International Journal of Innovations in Science & Technology

Prediction of Brain Stroke Using Federated Learning

Rabia Tehseen[1*], Uzma Omer[2], Rubab Javaid[1], Maham Mehr[1], Madiha Yousaf[1], Ayesha Zaheer[1]
[1]University of Central Punjab, Lahore, Pakistan 57400
[2]University of Education, Lahore, Pakistan, 57400
Corresponding author: Rabia Tehseen (e-mail: rabia.tehseen@ucp.edu.pk).
Citation | Tehseen. R, Omer. U, Javaid. R, Mehr. M, Yousaf. M, Zaheer. A, "Prediction of Brain Stroke using Federated Learning", IJIST, Vol. 06 Issue. 4 pp 1995-2013, Dec 2024
Received | Nov 15, 2024 **Revised |** Dec 10, 2024 **Accepted |** Dec 11, 2024 **Published |** Dec 13, 2024.

Stroke often arises from an abrupt blockage in the blood vessels supplying the brain and heart. Detecting early warning signs of stroke can significantly reduce its impact. In this study, we propose an early prediction method for stroke using various Machine Learning (ML) techniques, considering factors such as hypertension, body mass index, heart disease, average glucose levels, smoking habits, prior stroke history, and age. These attributes, rich in information, were utilized to train three distinct classifiers: Logistic Regression, Decision Tree, and K-nearest neighbors for stroke prediction. In this study, Federated Learning (FL) has been applied to combine the ML models (Logistic Regression (LR), Decision Tree (DT), and K-Nearest Neighbors (KNN)) from distributed medical data sources while preserving patient privacy. By aggregating locally trained models from multiple hospitals or devices, FL ensures the robustness of the weighted voting classifier without requiring direct data sharing, thereby enhancing stroke prediction accuracy across diverse datasets. Subsequently, the results from these base classifiers were combined using a weighted voting approach to achieve the highest accuracy. Our study demonstrated an impressive accuracy rate of 97%, with the weighted voting classifier outperforming the individual base classifiers. This model proved to be the more accurate in predicting strokes. Additionally, the Area Under the Curve (AUC) value for the weighted voting classifier was notably high, and it exhibited the lowest false positive and false negative rates compared to other classifiers. Consequently, the weighted voting classifier emerged as an almost ideal tool for predicting strokes, offering valuable support to both physicians and patients in identifying and preventing potential stroke incidents.

Keywords: Stroke, Machine Learning, CM, AUC, WV, Correlation Matrix

Machine Learning	ML
Confusion Matrices	CM
Area Under the Curve	AUC
Weighted Voting	WV
False Positive	FP
False Negative	FN
Federated Learning	FL
Logistic Regression	LR
Decision Tree	DT
K-Nearest Neighbors	KNN
World Health Organization	WHO
Deep Learning	DL
Stochastic Gradient Descent	SGD
Quadratic Discriminant Analysis	QDA
Multi-layer Perceptron	MLP
Gradient Boosting Classifier	GBC
XGBoost	XGB
Exploratory Data Analysis	EDA
Support Vector Machines	SVM
Neural Network	NN
Deep Neural Network	DNN

Introduction:

A stroke occurs when there is a disruption or reduction in blood flow to various regions of the brain, leading to a lack of essential nutrients and oxygen for the affected cells, which then begin to deteriorate. It is recognized as a medical emergency; prompt care is imperative. Early detection and appropriate management are crucial to minimize further damage to the affected brain area and other potential complications throughout the body. According to the World Health Organization (WHO), an alarming fifteen million people worldwide suffer from strokes every year, with individuals succumbing to this condition every 4-5 minutes.

Strokes are mainly classified into two types: ischemic and hemorrhagic. In ischemic strokes, blood flow is obstructed by clots, while hemorrhagic strokes involve the rupture of a weakened blood vessel, causing bleeding within the brain. Stroke prevention can be achieved through a healthy and balanced lifestyle, which involves the elimination of detrimental habits such as smoking and excessive alcohol consumption, as well as the management of Body Mass Index (BMI) and average glucose levels. Maintaining the well-being of the heart and kidneys is also vital. Predicting strokes is essential, as timely intervention can prevent permanent damage or even save lives. This paper focuses on the parameters of hypertension, BMI level, heart disease, and average glucose levels for stroke prediction. Furthermore, ML plays a pivotal role in the decision-making processes within this prediction system [1] [2] [3]. In the existing literature, there is a scarcity of documented works that utilize ML models for stroke prediction [4] [5] [6] [7] [8] [9].

Rahman et al. [1] used DL and ML techniques for the prediction of brain stroke at an early stage. They collected a dataset from Kaggle and trained several classification models. The highest classification accuracy was achieved by the Random Forest classifier, which was 99%. The comparative results of the study demonstrated that ML methods performed better than deep neural networks. Shobayo et al. [2] proposed the random forest algorithm and demonstrated that random forest performs better than other ML models for example DT and LR. They also focused that BMI is also important indicator for the occurrence of stroke.

Gahiwad et al. [3] used a convolutional neural network to identify brain strokes using CT-Scan pictures. They achieved the greatest accuracy of 90% after training and testing the model on a CT-scan dataset consisting of 2551 pictures. Singhal et al. [4] suggested a model that forecasts whether the patient is likely to experience a stroke or not. This assessment was based on input variables such as age, average glucose level, smoking status, BMI, etc. The dataset has been trained using a variety of ML and DL methods. The DT approach exhibited highest accuracy rate, at 96.1%.

Ruban et al. [5] predicted the risk of stroke utilizing five ML approaches which were evaluated using a range of metrics. According to the study's findings, random forest algorithms are more accurate than other algorithms. Deepthi [6] collected the datasets of several patients, and various ML techniques are available for prediction. They used the K-Nearest Neighbor with Random Forest methods for prediction. Based on medical data, Vamsi Bandi et. al. [7] employed ML approaches to forecast strokes. They proposed the Stroke Prediction (SPN) technique, which uses an adapted random forest to analyze the degrees of risk gained inside strokes.

As compared to other techniques, this model improved the prediction accuracy to 96.97%. Gangavarapu Sailasya et al. [8] employed a variety of ML algorithms to predict the likelihood of a brain stroke. They trained 5 different models using ML techniques to provide correct predictions. Among them, Naive Bayes was the most effective algorithm for this task, with an efficiency of about 82%. The increasing rate of brain stroke is alarming. Jalal et al. [9] discussed the causes of stroke and developed an integrated software system that will use ML and artificial NN to predict the stroke. They suggested that teaching individuals might lower its frequency. Saini et al. [10] used four different classifiers for brain stroke prediction and by comparing results concluded that Random forest is very good in performance and achieved 95.02% accuracy [11].

Objective:

The objective of this research is to develop a privacy-preserving, collaborative model that can accurately identify the risk of stroke based on patient data distributed across multiple healthcare institutions. This approach ensures data security and compliance with privacy regulations while leveraging diverse, large-scale datasets to improve prediction accuracy. Key goals include:

1. **Privacy Preservation:** Protectection of sensitive patient data by training models locally and sharing only model updates.
2. **Improved Prediction Accuracy:** Utilization of distributed datasets to enhance model generalizability and reduce bias.
3. **Data Compliance:** Adhering to regulations like GDPR and HIPAA by avoiding centralized data storage.
4. **Cross-Institutional Collaboration:** Assistance of cooperative research without compromising individual data ownership.
5. **Timely Stroke Risk Assessment:** Provision of reliable predictions to enable early intervention and reduce stroke-related morbidity and mortality.

Novelty:

The novelty of our research lies in several aspects, such as:

1. **Privacy-Preserving Collaboration:** Utilizing FL ensures that sensitive medical data remains localized, avoiding direct sharing of patient information while enabling multi-institutional model training.
2. **Enhanced Generalization:** The model leverages diverse data from multiple institutions, improving its ability to generalize across different populations and reducing dataset bias, a challenge in traditional centralized approaches.
3. **Regulatory Compliance:** The approach adheres to strict healthcare data privacy regulations (e.g., GDPR, HIPAA) by keeping data decentralized.
4. **Domain-Specific Optimization:** Incorporating stroke-specific medical features, such as clinical histories, imaging biomarkers, or genetic data, into FL models for superior stroke prediction accuracy.
5. **Scalable and Efficient Training:** Innovations in FL frameworks tailored to the medical domain, such as handling data heterogeneity or optimizing communication efficiency for real-world healthcare systems.
6. **Real-Time Deployment:** Designing models that can be deployed in real-time clinical settings, providing predictions to clinicians without data transfer delays or risk.
7. **Interdisciplinary Integration:** Combining state-of-the-art FL techniques with domain expertise in neurology to address unique challenges in stroke prediction.

The rest of the article is arranged in the following sections: section II presents related work, Section III provides methodology, an experiment is performed in Section IV, section V presents results, and Section VI provides a discussion the results. In the end, section VII concludes the research and provides future directions.

