of the features. This randomness helps to reduce the risk of overfitting, which is a common problem in machine learning where the model is too complex and performs well on the training data but poorly on new data.

Random Forest also has several other advantages that make it a popular choice for machine learning tasks. For example, it can handle both numerical and categorical data, and it is relatively easy to tune the model to improve performance. Additionally, Random Forest can provide information on feature importance, which can help understand the factors that are driving the predictions.

Overall, Random Forest is a powerful machine-learning algorithm that is often able to achieve high accuracy on a variety of tasks. However, as with any machine learning algorithm, it is important to carefully consider the specific characteristics of the problem at hand and the available data before selecting the best algorithm to use.

Table 5: Confusion Matrix, Precision Matrix, and Recall Matrix for all Models.

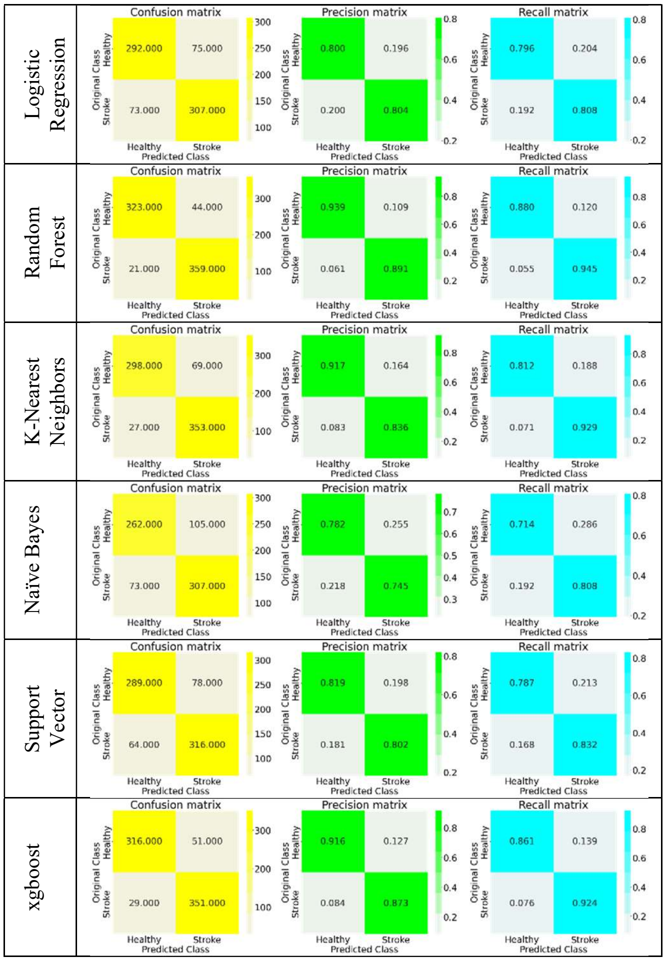


Table 5 depicts the Confusion Matrix, Precision, and Recall with three different colors for all the ML models.

Table 6: Precision, Recall, and F1-score respectively a, b, and c

Model		LR	RF	KNN	SVC	NB	Xgb
Precision	0	0.81	0.93	0.92	0.82	0.79	0.90
	1	0.79	0.88	0.83	0.79	0.74	0.88

a. Precision

Model		LR	RF	KNN	SVC	NB	Xgb
Recall	0	0.77	0.87	0.69	0.77	0.87	0.87
	1	0.83	0.94	0.82	0.84	0.91	0.91

b. Recall

Model		LR	RF	KNN	SVC	NB	Xgb
F1-score	0	0.79	0.90	0.86	0.79	0.74	0.89
	1	0.81	0.91	0.88	0.81	0.78	0.89

c. F1-Score

Table 6 represents the precision, recall, and F1-score for all machine learning models used in this article. a for Precision, b for recall, and c for F1-score. We have two classes where the value "0" for "Healthy" and "1" for "Stroke".

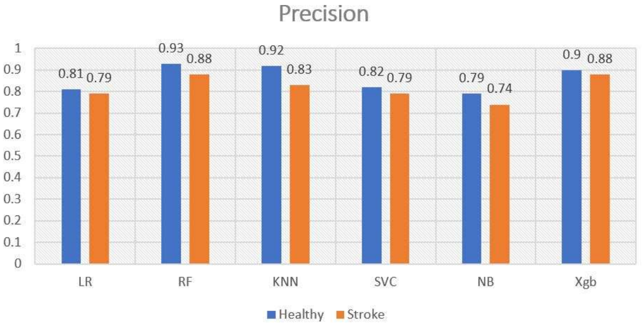


Figure 32: Precision Value for all ML models

From Figure 32, we can see that RF has the highest precision values for both classes, with a precision of 0.93 for class 0 (No) and 0.88 for class 1 (Yes). LR and KNN also have high precision values, with a precision of 0.81 and 0.92 respectively for class 0 (No). NB has the lowest precision values for both classes, with a precision of 0.74 for class 1 (Yes).

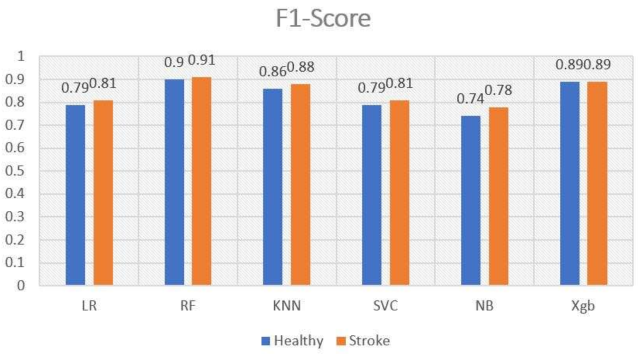


Figure 33: Recall Values for all ML models

Interpreting Figure 33, we can see that RF and Xgb have the highest F1-scores for both classes, with RF having the highest score for Class 0 and Xgb having the highest score for Class 1. NB has the lowest F1-score for both classes. Overall, the plot provides a clear visual comparison of the performance of each model for each class.