Related Work:

Mushtaq et al. [12] reviewed different research articles on brain stroke prediction using ML and suggested that the most commonly used methods for the prediction are Support Vector Machine, Stacking, DT, Weighted Voting, Random Forest, NN, and Naive Bayes. Victor [13] proposed a system for brain stroke prediction based on ML. In order to maintain the patient's data privacy FL was incorporated into the framework. Ritesh Kumari and Hitendra Garg [14], used various ML algorithms for brain stroke prediction and compared their results. According to the findings, Gradient Boost comes in second to Random Forest in terms of prediction accuracy.

Chandrabhatla et al [15] stated that ML-based FDA-approved technologies and devices can assist medical professionals in more accurately diagnosing and treating stroke. Zhang et al. [16] reviewed the literature to assess the significance of DL techniques and concluded that the diagnosis, therapy, and prognosis of stroke are significantly impacted by DL methods. To train five different models for precise brain stroke prediction, the authors in [8] analyzed a variety of physiological factors and used ML algorithms. Naive Bayes achieved the highest accuracy of about 82%. Akter et al. [17] suggested a model i.e. Random forest for predicting brain stroke with 95.30% accuracy. Chavva et al. [18] presented that DL techniques are very helpful in acute stroke management as they aid decision-making. In the proposed framework, DL techniques help physicians to solve problems in clinical practice. Ferdib-Al-Islam and Mounita Ghosh [19] used the oversampling method with various ML classifiers for brain stroke prediction. Among all, the random forest model achieved 99.07% accuracy, 99.0% precision, and 99.0% recall.

Kaur et al. [20] suggested a noninvasive method for the early diagnosis of strokes. The goal was to develop a method for forecasting strokes using EEG data. GRU outperforms all other algorithms used in the research with 95.6% accuracy. Premisha et al. [21] into an ensemble model, which predicted stroke severity with 95.76% accuracy. Sirsat et al. [22] classified the state-of-the-art ML approaches for brain stroke into 4 groups based on their functionality or resemblance. They highlighted the value of various ML techniques used in brain stroke. The effective methods employed for each category include SVM and Random Forest. Gaidhani et al. [23], identified the brain strokes from MRI images using CNN and DL models. Experimental findings revealed that DL models are effective not only for non-medical images but also for providing reliable results in medical imaging, especially in identifying brain strokes. Heo et al. [24] employed a variety of ML algorithms based on certain attributes and concluded that Deep Neural Networks (DNNs) have an impact on long-term predicting and are frequently used for ischemic stroke patients.

A.P.V. Rohit et al [25] used two ML approaches i.e. Naïve Bays and DT and concluded that age, heart disease, average glucose level, and hypertension are the most important factors for detecting stroke in patients. Dr. V. Jyothsna et al [26] ML techniques can aid in the early detection of stroke symptoms, with results suggesting the Random Forest (RF) model achieves a higher accuracy of 96.34% compared to other methods. For the investigation of the best-supervised ML for stroke prediction,[27] this paper has trained the 9 models, compared their accuracies, and achieved the best accuracy of 97% through the LR Model. Another Study [28] developed an ML model to precisely predict whether a person suffering from a stroke or not based on the K-Nearest Neighbor (KNN) Algorithm, Random Forest dataset LR.

A comparison between four algorithms was done in [29] to detect brain stroke with better accuracy. Guangtong Yang et al [30] in a study included a data set of 244 patients and extracted 35 features for model development to capture stroke-associated pneumonia. They assessed the performance of 3 different ML models such as LR, support vector machine, and random forest for the prognosis of the disease. Mouli [31] compared different CNN-based DL models and suggested that Enhanced DenseNet121 exhibited a maximum accuracy of 99.82%. In [32] authors conducted a systematic literature which included 12 studies. Almost fifteen different algorithms were used to predict stroke in patients in them, among which the most commonly used ML algorithm was Support Vector Machine (SVM). The author [33] compared 11 classifiers to evaluate the best model and concluded that SVM provides the highest accuracy of 98.18% in predicting brain stroke in patients. In another research by Hamza Al-Zubaidi et al [34], the Random Forest classifier model produced the best accuracy of 94.6% in stroke prediction using ML. In an investigational study [35], a set of 2501 brain stroke CT images were used to test and train the model. CNN classifiers were used to predict stroke but the model that produced the best accuracy, precision, recall, and F-score of 97%-96.49% was VGG-19. Wang, K., Shi, et al. [36] collected a data set of 645 AIS patients over 1 year, and 6 models were used for the prediction of stroke occurrence.

A framework was proposed in [37] by Uddin et al for the detection of stroke and achieved an accuracy of 99.90 % through this framework. Indicators. Average Glucose Level, Heart Disease, BMI, and Age were identified as keystrokes Clinton [38] developed a prediction model specifically targeting adults aged 25-64 years and achieved 76% recall using a stacking classifier. In [39] classification algorithms were combined with ANOVA to predict brain stroke. Kumar et al. [40] developed a DL-based stroke prediction scheme. Several ML methods were used in [41] for predicting stroke. Z Chen [34] used LR and random forest algorithms for the prediction of stroke risk in patients. AK Uttam [42] analyzed all elements that can cause and influence brain stroke and constructed a model to forecast brain stroke in patients. C Sharma et al. [43] achieved an accuracy of 98.94 % by using a random forest algorithm for predicting stroke prediction. Elias Dritsas and Maria Trigka [44] used stacking classification and achieved 98.9% in predicting the risk of stroke. Hao Ming Xia and Ramin Ramezani [45] used transformer-based models for forecasting brain stroke and compared them with other models.

Sushila Paliwal et al. [46] analyzed a wide range of ML algorithms for predicting brain stroke. Multiple ML methods were used in [47] for stroke prediction. Tianyu Liu, Wenhui Fan, and Cheng Wu [48] used a hybrid ML approach to predict cerebral stroke.

Methodology:

This section is divided into three parts: Data Description, ML Classifiers & Evaluation Matrices, and Implementation Procedures. The processes of these sections is described in Figures 1a, 1b, and 1c.

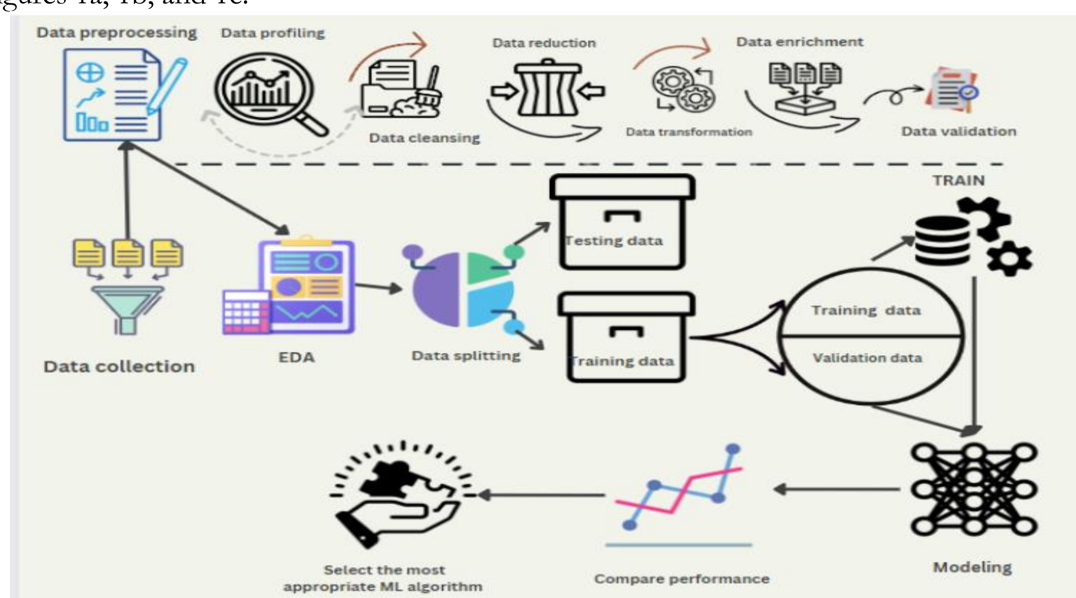

Figure 1a. Flow diagram of Proposed work

A. Dataset Description:

The classified dataset was selected from Kaggle with the title name ("Stroke Prediction Dataset"). It comprises records from 5110 individuals, and each attribute in the dataset is described below [49]:

1. Age: This attribute represents the age of an individual as numerical data.

2. Gender: This attribute denotes an individual's gender as categorical data.

3. Hypertension: This attribute indicates whether the person has hypertension or not, represented as numerical data.

4. Work Type: This attribute describes the individual's work scenario as categorical data.

5. Residence Type: This attribute reflects the person's living scenario as categorical data.

6. Heart Disease: This attribute indicates whether the person has a heart disease or not, represented as numerical data.

7. Avg Glucose Level: This attribute signifies the individual's glucose level, represented as numerical data.

8. BMI (Body Mass Index): This attribute represents the body mass index of an individual and is represented as numerical data.

9. Ever Married: This attribute indicates the marital status of the person as categorical data.

10. Smoking Status: This attribute denotes an individual's smoking condition as categorical data.

11. Stroke: This attribute indicates whether the person has previously had a stroke or not and is represented as numerical data.

Among these attributes, "Stroke" serves as the decision class, while the rest of the attributes are considered response classes.

B. Classifiers Trained:

This section described the ML algorithms used for analysis and the evaluation metrics employed to assess their performance. It provided details about the classifiers selected and the criteria used to evaluate their effectiveness. In this section, we presented the ten ML classifiers used to develop stroke prediction models. These classifiers are: (1) LR, (2) Stochastic Gradient Descent (SGD), (3) Decision Tree Classifier (DTC), (4) AdaBoost, (5) Gaussian Naive Bayes, (6) Quadratic Discriminant Analysis (QDA), (7) Multi-layer Perceptron (MLP), (8) K-Neighbors, (9) Gradient Boosting Classifier (GBC), and (10) XGBoost (XGB). These classifiers were selected due to their widespread use in building vulnerability predictors and their application in numerous similar research studies. We selected these ten classifiers based on their proven effectiveness in previous works [21], [22], which are closely related to our research. Additionally, the performance of these models was assessed using CM to evaluate their accuracy and effectiveness.

C. Implementation Procedure:

This section outlines the implementation process of the study. The analysis was conducted using Python and the Scikit-learn libraries.

1. **Input Data**: We collected data from 5,110 patients, documenting various health conditions related to stroke occurrences. The data was gathered from several hospitals across Bangladesh. All procedures involving human participants adhered to the ethical standards of institutional and national research committees, following the 1964 Helsinki Declaration and its later amendments. Approval for the study was granted by the Non-invasive Ethical Committee of Jahangirnagar University (JU), Dhaka, Bangladesh. All participants provided the necessary consent in compliance with JU's ethical standards.
2. **Data Pre-processing**: The first step in data processing involves checking for missing and duplicate values. Missing values were handled by replacing them with the mean or median of other values. For the attribute "smoking status," which had missing data, values were filled based on the corresponding age group. No duplicate values were found in the dataset. We then normalized the dataset and applied label encoding to convert categorical variables into numerical ones, preparing the data for further analysis.

3. **Data Splitting**: The dataset was divided into training and testing sets using a split technique to ensure effective model evaluation. This separation was essential for model training and subsequent testing.

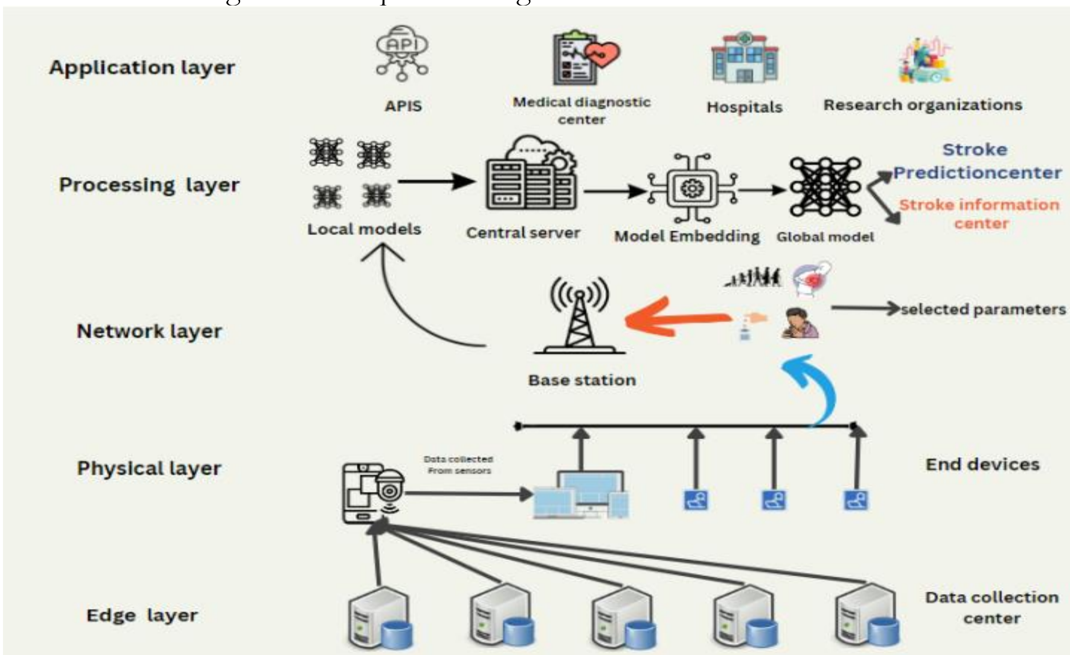

Figure1b. Proposed Methodology.

4. **Base Algorithms**: We used ten different ML algorithms as base models to train and test the proposed approach.
5. **Weighted Voting**: After running all the classifiers, a WV classifier was implemented to enhance the overall accuracy of the individual classifiers.
6. **Model Optimization**: CM was calculated for each model to derive metrics such as precision, recall, F1-score, accuracy, False Positive (FP) rate, and False Negative (FN) rate.
7. **Best Model**: Finally, the accuracy of the ten algorithms was evaluated, and the best model was identified using the WV classifier, which produced the highest accuracy among the tested models.

Generally, the flow of this research is presented in Figure 1c.

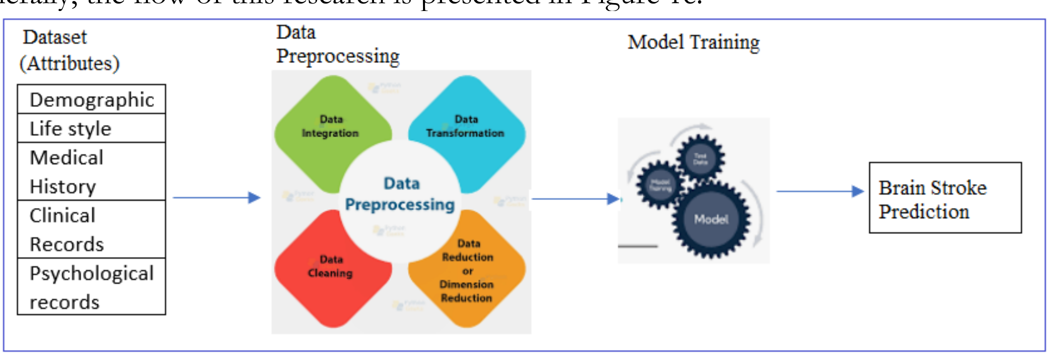

Figure1c. Flow of research

Experimental Verification: During the preprocessing step, the data for training and testing was read, and a classifier was utilized. For this purpose, the following libraries were required to be installed.

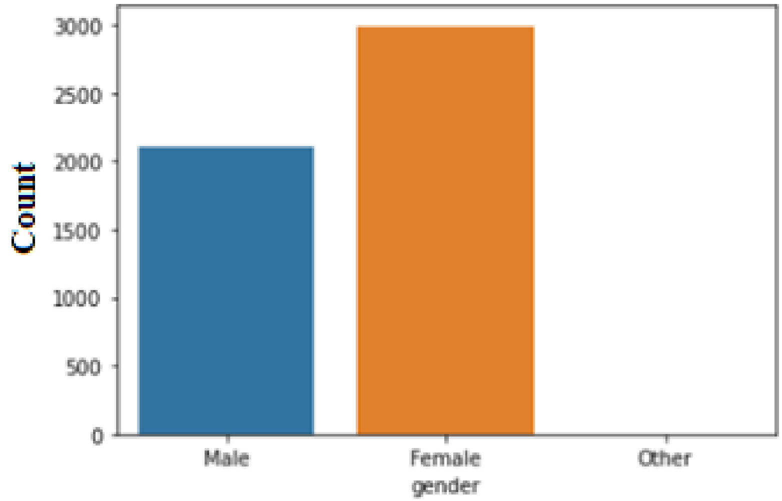

Figure 2. Count plot for gender

Software requirements:

Libraries used

1) pandas 2) matplotlib 3) numpy 4) seaborn 5) sklearn

For dealing with the dataset and the model preparation process, Python programming language was utilized in the Jupyter Note pad stage, which improves information control and representation. Python is a high-level interpreted programming language that is very easy to learn while still capable of leveraging the power of low-level programming languages when needed. Besides these advantages, the local area around the accessible instruments and libraries makes it especially appealing for jobs in information science, AI, also, logical registering [15]

Data Understanding:

Statistical techniques were used to analyze the collected information and to investigate the records/data. The Exploratory Data Analysis (EDA) was applied with Python tools i.e., pandas and NumPy in Jupiter notebook. The summary was calculated that provided the mean, max, min, and std of each attribute with total and class wise. After calculating the summary, the correlation was calculated to find the relation of attributes with each other. The data visualization is important to deeply know about the attributes so the frequency bar chart shows the frequency of each class in the dataset. Figure 2 presents that women are more suspected for brain stroke.