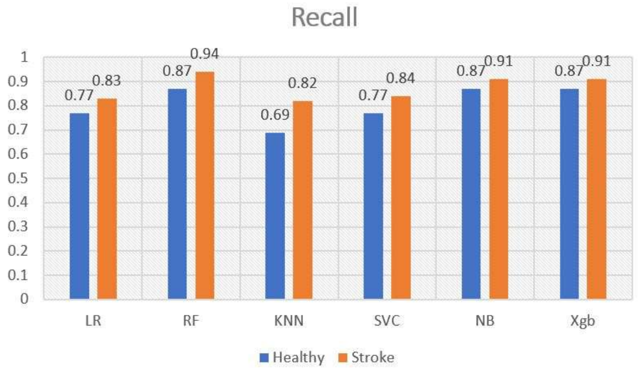
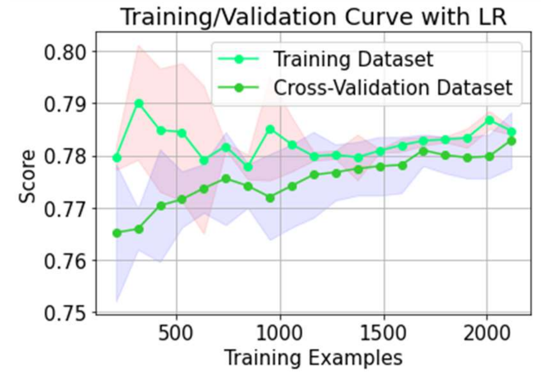
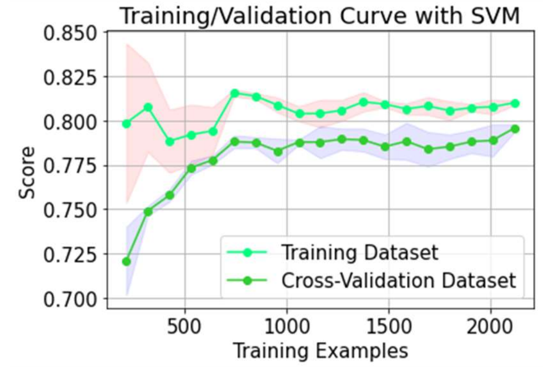
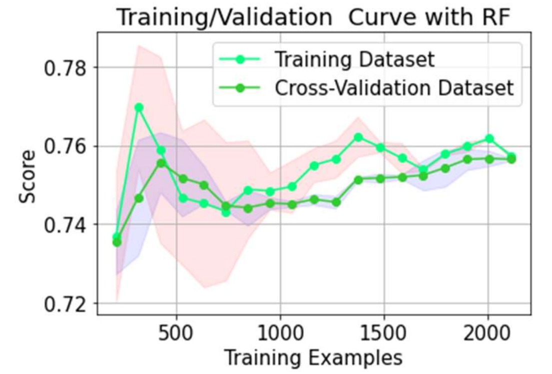
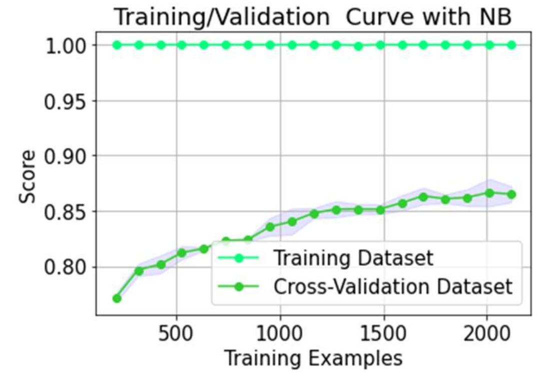
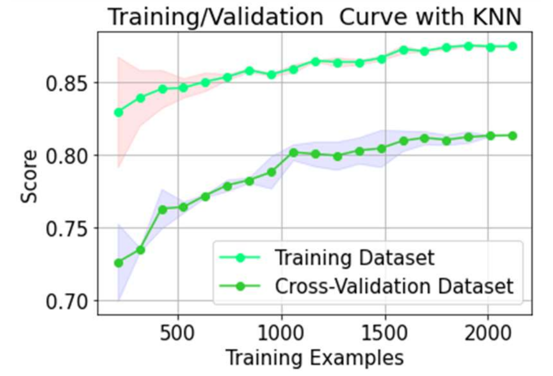
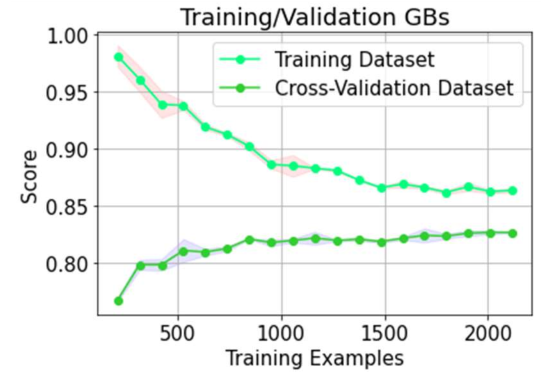


Figure 34: Recall Values for all ML models

From Figure 34, we can see that the RF and Xgb models have the highest recall scores for both classes, with RF having the highest score for Class 0 and Xgb having the highest score for Class 1. The KNN and SVC models have relatively lower recall scores for both classes, while the LR and NB models have more varied recall scores across the two classes.

Table 7: Training vs Validation Accuracy curve for all the models. These curves describe the model learning for all the samples.

Model	Training vs Validation Accuracy
Logistic Regression	
Support Vector Machine	
Random Forest	
Naïve Bayes	

K Nearest Neighbors	
xgboost	

The training and validation accuracy curves shown in Table 7 are visual representations of how well a model is learning and generalizing to new data [31]. The training accuracy curve represents the accuracy of the model on the training dataset, while the validation accuracy curve represents the accuracy of the model on a validation dataset that it has not seen during the training process.

This section emphasizes the significance of validation accuracy in evaluating a machine learning model's performance since it shows how effectively the model generalizes to new, untried data. The overfitting and underfitting problems, which can result in a model with poor performance, are also highlighted in the paragraph. The paragraph advises visualizing the training and validation accuracy curves to understand the model's behavior and enhance its performance.

To identify underfitting, we look for signs that our model is not capturing the complexity of the data. This can manifest as poor performance on both the training and test datasets, with the model unable to accurately predict the target variable. In this case, we may need to revisit our feature selection or engineering techniques to ensure that we are capturing the relevant information in the data.

On the other hand, overfitting occurs when our model is too complex and begins to memorize the training data rather than generalize to new data. This can result in excellent performance on the training dataset but poor performance on the test dataset. To mitigate overfitting, we can use techniques such as regularization or early stopping during the training process.

It is important to note that the risk of underfitting or overfitting can depend on the specific circumstances of the dataset and the modelling approach. For example, a small dataset may be more prone to overfitting, while a very large dataset may be more prone to underfitting if the model is not complex enough. Additionally, certain modelling techniques may be more or less prone to overfitting depending on their inherent flexibility.

Overall, we carefully monitor the performance of our models during training and testing to identify any signs of underfitting or overfitting and adjust our approach as needed to optimize performance on new data.

C. X-AI performance on Machine Learning

The significance of Explainable Artificial Intelligence (XAI), particularly in medical contexts, is discussed in the article. XAI offers concise and accessible explanations for predictions provided by machine learning models. It emphasizes that XAI approaches like SHAP and LIME can aid in delivering straightforward explanations because patients and doctors without technical backgrounds may struggle to grasp these forecasts.