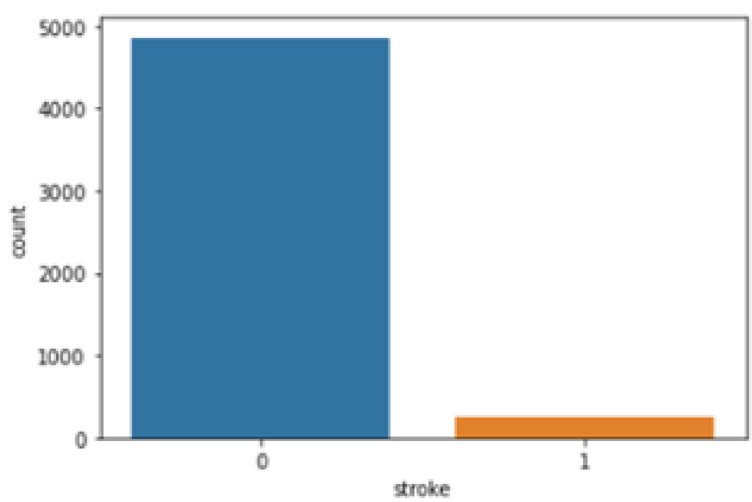

Figure 3. Count plot for stroke.

Figure 3 presents the overall view of brain stroke reported from the selected dataset. Figure 4 presents the work type that is mostly causing brain strokes among the workers.

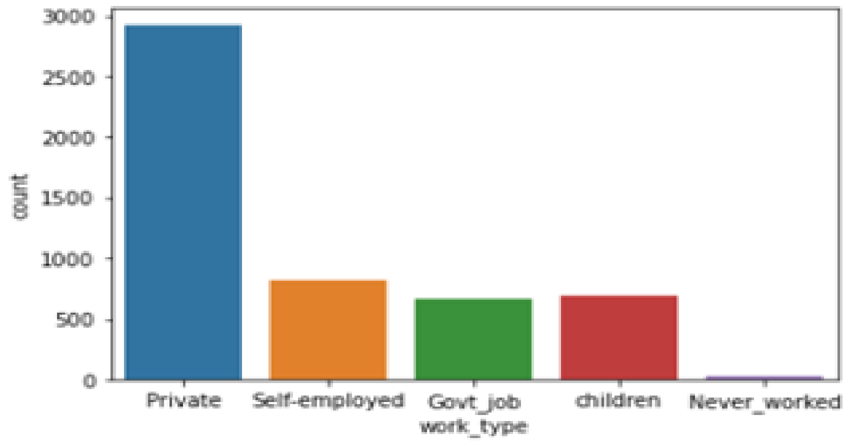

Figure 4. Work type

Figure 5 indicates the frequency of brain stroke occurrence in people living in rural areas and urban areas.

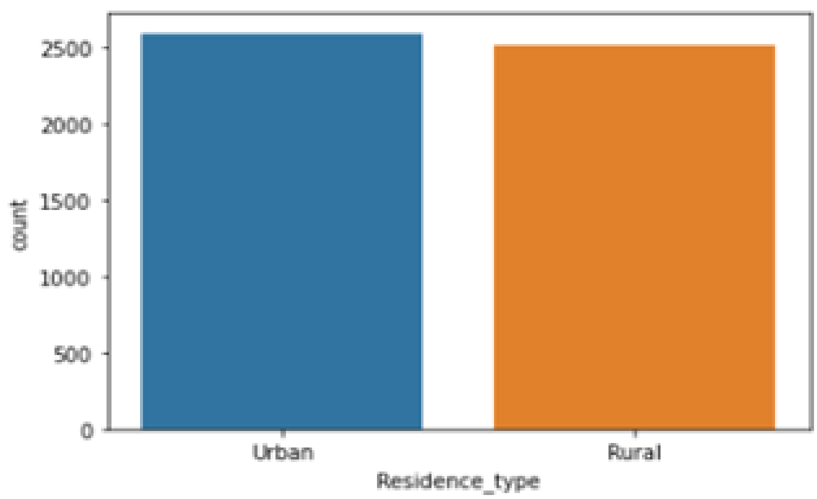

Figure 5. Residence

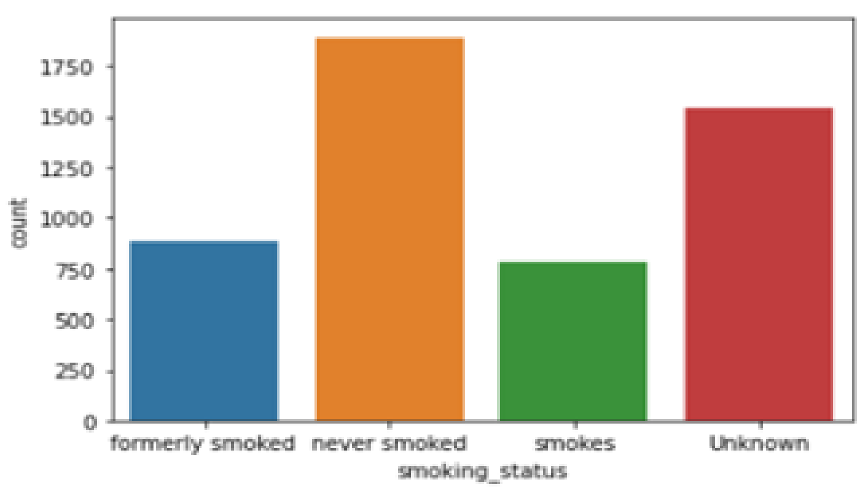

Figure 6. Smoke

Figure 7 shows the frequency of brain stroke in married and unmarried people. After performing data analysis, the model training process as presented in Figure 8 starts. Dependencies between variables are determined and presented in Figure 9.

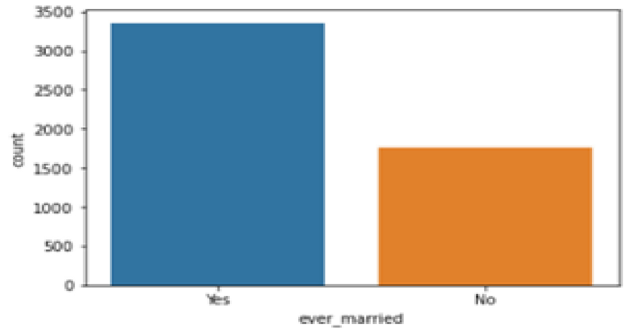

Figure 7. Married

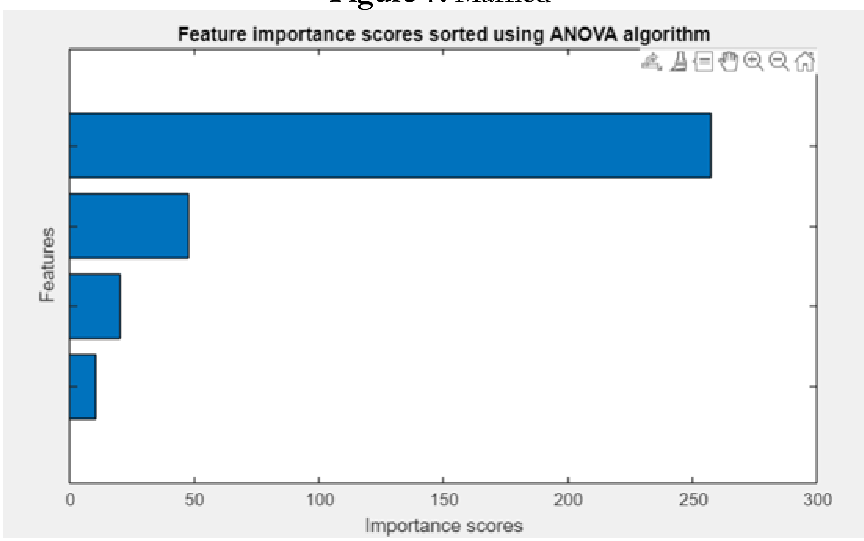

Figure 8. Training of dataset

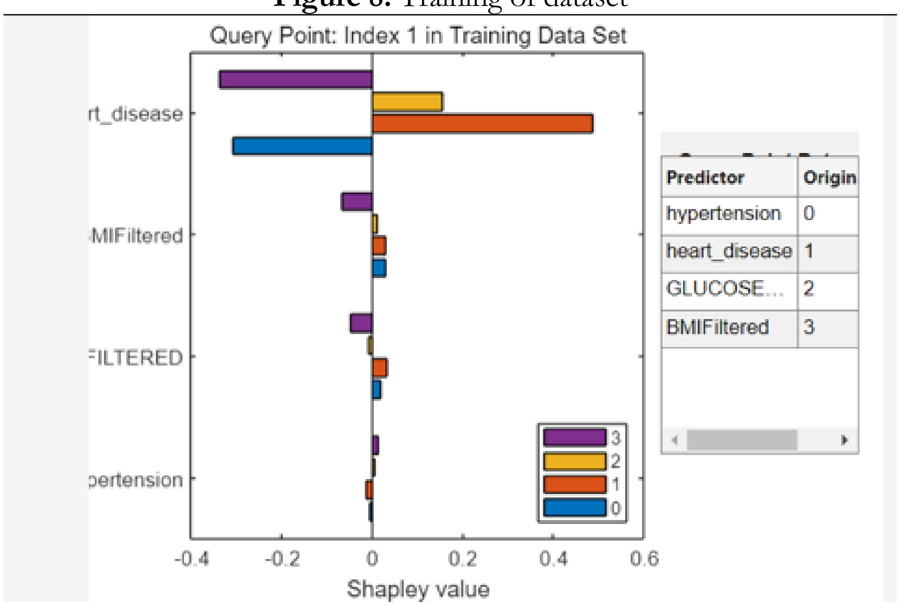

Figure 9. Feature dependency graph

Figure 10 and Figure 11 determine the overall importance of features and refer to the most important and relevant feature to be focused on for further analysis. We applied ANOVA and the Chi-Square method to segregate the most important features from the huge dataset.