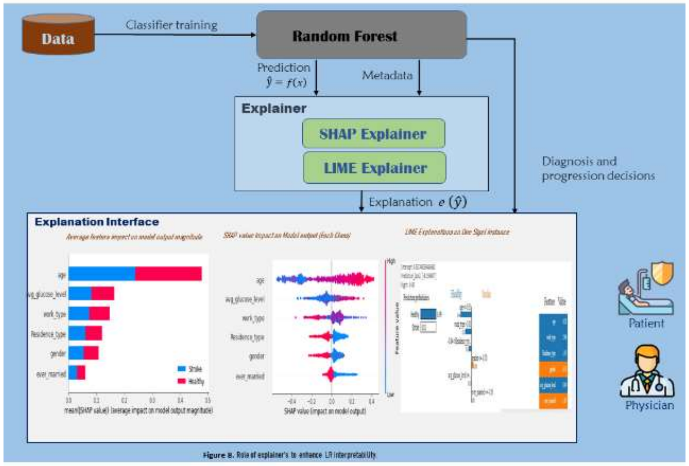


Figure 35: Role of Explainer to enhance the interpretability

Figure 35 describes the pipeline of XAI output from Data loading to the patient’s question. Here, we introduce LIME and SHAP output got from the model explanation to answer the patient's questions “how” and “why”. No only patients, it is mostly helpful for the physician to interpret our patient report to reach the final prediction.

C.1 Global Explanations: Each predictor's effect on the result of the complex model must be ascertained, we calculate the mean values of the random forest's Shapley Additive Explanations (SHAP) method. Figure 9. illustrates the common feature impact of the created RF classifier.

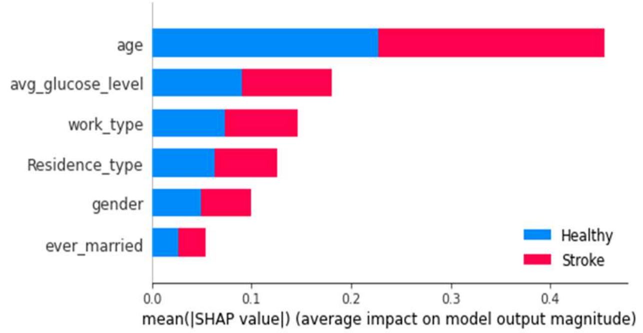


Figure 36: Average feature impact on model output magnitude

It is found that the six factors with the greatest effects are Age, Average Glucose Level, Work type, Residence type, Gender, and Ever Married. Theoretical explanations for the feature focus are generally in line with existing knowledge from hepato-biliary experts as well as the literature.

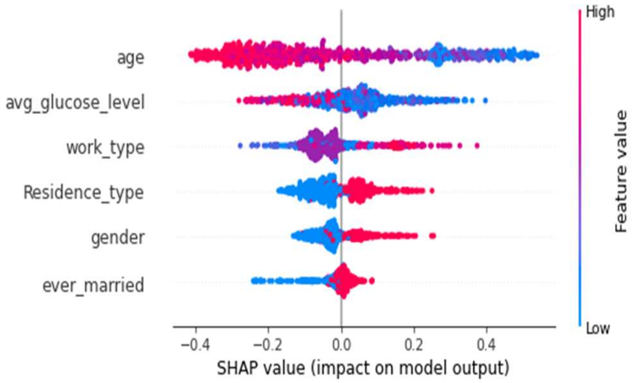
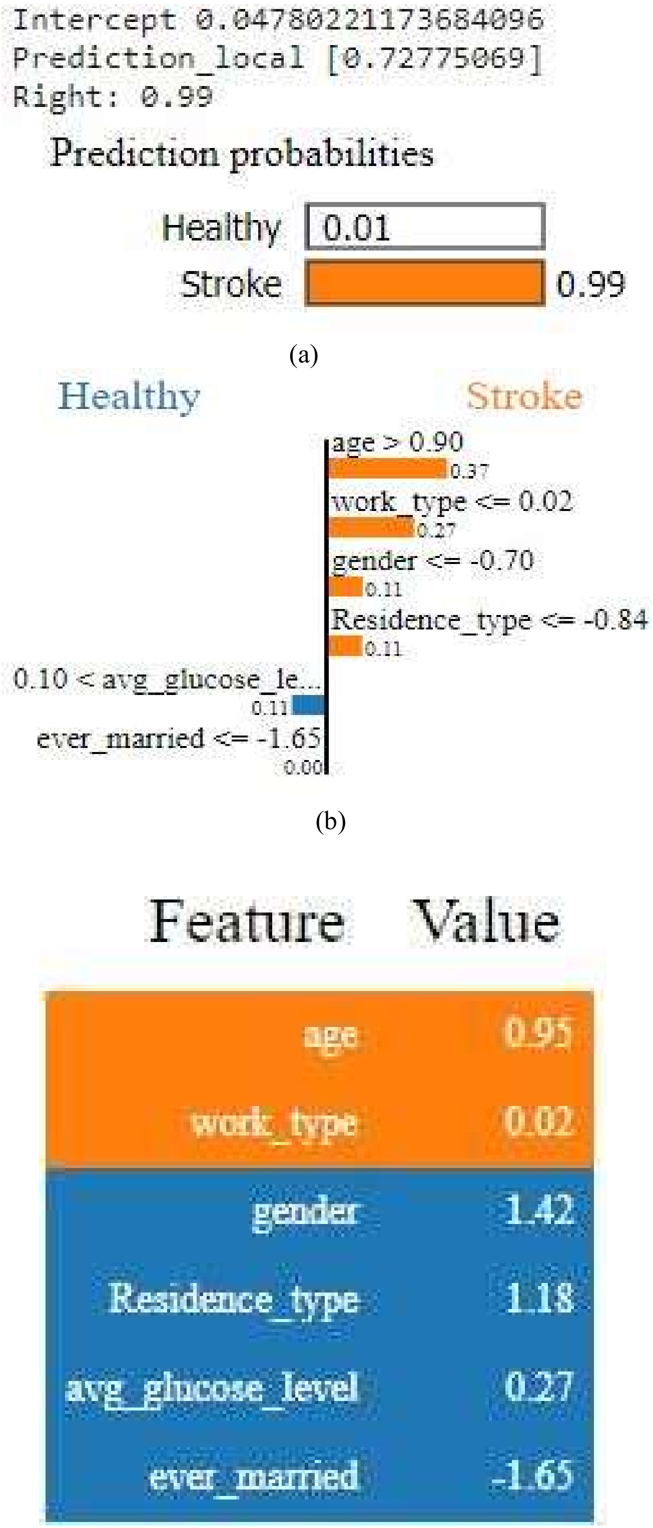


Figure 37: SHAP value impact on Model output (Each Class)

The random forest classifier's mean feature-importance estimates for each class (0, 1) are shown in Figures 36 and 37. The average feature value contributes very little to the total potential collaborations of the output, as shown by the Shapley value, which is displayed on the horizontal axis (x-axis). When the Shapley value is lower than 0, equal to 0, or higher than 0, respectively, a negative, zero, or positive contribution is displayed. The left longitudinal coordinate displays the attributes that are ordered by relevance in reverse order (y-axis). The right longitudinal coordinate, which ranges from low to high, represents the value of the characteristics.

C.2: Local Explanations: Model agnosticism characterizes this approach to constructing local model explanations. Use this model to get a justification for one data point and one classifier. By evenly and randomly choosing the locality surrounding the one selected data point, a collection of changed data points is created along with their corresponding estimate from the model we desire to understand.


```
Intercept 0.04780221173684096
Prediction_local [0.72775069]
Right: 0.99
```

(a)

(b)

Feature	Value
age	0.95
work_type	0.02
gender	1.42
Residence_type	1.18
avg_glucose_level	0.27
ever_married	-1.65

(c)

Figure 38: (a), (b), (c): LIME Explanations of one Instance

Figure 38 illustrates two instances of the dataset's explanations. Even if other classes have a chance of being predicted, the initial result, in this case, the first prediction belongs to the category "Healthy" and from the feature value, we can see the important features that have more impact in this prediction. It is the same as for another instance of predicting where another prediction is “Stroke” and thus the XAI explains the prediction by providing the features and important information so that the physician and patients understand and believe in computer-aided diagnoses. The local interpretation is really difficult with ML models. LIME, though, does a great job here. If we investigate how the model entered this result, we can see that there is a strong push and pull impact on the location of the forecast.

The Average feature impact calculates the average impact of each feature on the model's output across Average feature impact calculates the average effect of feature on the model's output across all instances in the dataset. It measures how much the model's output changes when the value of a feature is changed while keeping the values of all other features constant. It doesn't consider interactions between features and is typically used for linear models or decision trees.

On the other hand, SHAP value impact is a local measure that calculates the contribution of each feature for a specific instance, taking into account the interaction between features. SHAP values can be used for any model and don't rely on the model's assumptions or architecture. In multi-class classification problems, SHAP values can be calculated for each class separately, which can provide more detailed insights into the model's behaviour.

V. Integration of the Model with Web Technology to Create a Web-application

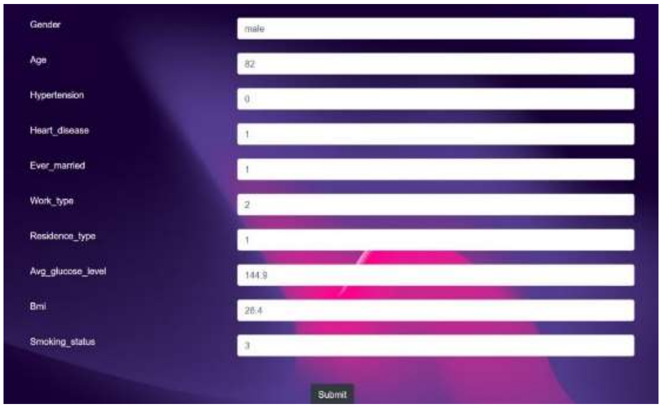
This web application is developed to identify whether a particular person is diagnosed with a stroke or not, which uses the Machine Learning model that we have made, and based on that whatever input the user has provided to the input fields in the web application will predict the results and redirect to the output page based on the output from machine learning model. To accomplish this task we have used HTML, CSS, JS, BootStrap, Flask, and Python. It doesn't matter how great the model is if the normal users or the targeted audience don't understand or know how to use it. Here we are addressing the issues of normal users where they can check their status of stroke whenever necessary, it is much advised that those tests and results are supervised by doctors or specialists.

Flask is a web application framework written in python, which helps end users interact with python code which is our ML model without needing the necessary libraries, code users can use directly without hassle. Flask is based on two components: the WSGI toolkit and the Jinja template engine, this toolkit is a specification for web apps and the template engine is to render web pages. Flask helps the user to take input from the browser and run the models, in our case if the prediction value is "0" in the model we redirect the browser to a new page that says no chances of stroke else to the page which says the risk of stroke. This web application is tested and run on a local machine. Bootstrap and CSS are mainly used for designing web pages and adding styles that make everything interesting. The folder structure goes as follows:

Templates: - This folder contains the HTML files that would be used by our main file (app.py) to generate the front end of our application

app.py: - This is the main application file, where all our code resides and it binds everything together.

Model: - This folder contains models, that we would be using, in this case, we have trained already.

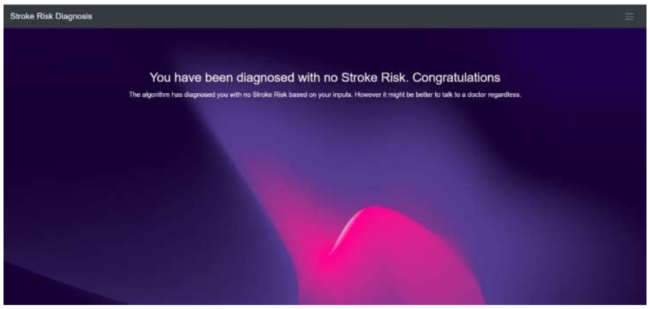


Gender	male
Age	82
Hypertension	0
Heart_disease	1
Ever_married	1
Work_type	2
Residence_type	1
Avg_glucose_level	144.9
Bmi	28.4
Smoking_status	3

Figure 39: User Interface Home Page:

In Figure 39, we have home.html which contains ten input fields and a submit button. It is designed and styled with the help of Bootstrap, Html, and CSS. There is a navigation bar at the top which will help to navigate from one page to another. In the home.html page, it contains the following input fields: Gender, Age, Hypertension, Heart_Disease, Ever_Married, Work_Type, Residence_Type, Avg_Glucose_Level, BMI, Smoking_Status.

In each field, users have to enter the numbers only, otherwise, the form won’t submit. In the Gender field if the user is male then enter ‘1’ else if the user is female, then enter ‘0’. In the Age field, users can simply enter their age in number. For Hypertension, Heart_Disease, and Ever_Married enter ‘1’ for Yes and ‘0’ for No. Likewise, if the user is working for a government job enter ‘0’, ‘1’ if the user is unemployed till now, ‘2’ for working in the private sector, ‘3’ for self-employed, and ‘4’ for children in the field Work_Type. In the case of Residence_Type enter ‘0’ if the user is living in a Rural area, if not then enter ‘1’. For Avg_Glucose_Level just enters the glucose level in number and the same goes for BMI too. The last field is for inserting the Smoking_Status where the user needs to enter ‘0’ for unknown and ‘1’ for formerly smoked, ‘2’ for never smoked, and ‘3’ for smoking daily. At the bottom of the page, there is a submit button, after filling in all the fields the user can submit the form and get the result on the next page.