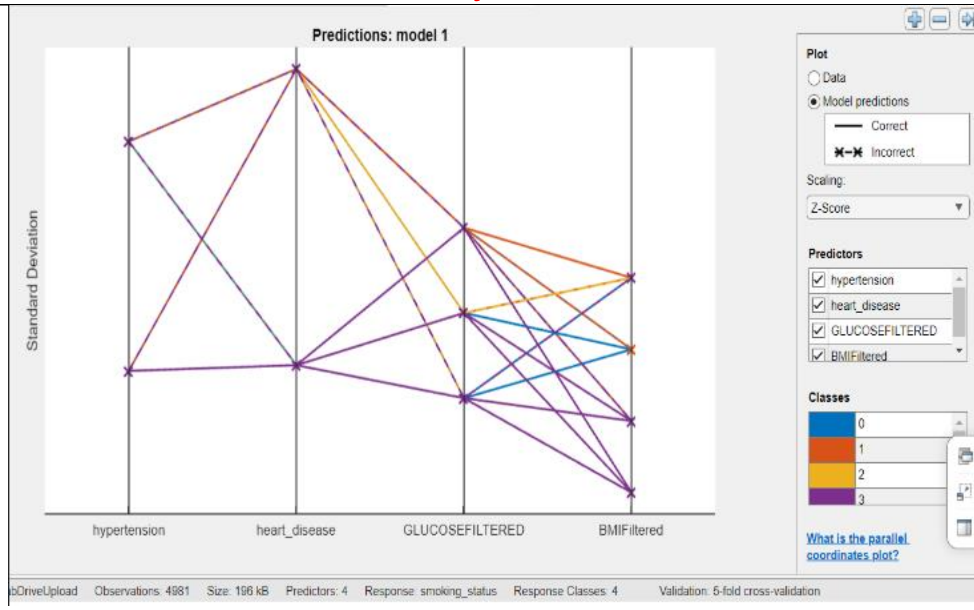

Figure10. Feature importance calculation using ANOVA

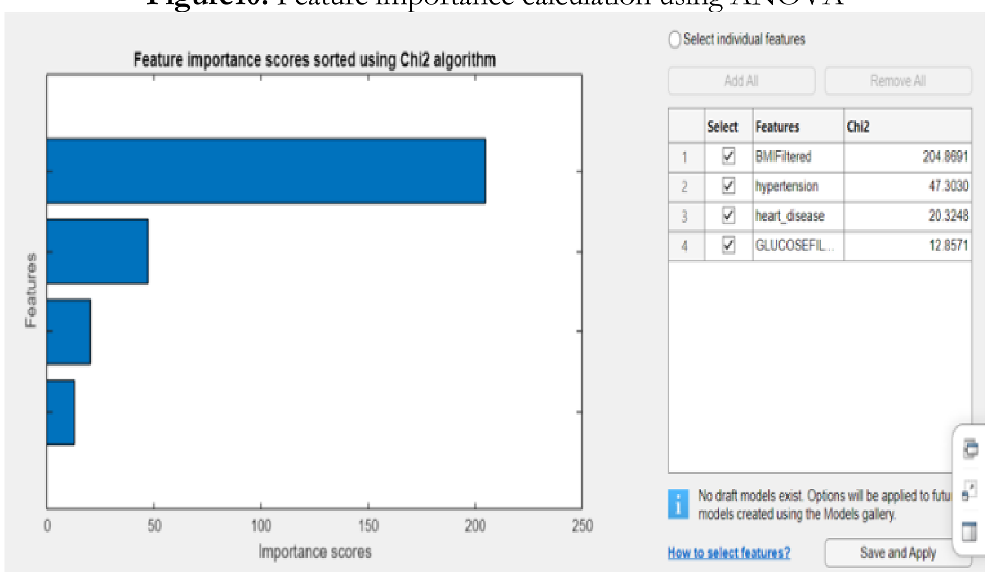

Figure 11. Feature importance calculation using Chi-Square

Correlation between variables:

Figure 12 depicts a correlation matrix, which visualizes the relationships between various variables in a dataset. The color gradient represents the strength and direction of the correlation, ranging from 1 to -1:

1 indicates a perfect positive correlation (yellow color).

-1 indicates a perfect negative correlation (purple color).

0 indicates no correlation (green color).

Details based on the correlation values are listed below:

1. Age and Ever Married: A strong positive correlation (0.68), indicated that older individuals are more likely to be married.

2. Age and Hypertension: There is a moderate positive correlation (0.28), suggesting that hypertension is more common with age.

3. BMI and Hypertension: A moderate correlation (0.33), indicated that higher BMI is associated with hypertension.

4. Heart Disease and Age: Moderate positive correlation (0.26), showed that heart disease is more prevalent in older individuals.

5. Stroke and Age: The correlation between stroke and age is notable (0.25), indicating an increased risk of stroke with age.

6. Average Glucose Level and Heart Disease: A moderate positive correlation (0.16), suggested a relationship between glucose levels and heart disease. This correlation matrix visualizes the pairwise Pearson correlation coefficients between different features in a dataset.

Key Observations:

1. **Diagonal Values**: The diagonal values are all 1 because each feature has a perfect correlation with itself.
2. **Correlation with Stroke**: The "stroke" column/row shows the correlation of each feature with the target variable, stroke. Features like age, heart disease, average glucose level, and hypertension appear to have moderate positive correlations with stroke. Features such as gender, work_type, Residence_type, and smoking_status show weak or near-zero correlations with stroke.
3. **Feature Relationships**: There are strong correlations between some features, such as age and ever_married (positive correlation): Older individuals are more likely to be married.

Heart disease and hypertension (moderate correlation). Individuals with heart disease often have hypertension. Some features like Residence_type and work_type show weak correlations with others, indicating potential redundancy.

1) Feature Selection:

Based on this matrix, feature selection decisions could involve:

- **Retain**:

Age, heart_disease, avg_glucose_level, and hypertension: These have notable correlations with stroke.

- **Consider Dropping**:

Residence_type, work_type, and potentially gender: These show weak correlations with stroke and might add noise rather than predictive power.

Smoking status: Unless domain knowledge suggests its importance, the weak correlation might justify its removal.

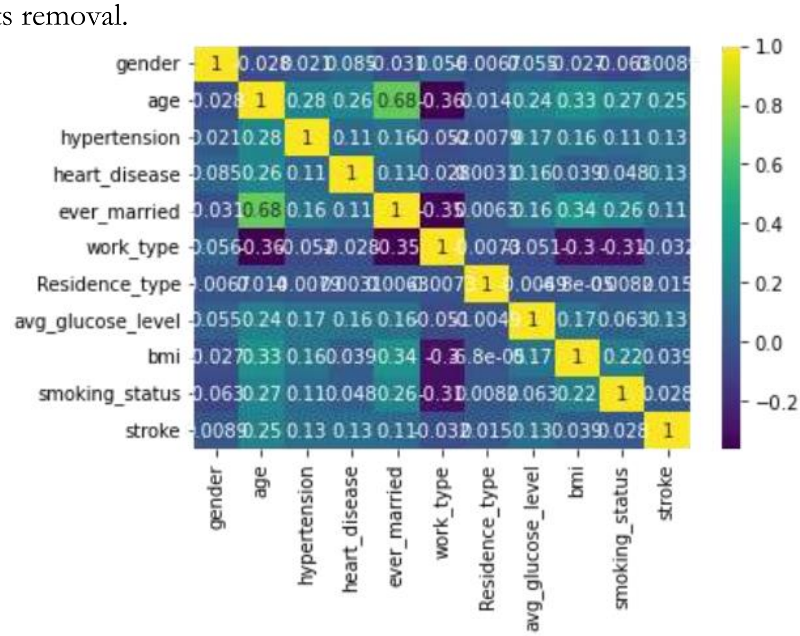

Figure 12. Correlation Matrix

Most other variables show weak or no significant correlation, as indicated by values close to zero and the corresponding green to light purple shades.

Federated Learning:

The goal of the algorithm is to securely determine the attribute values on each device without compromising its security or privacy. In this scenario, there are **N clients**, denoted by $C \subset \{C_0, C_1, \ldots, C_n\}$, the server model ($w_0$) is broadcast to all clients. In addition to distributing the server model, **three distinct learning rates** are transmitted to the edge devices. These learning rates are chosen from an array (ηm), whose values vary from [$1e^{-1}$, $1e^{5}$], allowing for flexibility in learning speed and convergence behavior across devices with varying computational power and data properties. Additionally, the sample size is randomly determined, and each edge device is initialized with the same value of w_0. Increasing the sample size for each device can further enhance training accuracy. This process is replicated for all values and monitored separately throughout the entire iteration. The complexity, size, ambiguity, and variation in the data characteristics are unique to each edge device. As a result, hyper-parameters are selected with care, as these data properties significantly impact the training process. Only the model with the lowest loss w_0min, is chosen out of the models at every individual device. Each edge device will provide the values w_{0min}, η_{min}, and $losses_{min}$. Because they can display data about specific edge devices, these statistics are crucial. On the server, the models "w_0 " learning rates "0," "1",... "n" and their associated losses are obtained. Using the model aggregation technique, the edge device models are combined to create the server model. The model weights (w_0 n) are incrementally added. Figure 13 presents the FL process.

Algorithm 1 Clustering and Initial Broadcasting

```
Create a server
For every client
while j = 3 do
    while j ≠ clients do
            k=0
            while j ≠ 3 do
                    w_j^0 on η_j
                    los = loss(w_j^0)
return train_losses, val_losses
w^0 ← Σ(w_n^0)/n
```

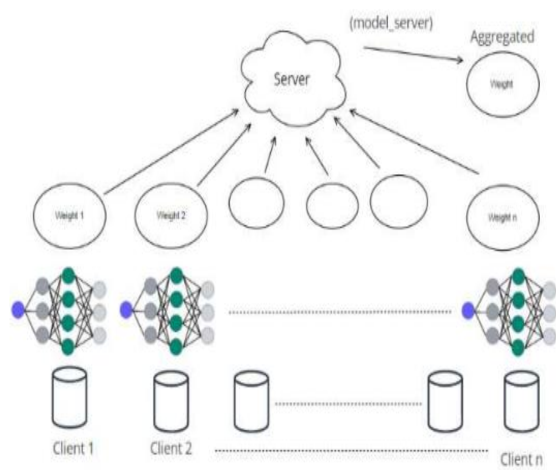

Figure 13. FL process.

Results:

The EDA provided a comprehensive numerical and visualization analysis. The next step was to design a model to predict seed types based on the given characteristics. Since the class is categorical, the Random Forest model was chosen for training due to its high accuracy in classified datasets where human involvement is not required. Thus, the 80% training and 20% testing dataset was selected to train the model. After training the model with the training dataset and testing it with the testing dataset, the accuracy was 99% on the training dataset and approximately 97% on the testing dataset.

The random forest model showed the highest accuracy with a precision of 99% recall of 94% and an f1 score of 94% outperforming the state-of-the-art models like KNN, DT model, Naïve Bayes model, SVM. The utilized dataset was imbalanced, thus smote feature engineering was used to process the data. Results are presented in Table 1.

Table 1. Comparison of Result

Model	Precision	Recall	F1-Score
DT	95	94	94
RF	95	99	97
CNN	90	80	88
Proposed Model	93	92	95

The Random Forest model was selected to train the global model to achieve the true essence of FL. After training the model to use the training dataset and then testing the model with a testing dataset, the accuracy was 99% with the training dataset and around 97% with a testing dataset. This information is shown in the accuracy curve in Figure 14.

Table 2. Validation of ML models

Model Type	Status	RMSE (Validation)	MSE (Validation)	RSquared (Validation)	MAE (Validation)
Linear Regression	Trained	1.2133	1.4721	0.067596	1.0823
Neural Network	Trained	1.2022	1.4454	0.084537	1.0692
Neural Network	Trained	1.2039	1.4493	0.082029	1.0707
Neural Network	Trained	1.2042	1.4502	0.081466	1.0712
Neural Network	Trained	1.2037	1.449	0.082234	1.0708
Neural Network	Trained	1.2032	1.4478	0.083023	1.0704
SVM	Canc...	-	-	-	-
SVM	Trained	1.2206	1.4898	0.056374	1.0527
SVM	Trained	1.2296	1.5118	0.042432	1.054
SVM	Trained	1.2458	1.552	0.017003	1.065
SVM	Trained	1.2413	1.5409	0.024044	1.0622
SVM	Trained	1.2305	1.514	0.041062	1.0519

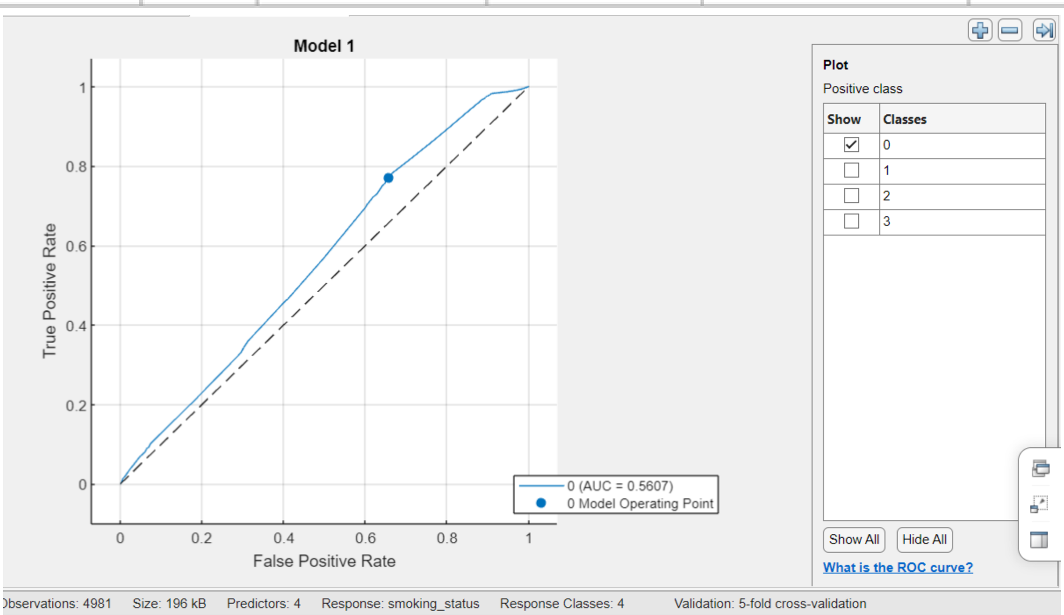

Figure 14. Accuracy Curve

Discussion:

In this research, three different models namely Linear Regression, NN, and SVM have been applied to predict the occurrence of brain stroke, which is very essential as early diagnosis can make the treatment rapid.

A. Hyper-parameter optimization:

Hyperparameter optimization is a crucial step in enhancing the performance of ML models, including those deployed in FL settings. By fine-tuning parameters such as learning rate, batch size, number of communication rounds, and model architecture, hyperparameter optimization can significantly improve the model's predictive accuracy, convergence speed, and overall robustness. In brain stroke prediction, optimal hyperparameters ensure that the model effectively captures the complex patterns in distributed datasets, while minimizing issues such as overfitting or underfitting.

B. Model Validation:

Table 2 presents a comparison of various ML models based on their validation metrics. The table includes different model types such as Linear Regression, NN, and SVM. Each model has been evaluated using several key validation performance indicators:

1. **RMSE (Root Mean Squared Error)**: This metric shows the error between predicted and actual values. Lower RMSE values indicate better model performance. The models have RMSE values ranging from **1.2022** to **1.2413**.
2. **MSE (Mean Squared Error)**: Similar to RMSE but without the square root, MSE measures the average squared difference between predicted and actual values. The values range from **1.449** to **1.514**, reflecting model performance.
3. **R-Squared**: This is a statistical measure that indicates how well the model predictions fit the actual data, with higher values (closer to 1) representing a better fit. The models' R-ssquared values range from **0.040** to **0.082**, which suggests varying levels of performance in capturing the variance in the data.
4. **MAE (Mean Absolute Error)**: MAE measures the average absolute difference between predicted and actual values. Lower values indicate more accurate predictions, with the models showing MAE values ranging from **1.0519** to **1.0713**.