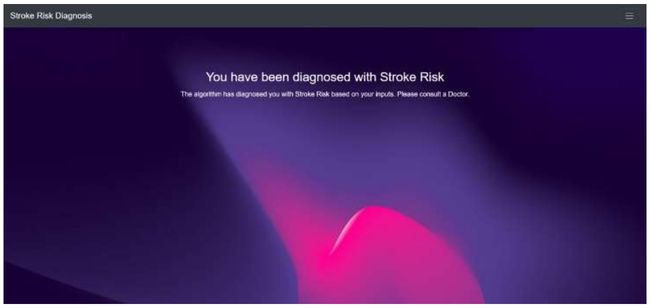
Stroke Risk Diagnosis

You have been diagnosed with no Stroke Risk. Congratulations

The algorithm has diagnosed you with no Stroke Risk based on your inputs. However it might be better to talk to a doctor regardless.

Figure 40: Output for Stroke Positive patients

In Figure 40, we have an output page that tells you have been diagnosed with no stroke risk. Based on the data you have given us the machine learning model predicts no chances of a stroke at the moment and is displayed on this page.



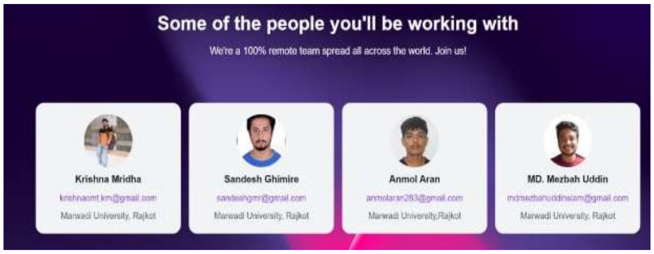
Stroke Risk Diagnosis

You have been diagnosed with Stroke Risk

The algorithm has diagnosed you with Stroke Risk based on your inputs. Please consult a Doctor.

Figure 41: Output for Healthy patients

In Figure 41, the user sees an output page that tells if he/she has been diagnosed with stroke risk or if there are chances of having a stroke. Based on the data the user has provided, the machine learning model predicts if the user has been diagnosed with a stroke or if the user has not been diagnosed with the stroke. However, it is best to get an appointment with a doctor regardless.



Some of the people you'll be working with

We're a 100% remote team spread all across the world. Join us!

Krishna Mridha
krishnaomt.km@gmail.com
Marwadi University, Rajkot

Sandesh Ghimire
sandeshgmr@gmail.com
Marwadi University, Rajkot

Anmol Aran
anmolaran283@gmail.com
Marwadi University,Rajkot

MD. Mezbah Uddin
mdmezbahuddinislam@gmail.com
Marwadi University, Rajkot

Figure 42: Team Member

Figure 42, the details of our research team members or those who have put their effort into research and development to develop this model and implement it on the web.

VI. Discussion and Future Scopes

The study investigated which factors may have been important in how a black box model reacts to changing images. However, users must be able to interpret the information when it is provided to them. The study only employed one trained model, therefore various model designs and training datasets may yield different results. To give domain-specific explanations for ML models, a method for doing so has to be developed. In the future, it's possible to combine different samples when using the explanation technique. Further research and the exploration of other metrics for assessing explanations are necessary given the observed importance of feature dimensions to the real score when applying the ABCD rule.

The computational cost of implementing a stacked cross-validation approach is substantial. However, this is a one-time offline operation, and the trained model has a quick inference time, allowing for near-real-time deployment. The proposed analysis used the entire feature set, which could be seen as a limitation, but no feature selection was performed to determine the contribution of each feature (via SHAP values) to the stroke prediction outcome, which is useful for optimizing future experimental designs based on the most relevant risk factors. The lack of an external validation dataset to test the generalization of the best ML model is a restriction. Our next work will include the creation of an end-to-end smart stroke prediction system that will use an android and iOS application in the patient and real hand. The program is user-friendly for the patient. They can either upload photographs from memory or screen the damaged area to categorize the skin illness. Not only do they obtain the categorization report, but they may also call a doctor from the medical list and receive medicine and counselling. If necessary, the patient and physician can also track earlier recorded information to validate the updated information.

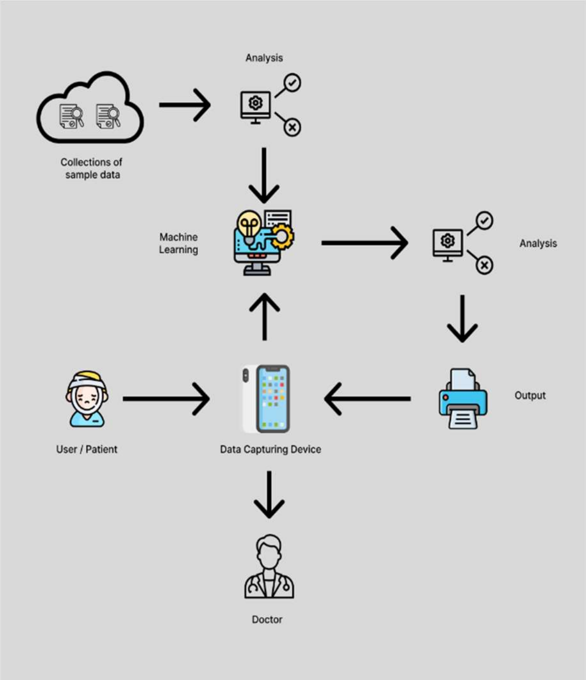
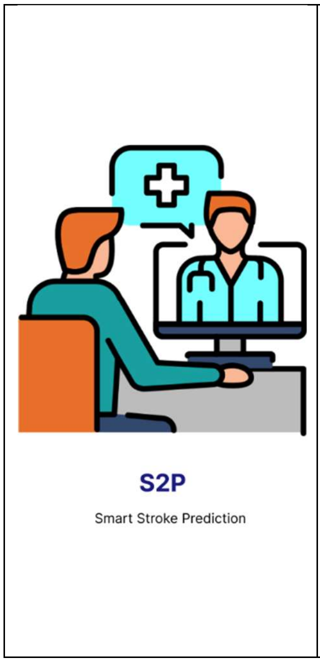


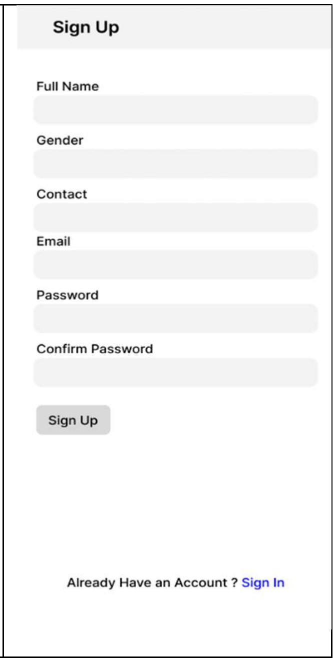
Figure 43: Proposed Future Work done by Hospitals

Table 8: Cross Platform (Android and iOS) user interface which will be used for Stroke diagnosis and Doctor Consultancy



S2P

Smart Stroke Prediction



Sign Up

Full Name

Gender

Contact

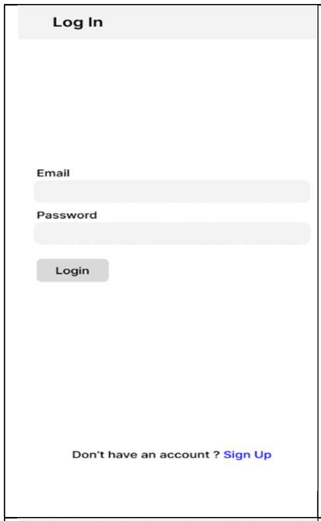
Email

Password

Confirm Password

Sign Up

Already Have an Account ? Sign In



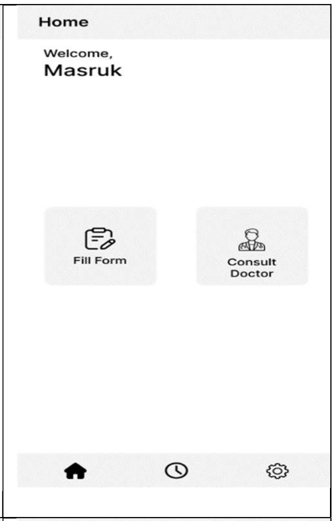
Log In

Email

Password

Login

Don't have an account ? Sign Up

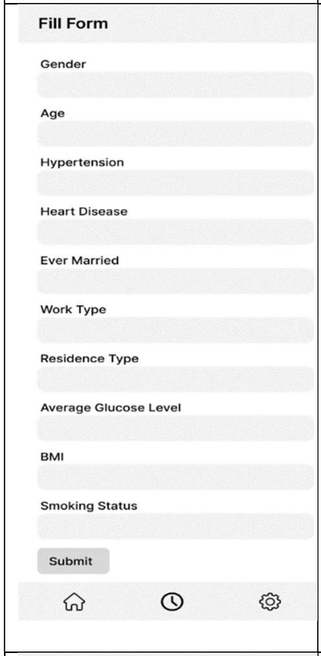


Home

Welcome,
Masruk

Fill Form

Consult Doctor



Fill Form

Gender

Age

Hypertension

Heart Disease

Ever Married

Work Type

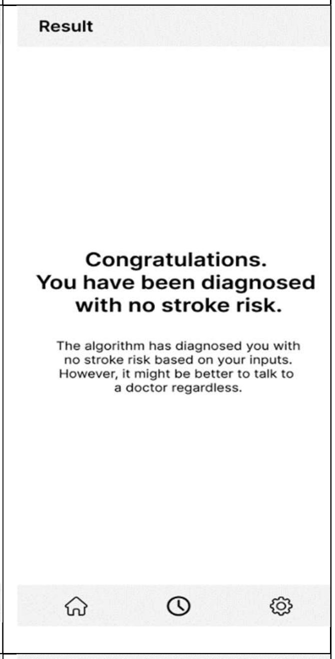
Residence Type

Average Glucose Level

BMI

Smoking Status

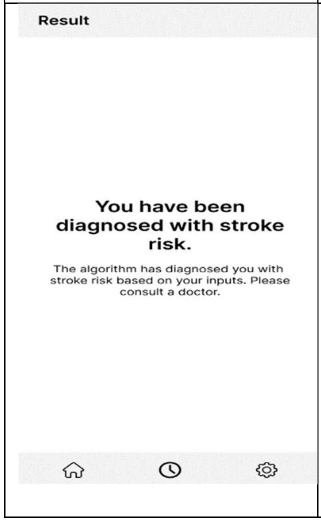
Submit



Result

Congratulations. You have been diagnosed with no stroke risk.

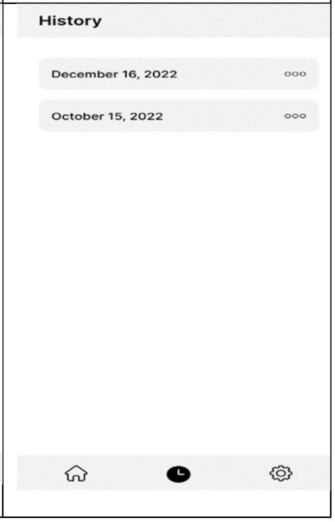
The algorithm has diagnosed you with no stroke risk based on your inputs. However, it might be better to talk to a doctor regardless.



Result

You have been diagnosed with stroke risk.

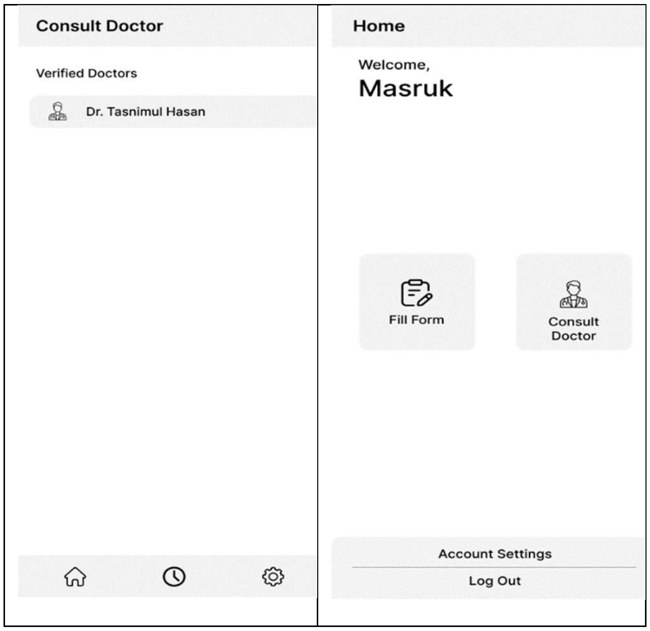
The algorithm has diagnosed you with stroke risk based on your inputs. Please consult a doctor.



History

December 16, 2022

October 15, 2022



Working Model

Registration: The user needs to signup/ register to be able to access the features of this application. Once the account has been created, they can sign in to the application using the same.

Data Form: The logged-in user has the option to fill up the form with the form fields with the dataset known to him/her. After filling up the necessary details, the user needs to submit the form.

Dataset analysis: The dataset provided by the user will be analyzed through the machine learning model and the output result will be provided to the user after the dataset is analyzed.

Result: The user will be able to check the result of the test once the analysis has been completed. Users will have a history section where his/her past reports will be easily visible.

Consulting a doctor: Once the report has been given to the user, the user will be able to share the report file with the doctor. The doctor will look after the report and consult with the user/patient.

VII. Conclusion

A clinical decision support system's Stroke Prediction can serve as a second option in Computer Aided Diagnosis. Although a large research community has helped, these AI-based systems can only make predictions and cannot explain their rationale. This is where XAI approaches come in. We demonstrated how to approach Stroke Prediction in a domain-specific manner. For example, if a physician identifies as a Stoke patient but the model labels it as healthy, both the doctor and the patient may wonder "why?" Our method includes explanations such as "if the age of the patient is between 62 and 84, the prediction confidence in healthy diagnosis drops." The clinician may then notice the age limit in the electronic health records, which is not evident in the disease, and figure out why the model was predicted incorrectly. Whether the clinical decision support system supports or opposes the physician's diagnosis, offering human-readable reasons fosters confidence and improves system knowledge. Furthermore, our perturbation-based explanation technique for diagnosis employing medically relevant and irrelevant characteristics may have implications in other medical domains.

ACKNOWLEDGEMENTS AND FUNDING

This work was supported by the Competitive Research Fund of The University of Aizu, Japan.

REFERENCES

[1] Learn about Stroke. Available online: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learnabout-stroke (accessed on 25 May 2022).

[2] Elloker, T.; Rhoda, A.J. The relationship between social support and participation in stroke: A systematic review. Afr. J. Disabil. 2018, 7, 1–9.

[3] Katan, M.; Luft, A. Global burden of stroke. In Seminars in Neurology; Thieme Medical Publishers: New York, NY, USA, 2018; Volume 38, pp. 208–211.

[4] Bustamante, A.; Penalba, A.; Orset, C.; Azurmendi, L.; Llombart, V.; Simats, A.; Pecharroman, E.; Ventura, O.; Ribó, M.; Vivien, D.; eta. Blood biomarkers to differentiate ischemic and hemorrhagic strokes. Neurology 2021, 96, e1928–e1939.

[5] Xia, X.; Yue, W.; Chao, B.; Li, M.; Cao, L.; Wang, L.; Shen, Y.; Li, X. Prevalence and risk factors of stroke in the elderly in Northern China: Data from the National Stroke Screening Survey. J. Neurol. 2019, 266, 1449–1458.

[6] Alloubani, A.; Saleh, A.; Abdelhafiz, I. Hypertension and diabetes mellitus as a predictive risk factor for stroke. Diabetes Metab.

[7] Syndr. Clin. Res. Rev. 2018, 12, 577–584. Boehme, A.K.; Esenwa, C.; Elkind, M.S. Stroke risk factors, genetics, and prevention. Circ. Res. 2017, 120, 472–495.

[8] Mosley, I.; Nicol, M.; Donnan, G.; Patrick, I.; Dewey, H. Stroke symptoms and the decision to call for an ambulance. Stroke 2007, 38, 361–366.

[9] Lecouturier, J.; Murtagh, M.J.; Thomson, R.G.; Ford, G.A.; White, M.; Eccles, M.; Rodgers, H. Response to symptoms of stroke in the UK: A systematic review. BMC Health Serv. Res. 2010, 10, 1–9.

[10] Gibson, L.; Whiteley, W. The differential diagnosis of suspected stroke: A systematic review. J. R. Coll. Physicians Edinb. 2013, 43, 114–118.

[11] Murray, N.M.; Unberath, M.; Hager, G.D.; Hui, F.K. Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: A systematic review. J. NeuroInterv. Surg. 2020, 12, 156–164.

[12] Zhao, Y.; Fu, S.; Bielinski, S.J.; A Decker, P.; Chamberlain, A.M.; Roger, V.L.; Liu, H.; Larson, N.B. Natural Language Processing and Machine Learning for Identifying Incident Stroke from Electronic Health Records: Algorithm Development and Validation. J. Med. Internet Res. 2021, 23, e22951.

[13] McDermott, B.J.; Elahi, A.; Santorelli, A.; O'Halloran, M.; Avery, J.; Porter, E. Multi-frequency symmetry difference electrical impedance tomography with machine learning for human stroke diagnosis. Physiol. Meas. 2020, 41, 075010.

[14] Bivard, A.; Churilov, L.; Parsons, M. Artificial intelligence for decision support in acute stroke—Current roles and potential. Nat. Rev. Neurol. 2020, 16, 575–585.

[15] Wang, W.; Kiik, M.; Peek, N.; Curcin, V.; Marshall, I.J.; Rudd, A.G.; Wang, Y.; Douiri, A.; Wolfe, C.D.; Bray, B. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS ONE 2020, 15, e0234722.

[16] Sirsat, M.S.; Fermé, E.; Câmara, J. Machine learning for brain stroke: A review. J. Stroke Cerebrovasc. Dis. 2020, 29, 105162.

[17] Arslan, A.K.; Colak, C.; Sarihan, M.E. Different medical data mining approaches-based prediction of ischemic stroke. Comput. Methods Programs Biomed. 2016, 130, 87–92.

[18] Islam, M.S., Hussain, I., Rahman, M.M., Park, S.J. and Hossain, M.A., 2022. Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal. *Sensors*, *22*(24), p.9859.

[19] Dritsas, E. and Trigka, M., 2022. Stroke risk prediction with machine learning techniques. *Sensors*, *22*(13), p.4670

[20] Kokkotis, C., Giarmatzis, G., Giannakou, E., Moustakidis, S., Tsatalas, T., Tsiptsios, D., Vadikolias, K. and Aggelousis, N., 2022. An Explainable Machine Learning Pipeline for Stroke Prediction on Imbalanced Data. *Diagnostics*, *12*(10), p.2392

[21] Islam, R., Debnath, S. and Palash, T.I., 2021, December. Predictive Analysis for Risk of Stroke Using Machine Learning Techniques. In *2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.

[22] Darabi, N., Hosseinichimeh, N., Noto, A., Zand, R. and Abedi, V., 2021. Machine learning-enabled 30-day readmission model for stroke patients. *Frontiers in neurology*, *12*, p.638267.

[23] Youngkeun Choi, Jae Won Choi. Stroke Prediction Using Machine Learning based on Artificial Intelligence. International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Volume 9, No.5, September - October 2020.

[24] Tazin, T., Alam, M.N., Dola, N.N., Bari, M.S., Bourouis, S. and Monirujjaman Khan, M., 2021. Stroke Disease Detection and Prediction Using Robust Learning Approaches. *Journal of Healthcare Engineering*, *2021*.

[25] H. K V, H. P, G. Gupta, V. P, and P. K B, "STROKE PREDICTION USING MACHINE LEARNING ALGORITHMS," International Journal of Innovative Research in Engineering &amp; Management, vol. 8, no. 4, Jul. 2021, doi: 10.21276/ijirem.2021.8.4.2.

[26] Dev, S., Wang, H., Nwosu, C.S., Jain, N., Veeravalli, B. and John, D., 2022. A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics*, *2*, p.100032.

[27] Ali, A.A., 2019. Stroke prediction using distributed machine learning based on Apache spark. *Stroke*, *28*(15), pp.89-97.

[28] K. Mridha, S. Kumbhani, S. Jha, D. Joshi, A. Ghosh, and R. N. Shaw, "Deep Learning Algorithms are used to Automatically Detection Invasive Ducal Carcinoma in Whole Slide Images," 2021 IEEE 6th International Conference on Computing, Communication, and Automation (ICCCA), Arad, Romania, 2021, pp. 123-129, doi: 10.1109/ICCCA52192.2021.9666302.

[29] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information fusion, 58, pp.82-115.

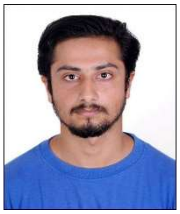
[30] K. Mridha, M. M. Shorna, N. Arefin, A. Ritu, M. M. Alam Chowdhury, and M. I. Islam, "DBNet: Detect Diabetic Retinopathy to Stop Blindness Before it's Too Late," 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2022, pp. 1-6, doi: 10.1109/ICRITO56286.2022.9964937.

[31] K. Mridha, M. I. Islam, M. M. Shorna and M. A. Priyok, "ML-DP: A Smart Emotion Detection System for Disabled Person to Develop a Smart City," 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2022, pp. 1-6, doi: 10.1109/ICRITO56286.2022.9965131.

[32] K. Mridha, A. Ritu, M. M. A. Chowdhury, and N. Arefin, "ML-MT: A Study of e-Health Application Framework by Machine Learning Techniques," 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), Goa, India, 2022, pp. 337-342, doi: 10.1109/ICCCMLA56841.2022.9989049.

[33] T. Dhar, R. V. Satapathy, N. R. Pal, and T. Acharya, "Challenges of Deep Learning in Medical Image Analysis-Improving Explainability and Trust," IEEE Transactions on Technology and Society, vol. 1, no. 1, pp. 46-54, Jan. 2023.

Krishna Mridha (Member, IEEE) was born in Madaripur, Dhaka, Bangladesh in 1997. Currently, he is working as a Research Assistant under the Department of Computer Engineering - Artificial Intelligence at Marwadi University, Rajkot, Gujarat, India. From 2017 to 2018, he was a Junior Instructor of Information and Communication Technology at a private medical technology institute in Dhaka, Bangladesh. At the end of 2019 he came to India to pursue his bachelor's degree in Computer Engineering through the Government of India Initiative called "Study in India". In his bachelor's journey, he published some research articles related to medical diagnosis. He won the IEEE student beset student research paper award in 2021.



Sandesh Ghimire is a computer Engineering student with extensive experience and programming skills in his field. He is from Dandabazar, Dhankuta, Nepal. Currently, He is pursuing his Bachelor's degree in Computer Engineering in India at Marwadi University through the Study in India Scholarship which is sponsored by the Indian government in the year 2019. Before joining Marwadi University, Mr. Sandesh was awarded first prize by the Nepalese government in the competition Talent Hunt program in the field of science and technology in the year 2017. Mr. Sandesh's area of expertise is in Machine Learning and Web Development. His research interest is focused on the field of Artificial Intelligence, Machine Learning in healthcare. His future goal is to pursue a Ph.D. in Artificial Intelligence in the healthcare sector.

Jungpil Shin (Senior Member, IEEE) received a B.Sc. in Computer Science and Statistics and an M.Sc. in Computer Science from Pusan National University, Korea, in 1990 and 1994, respectively. He received his Ph.D. in computer science and communication engineering from Kyushu University, Japan, in 1999, under a scholarship from the Japanese government (MEXT). He was an Associate Professor, a Senior Associate Professor, and a Full Professor at the School of Computer Science and Engineering, The University of Aizu, Japan, in 1999, 2004, and 2019, respectively. He has co-authored more than 300 published papers for widely cited journals and conferences. His research interests include pattern recognition, image processing, computer vision, machine learning, human-computer interaction, non-touch interfaces, human gesture recognition, automatic control, Parkinson's disease diagnosis, ADHD diagnosis, user authentication, machine intelligence, as well as handwriting analysis, recognition, and synthesis. He is a member of ACM, IEICE, IPSJ, KISS, and KIPS. He served as program chair and as a program committee member for numerous international conferences. He serves as an Editor of IEEE journals and for MDPI Sensors and Electronics. He serves as a reviewer for several major IEEE and SCI journals.

Anmol Aran was born in Manthali, Nepal. Currently, He is pursuing a Bachelor's degree in India. His research interests lie in the field of machine learning and deep learning in the healthcare sector. His future goal is to pursue Ph.D. in deep learning and continue to contribute to the development and implementation of machine learning and deep learning in different industries.

Md. Mezbah Uddin was born in Cumilla, Bangladesh. Currently, He is pursuing Bachelor's degree in Computer Engineering in India through the Study in India Scholarship, which is sponsored by the Indian government. His research interests lie in the field of Artificial Intelligence in healthcare. His future goal is to pursue a Ph.D. in Healthcare AI and continue to contribute to the development and implementation of AI solutions in the healthcare industry.

M. F. Mridha (Senior Member, IEEE) is currently working as an Associate Professor in the Department of Computer Science at American International University-Bangladesh (AIUB). Before that he worked as an Associate Professor and Chairman in the department of CSE of Bangladesh University of Business and Technology. He also worked as a CSE department faculty member at the University of Asia Pacific and as a graduate head from 2012 to 2019. He received his Ph.D. in AI/ML from Jahangirnagar University in the year 2017. His research experience, within both academia and industry, results in over 120 journal and conference publications. His research work contributed to the reputed Journal of Scientific Reports–Nature, Knowledge-Based Systems, Artificial Intelligence Review, IEEE Access, Sensors, Cancers and Applied Sciences, etc. His research interests include artificial intelligence (AI), machine learning, deep learning, and natural language processing (NLP). For more than 10 (Ten) years, he has been with the masters and undergraduate students as a supervisor of their thesis work. His research interests include artificial intelligence (AI), machine learning, natural language processing (NLP), big data analysis, etc. He has served as a program committee member in several international conferences/workshops. He served as an associate editor of several journals including PLOS ONE Journal. He has served as a reviewer of reputed journals and IEEE conferences like HONET, ICIEV, ICCIT, IJCCI, ICAEE, ICCAIE, ICSIPA, SCORED, ISIEA, APACE, ICOS, ISCAIE, BEIAC, ISWTA, IC3e, ISWTA, CoAST, icIVPR, ICSCT, 3ICT, DATA21, etc**.**