The status column indicates whether the model has been successfully trained, with one instance of an SVM model marked as "Cancelled." Overall, the neural network models perform better in terms of error metrics. Figure 15 presents a 3D surface plot comparing the performance of NN, SVM, and Linear Regression models. The x-axis represents Linear Regression, the y-axis represents SVM, and the z-axis represents Neural Network. The color gradient, ranging from cyan to purple, likely indicates the performance metric (such as loss, accuracy, or error), with cyan representing lower values (better performance) and purple representing higher values (worse performance). The surface formed between the axes shows how the models behave relative to each other based on the chosen metric. The central area with the peak represents a higher error or worse performance. As you move toward the base of the surface plot, the values decrease, indicating better performance. The label "FL" at the top of the surface, suggests the outcomes of a FL model's performance across different ML algorithms. The combination of NN, SVM, and Linear Regression shows how different models perform in an FL setup. Figure 15 demonstrates how performance varies across models (SVM, Neural Network, and Linear Regression) to one another, with the color and surface height giving insight into their relative effectiveness.

C. Real World Applications

The developed stroke prediction models offer significant real-world applications:

1. **Personalized Healthcare**: Enables early identification of high-risk patients for preventive measures.
2. **Telemedicine**: Facilitates remote risk assessment, especially for underserved areas.
3. **Hospital Optimization**: Helps prioritize and allocate resources for high-risk patients.
4. **Wearable Devices**: Embeds risk monitoring in smartwatches for real-time alerts.
5. **Population Health**: Analyzes demographic trends to guide public health policies.
6. **EHR Integration**: Provides stroke risk scores directly within clinical workflows.

7. **Insurance**: Aids in risk stratification for personalized insurance plans.
8. **Research**: Identifies high-risk participants for clinical trials.

These applications highlight the potential to improve healthcare delivery, reduce stroke incidence, and enhance patient outcomes while maintaining data privacy and scalability.

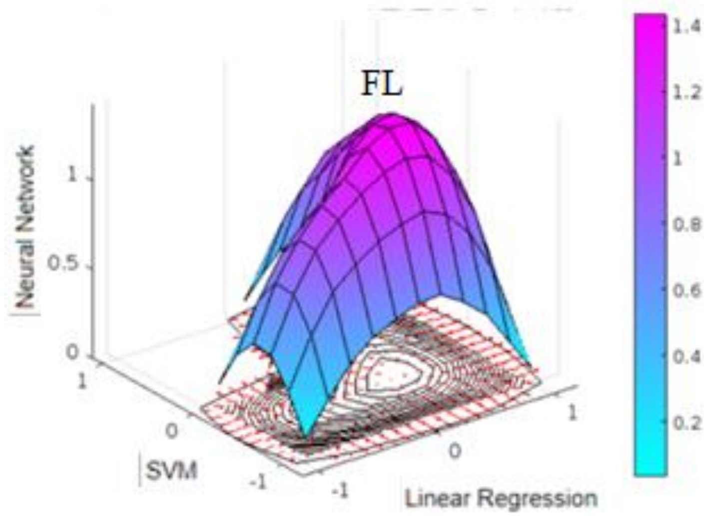

Figure 15. Performance of NN, SVN, and LR in FL setup

Conclusion:

This paper presented an ML approach to the stroke dataset. The Random Forest models showed the best accuracy with a precision of 95%, recall of 99%, and F1-score of 97%, outperforming the state-of-art models including LR, DT classifier, and K-NN. The utilized dataset is imbalanced; therefore, SMOTE feature engineering is used to process the data. In the future, we will plan to analyze the dataset using DL methods and try to enhance the accuracy.

Acknowledgments: The authors gratefully acknowledge the continuous support of Mr. M. Inayat Ali for their expertise and assistance throughout all aspects of our study and for their help in writing the manuscript.

Contributors: Rabia Tehseen (RT) led this research work and statistically analyzed the whole article. This manuscript presents the network architecture, taxonomies, and security model that are designed to measure diabetes disease effectively. Uzma Omer (UO) has supported the problem statement refinement with a state-of-the-art literature review. Moreover, RT and (Rubab Javaid) RJ have contributed to the selection of frequently used open-source software in the field of diabetes. The manuscript draft was proofread and revised three times before submitting to your prestigious journal by RT, UO, and Madiha Yousaf (MY). RT and Maham Mehr (MM) have s contributed to designing the methodology and proofreading of the manuscript. In the end, Nosheen Sabahat (NS) read and finalized the manuscript.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest to the research, authorship, and/or publication of this article.

Ethical approval: Ethical approval was not required as this research was a survey of the literature. **Funding:** The author(s) received no financial support for the research, authorship, and/or publication of this article.

Guarantor: RT

Reference:

[1] S. Rahman, M. Hasan, and A. K. Sarkar, "Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques," *Eur. J. Electr. Eng. Comput. Sci.*, vol. 7, no. 1, pp. 23–30, Jan. 2023, doi: 10.24018/EJECE.2023.7.1.483.

[2] O. Shobayo, O. Zachariah, M. O. Odusami, and B. Ogunleye, "Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm," *Anal. 2023, Vol. 2, Pages 604-617*, vol. 2, no. 3, pp. 604–617, Aug. 2023, doi: 10.3390/ANALYTICS2030034.
[3] P. Gahiwad, N. Deshmane, S. Karnakar, S. Mali, and R. Pise, "Brain Stroke Detection Using CNN Algorithm," *2023 IEEE 8th Int. Conf. Converg. Technol. I2CT 2023*, 2023, doi: 10.1109/I2CT57861.2023.10126125.
[4] M. Singhal, M. Pruthi, N. Singhal, P. Nagrath, R. Jain, and A. Kumar, "Comparison of machine-learning and deep-learning algorithms for stroke prediction," *Edge-AI Healthc. Trends Futur. Perspect.*, pp. 115–128, Jan. 2023, doi: 10.1201/9781003244592-9/COMPARISON-MACHINE-LEARNING-DEEP-LEARNING-ALGORITHMS-STROKE-PREDICTION-MAYANK-SINGHAL-MANAN-PRUTHI-NIVEDITA-SINGHAL-PREETI-NAGRATH-RACHNA-JAIN-ASHISH-KUMAR.
[5] S. Gupta and S. Raheja, "Stroke Prediction using Machine Learning Methods," *Proc. Conflu. 2022 - 12th Int. Conf. Cloud Comput. Data Sci. Eng.*, pp. 553–558, 2022, doi: 10.1109/CONFLUENCE52989.2022.9734197.
[6] T. N. Deepthi, S. Sharmila, M. Swarna, M. Gouthami, and C. L. Akshaya, "Prediction of Brain Stroke in Human Beings using Machine Learning," *Proc. 2023 2nd Int. Conf. Electron. Renew. Syst. ICEARS 2023*, pp. 1314–1320, 2023, doi: 10.1109/ICEARS56392.2023.10085128.
[7] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of brain stroke severity using machine learning," *Rev. d'Intelligence Artif.*, vol. 34, no. 6, pp. 753–761, Dec. 2020, doi: 10.18280/RIA.340609.
[8] G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 539–545, 2021, doi: 10.14569/IJACSA.2021.0120662.
[9] M. J. Uddin Chowdhury, A. Hussan, D. A. Islam Hridoy, and A. S. Sikder, "Incorporating an Integrated Software System for Stroke Prediction using Machine Learning Algorithms and Artificial Neural Network," *2023 IEEE 13th Annu. Comput. Commun. Work. Conf. CCWC 2023*, pp. 222–228, 2023, doi: 10.1109/CCWC57344.2023.10099245.
[10] "Performance Analysis of Machine Learning Approaches for Stroke Prediction in Healthcare | IEEE Conference Publication | IEEE Xplore." Accessed: Dec. 13, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10112284
[11] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," *Healthc. Anal.*, vol. 2, p. 100032, Nov. 2022, doi: 10.1016/J.HEALTH.2022.100032.
[12] S. Mushtaq and K. S. Saini, "A Review on Predicting Brain Stroke using Machine Learning," *Int. Conf. Comput. Sustain. Glob. Dev.*, 2023.
[13] N. Victor, S. Bhattacharya, P. K. R. Maddikunta, F. M. Alotaibi, T. R. Gadekallu, and R. H. Jhaveri, "FL-PSO: A Federated Learning approach with Particle Swarm Optimization for Brain Stroke Prediction," *Proc. - 23rd IEEE/ACM Int. Symp. Clust. Cloud Internet Comput. Work. CCGridW 2023*, pp. 33–38, 2023, doi: 10.1109/CCGRIDW59191.2023.00020.
[14] R. Kumari and H. Garg, "Interpretation and Analysis of Machine Learning Models for Brain Stroke Prediction," *2023 6th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2023*, 2023, doi: 10.1109/ISCON57294.2023.10112188.
[15] A. S. Chandrabhatla, E. A. Kuo, J. D. Sokolowski, R. T. Kellogg, M. Park, and P. Mastorakos, "Artificial Intelligence and Machine Learning in the Diagnosis and Management of Stroke: A Narrative Review of United States Food and Drug Administration-Approved Technologies," *J. Clin. Med.*, vol. 12, no. 11, Jun. 2023, doi: 10.3390/JCM12113755.
[16] S. Zhang *et al.*, "Research Progress of Deep Learning in the Diagnosis and Prevention of Stroke," *Biomed Res. Int.*, vol. 2021, p. 5213550, 2021, doi: 10.1155/2021/5213550.
[17] B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas, and U. Sara, "A Machine Learning Approach to Detect the Brain Stroke Disease," *Proc. - 4th Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2022*, pp. 897–901, 2022, doi: 10.1109/ICSSIT53264.2022.9716345.

[18] I. R. Chavva *et al.*, "Deep Learning Applications for Acute Stroke Management," *Ann. Neurol.*, vol. 92, no. 4, pp. 574–587, Oct. 2022, doi: 10.1002/ANA.26435.
[19] Ferdib-Al-Islam and M. Ghosh, "An Enhanced Stroke Prediction Scheme Using Smote and Machine Learning Techniques," *2021 12th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2021*, 2021, doi: 10.1109/ICCCNT51525.2021.9579648.
[20] M. Kaur, S. R. Sakhare, K. Wanjale, and F. Akter, "[Retracted] Early Stroke Prediction Methods for Prevention of Strokes," *Behav. Neurol.*, vol. 2022, no. 1, p. 7725597, Jan. 2022, doi: 10.1155/2022/7725597.
[21] P. Premisha, S. Prasanth, M. Kanagarathnam, and K. Banujan, "An Ensemble Machine Learning Approach for Stroke Prediction," *Proc. - Int. Res. Conf. Smart Comput. Syst. Eng. SCSE 2022*, pp. 165–170, 2022, doi: 10.1109/SCSE56529.2022.9905215.
[22] "Machine Learning for Brain Stroke: A Review - ScienceDirect." Accessed: Dec. 13, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1052305720305802
[23] B. R. Gaidhani, R. Rajamenakshi, and S. Sonavane, "Brain stroke detection using convolutional neural network and deep learning models," *2019 2nd Int. Conf. Intell. Commun. Comput. Tech. ICCT 2019*, pp. 242–249, Sep. 2019, doi: 10.1109/ICCT46177.2019.8969052.
[24] H. Zhang, "Stroke Prediction Based on Support Vector Machine," *Highlights Sci. Eng. Technol.*, vol. 31, pp. 53–59, Feb. 2023, doi: 10.54097/HSET.V31I.4812.
[25] A. P. . Rohit, M. U. Chowdary, G. B. . Ashish, V. Anitha, and S. Sana, "ML APPROACH FOR BRAIN STROKE PREDICTION USING IST DATABASE," *Int. J. Eng. Appl. Sci. Technol.*, vol. 7, no. 10, pp. 72–75, Feb. 2023, doi: 10.33564/IJEAST.2023.V07I10.008.
[26] A. Subudhi, P. Dash, M. Mohapatra, R. S. Tan, U. R. Acharya, and S. Sabut, "Application of Machine Learning Techniques for Characterization of Ischemic Stroke with MRI Images: A Review," *Diagnostics 2022, Vol. 12, Page 2535*, vol. 12, no. 10, p. 2535, Oct. 2022, doi: 10.3390/DIAGNOSTICS12102535.
[27] D. V. Desai, T. Jain, and P. Tiwari, "Supervised Machine Learning Approaches for Brain Stroke Detection," *11th Int. Conf. Internet Everything, Microw. Eng. Commun. Networks, IEMECON 2023*, 2023, doi: 10.1109/IEMECON56962.2023.10092374.
[28] S. Kanaga Suba Raja, B. Chandra, K. Kausalya, R. M. Ciddarth, and V. Gokul Ranjith, "Prognosis of Stroke using Machine Learning Algorithms," *Proc. - 7th Int. Conf. Comput. Methodol. Commun. ICCMC 2023*, 2023, doi: 10.1109/ICCMC56507.2023.10084158.
[29] S. Mushtaq, K. S. Saini, and S. Bashir, "Machine Learmusht for Brain Stroke Prediction," *2023 Int. Conf. Disruptive Technol. ICDT 2023*, pp. 401–408, 2023, doi: 10.1109/ICDT57929.2023.10151148.
[30] G. Yang, M. Xu, W. Chen, X. Qiao, H. Shi, and Y. Hu, "A brain CT-based approach for predicting and analyzing stroke-associated pneumonia from intracerebral hemorrhage," *Front. Neurol.*, vol. 14, 2023, doi: 10.3389/FNEUR.2023.1139048.
[31] D. V. V. V. C. Mouli, "Artificial Intelligence Enabled Framework for Automatic Brain Stroke Detection," *Proc. 10th Int. Conf. Signal Process. Integr. Networks, SPIN 2023*, pp. 659–664, 2023, doi: 10.1109/SPIN57001.2023.10116629.
[32] S. Ruksakulpiwat *et al.*, "Machine learning-based patient classification system for adults with stroke: A systematic review," *Chronic Illn.*, vol. 19, no. 1, pp. 26–39, Mar. 2023, doi: 10.1177/17423953211067435.
[33] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthc. Anal.*, vol. 2, p. 100116, Nov. 2022, doi: 10.1016/J.HEALTH.2022.100116.
[34] H. Al-Zubaidi, M. Dweik, and A. Al-Mousa, "Stroke Prediction Using Machine Learning Classification Methods," *Proc. - 2022 23rd Int. Arab Conf. Inf. Technol. ACIT 2022*, vol. 2022-January, 2022, doi: 10.1109/ACIT57182.2022.10022050.
[35] A. Diker, A. Elen, and A. Subasi, "Brain stroke detection from computed tomography images using deep learning algorithms," *Appl. Artif. Intell. Med. Imaging*, pp. 207–222, Jan. 2023,

doi: 10.1016/B978-0-443-18450-5.00013-X.
[36] K. Wang *et al.*, "A machine learning model for visualization and dynamic clinical prediction of stroke recurrence in acute ischemic stroke patients: A real-world retrospective study," *Front. Neurosci.*, vol. 17, 2023, doi: 10.3389/FNINS.2023.1130831.
[37] "(PDF) An Explainable Machine Learning Framework for Early Detection of Stroke." Accessed: Dec. 13, 2024. [Online]. Available: https://www.researchgate.net/publication/373424278_An_Explainable_Machine_Learning_Framework_for_Early_Detection_of_Stroke
[38] S. C. Manjunath, "Predicting Stroke at Adulthood Using Machine Learning Techniques," Jan. 2023.
[39] S. O. Abdulsalam, M. O. Arowolo, and O. Ruth, "Stroke Disease Prediction Model Using ANOVA with Classification Algorithms," pp. 117–134, 2023, doi: 10.1007/978-981-99-0369-6_8.
[40] V. Kumar, V. S., Thanmai, B., ManognaVyshnavi, A., Sujitha, I., & Jhansi, "Prediction of Stroke Risk Using a Hybrid Deep Transfer Learning Framework," *J. Eng. Sci.*, vol. 14, no. 04, 2023.
[41] A. Faskhutdinova, D. Grigorieva, B. Garafutdinov, and V. Mokshin, "Investigation of Machine Learning Methods for Stroke Prediction," *Proc. - 9th IEEE Int. Conf. Inf. Technol. Nanotechnology, ITNT 2023*, 2023, doi: 10.1109/ITNT57377.2023.10139121.
[42] A. K. Uttam, "Analysis of Uneven Stroke Prediction Dataset using Machine Learning," *Proc. - 2022 6th Int. Conf. Intell. Comput. Control Syst. ICICCS 2022*, pp. 1209–1213, 2022, doi: 10.1109/ICICCS53718.2022.9788309.
[43] C. Sharma, S. Sharma, M. Kumar, and A. Sodhi, "Early Stroke Prediction Using Machine Learning," *2022 Int. Conf. Decis. Aid Sci. Appl. DASA 2022*, pp. 890–894, 2022, doi: 10.1109/DASA54658.2022.9765307.
[44] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," *Sensors (Basel).*, vol. 22, no. 13, p. 4670, Jul. 2022, doi: 10.3390/S22134670.
[45] H. M. Xia and R. Ramezani, "Analyzing the Performance of TabTransformer in Brain Stroke Prediction," *J. Student Res.*, vol. 12, no. 1, Feb. 2023, doi: 10.47611/jsrhs.v12i1.3935.
[46] S. Paliwal, S. Parveen, M. A. Alam, and J. Ahmed, "Improving Brain Stroke Prediction through Oversampling Techniques: A Comparative Evaluation of Machine Learning Algorithms," Jun. 2023, doi: 10.20944/PREPRINTS202306.1444.V1.
[47] P. Savar Mattas, "Brain Stroke Prediction Using Machine Learning," *Int. J. Res. Publ. Rev.*, vol. 03, no. 12, pp. 711–722, 2022, doi: 10.55248/GENGPI.2022.31211.
[48] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artif. Intell. Med.*, vol. 101, p. 101723, Nov. 2019, doi: 10.1016/J.ARTMED.2019.101723.
[49] "Stroke Prediction Dataset." Accessed: Dec. 13, 2024. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset