Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

Optimizing Stroke Prediction: High-Accuracy ML Models and Explainable AI Techniques

[1]P.Subhan basha, [2] Vallepu Eswaraiah
[1]HOD,Associate professor, [2]M.Tech Student
[1,2]Department of CSE
[1,2]Sri Sai Institute Of Technology And Science, Masapeta, Rayachoty (M), Annamayya Dt, AP.
subhan.mahammad@gmail.com, eswar19970807@gmail.com

ABSTRACT

Stroke is a global health crisis requiring early intervention to mitigate severe outcomes. To address this, researchers are developing automated prediction algorithms to identify at-risk individuals, which is increasingly important with an aging population. This study compares various machine learning classifiers, assessing their generalization and accuracy. We also use explainable techniques like SHAP and LIME to interpret model decisions. Our results indicate that complex models outperform simpler ones, with the best model achieving nearly 91% accuracy. Incorporating CATBOOST, a group-based classifier, further boosts prediction accuracy to 95%. This comprehensive approach offers valuable insights into model decisions, enhancing stroke care and treatment protocols.

Index Terms: Machine Learning, AI

INTRODUCTION:

Stroke incidence is rising globally, making it a leading cause of death and disability. Early intervention is vital to prevent long-term consequences, yet traditional stroke risk prediction methods are often slow and error-prone. Recently, machine learning algorithms have shown significant promise in accurately predicting stroke risk from clinical data. These models enable clinicians to identify high-risk patients early, potentially reducing complications and improving outcomes. Additionally, the need for transparency in machine learning models is growing in healthcare. Interpretable models provide insights into factors contributing to stroke risk, aiding treatment decisions and improving patient care.

LITERATURE SURVEY

L. Gibson *et al*

This meta-analysis assesses the proportion of confirmed stroke cases among patients with suspected stroke across various healthcare settings. Among 8,839 patients from 29 studies, approximately 74% received a stroke diagnosis. However, significant heterogeneity exists in this estimate. The study also identifies common non-stroke diagnoses like seizures and syncope, crucial for accurate differential diagnosis in suspected stroke cases.

2.2 N. M. Murray *et al*

This systematic review evaluates the role of artificial intelligence (AI) in identifying and triaging acute large vessel occlusion (LVO) strokes. Using machine learning (ML) methods like random forest learning (RFL)

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

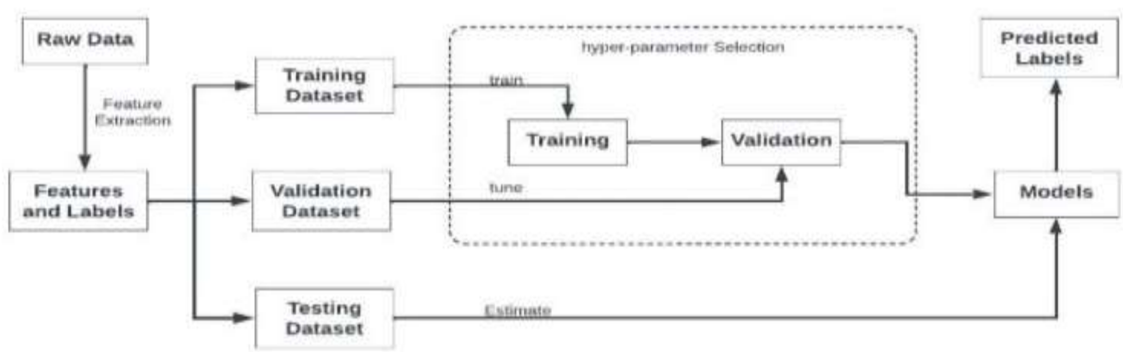
and convolutional neural networks (CNNs), AI enhances LVO detection. While CNNs exhibit higher sensitivity, standardization of AI algorithm metrics remains a challenge. Notable AI software platforms include Brainomix, iSchemaView, and Viz.ai, revolutionizing stroke care.

PROBLEM STATEMENT:

Stroke, a severe condition caused by disrupted blood flow to the brain, is a major health threat. Timely and precise detection can save lives and prevent strokes, but current methods are resource-intensive and slow. To address these challenges, machine learning algorithms have been developed, offering high accuracy in medical predictions. However, these methods often suffer from issues like data leakage and poor handling of missing or categorical data. Moreover, they lack explainability, failing to highlight which features—such as smoking, age, or BMI—are most critical in stroke detection. Incorporating Explainable AI (XAI) could provide valuable insights, allowing doctors to focus on key risk factors for quicker intervention.

PROPOSED METHOD:

The author of this paper applies several preprocessing techniques, including handling missing values, addressing class imbalance with SMOTE, and selecting relevant features using the CHI2 algorithm. These processed features are then used to train six different algorithms: Random Forest, KNN, SVM, Logistic Regression, XGBoost, and Naive Bayes. Among these, Random Forest delivers the highest accuracy. The performance of each algorithm is assessed using metrics such as accuracy, precision, recall, and F1-score. To facilitate understanding, various graphs are used to visualize stroke patient data. The top-performing algorithm is further analyzed using SHAPLEY Explainable AI (XAI) to highlight the key features influencing stroke prediction.



ARCHITECTURE

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

```
id,gender,age,hypertension,heart_disease,ever_married,work_type,Residence_type,avg_glucose_level,bmi,smoking_status,stroke
9046,Male,67,0,1,Yes,Private,Urban,228.69,36.6,formerly smoked,1
51676,Female,61,0,0,Yes,Self-employed,Rural,202.21,N/A,never smoked,1
31112,Male,80,0,1,Yes,Private,Rural,105.92,32.5,never smoked,1
60182,Female,49,0,0,Yes,Private,Urban,171.23,34.4,smokes,1
1665,Female,79,1,0,Yes,Self-employed,Rural,174.12,24,never smoked,1
56669,Male,81,0,0,Yes,Private,Urban,186.21,29,formerly smoked,1
53882,Male,74,1,1,Yes,Private,Rural,70.09,27.4,never smoked,1
10434,Female,69,0,0,No,Private,Urban,94.39,22.8,never smoked,1
27419,Female,59,0,0,Yes,Private,Rural,76.15,N/A,Unknown,1
60491,Female,78,0,0,Yes,Private,Urban,58.57,24.2,Unknown,1
12109,Female,81,1,0,Yes,Private,Rural,80.43,29.7,never smoked,1
12095,Female,61,0,1,Yes,Govt_job,Rural,120.46,36.8,smokes,1
12175,Female,54,0,0,Yes,Private,Urban,104.51,27.3,smokes,1
8213,Male,78,0,1,Yes,Private,Urban,219.84,N/A,Unknown,1
5317,Female,79,0,1,Yes,Private,Urban,214.09,28.2,never smoked,1
58202,Female,50,1,0,Yes,Self-employed,Rural,167.41,30.9,never smoked,1
56112,Male,64,0,1,Yes,Private,Urban,191.61,37.5,smokes,1
34120,Male,75,1,0,Yes,Private,Urban,221.29,25.8,smokes,1
27458,Female,60,0,0,No,Private,Urban,89.22,37.8,never smoked,1
25226,Male,57,0,1,No,Govt_job,Urban,217.08,N/A,Unknown,1
70630,Female,71,0,0,Yes,Govt_job,Rural,193.94,22.4,smokes,1
13861,Female,52,1,0,Yes,Self-employed,Urban,233.29,48.9,never smoked,1
68794,Female,79,0,0,Yes,Self-employed,Urban,228.7,26.6,never smoked,1
64778,Male,82,0,1,Yes,Private,Rural,208.3,32.5,Unknown,1
4219,Male,71,0,0,Yes,Private,Urban,102.87,27.2,formerly smoked,1
70822,Male,80,0,0,Yes,Self-employed,Rural,104.12,23.5,never smoked,1
38047,Female,65,0,0,Yes,Private,Rural,100.98,28.2,formerly smoked,1
61843,Male,58,0,0,Yes,Private,Rural,189.84,N/A,Unknown,1
```

STROKE PREDICTION DATASET

STROKE dataset from KAGGLE

METHODOLOGY:

Reading and Displaying Dataset

Load the dataset into a Pandas DataFrame and display its first few rows to understand its structure. Address any missing values and apply label encoding to convert categorical variables into numerical format.

Exploratory Data Analysis (EDA)

Perform exploratory data analysis to understand the distribution of labels and identify any class imbalance.

1. **Distribution of Labels:** Plot the distribution of 'Normal' and 'Stroke' labels to visualize class imbalance.

2. **Class Imbalance Visualization:** Use both bar and pie charts to illustrate the imbalance between the classes.

Cluster Features Correlation Graph

Create a correlation matrix to visualize the relationships between features and identify highly correlated features.

Gender and Age Relationship

Generate a graph to show how gender relates to stroke occurrences across different age groups.

Age-Based Stroke Counts

Use a stacked bar graph to display stroke counts by age, differentiating between genders.

Gender and BMI on Stroke Patients

Visualize the relationship between gender, BMI, and stroke occurrence.

Hypertension and Heart Disease Counts

Create separate graphs to display the number of stroke patients with hypertension and heart disease.

Average Glucose Level by Gender

Generate a graph illustrating the average glucose level by gender for stroke patients.

Smoking Status and Residence Type Visualization

Display the number of stroke patients based on smoking status and residence type.

Converting Categorical Data and Normalizing

Convert all remaining categorical data to numeric format and normalize the features to standardize the dataset.

Handling Class Imbalance using SMOTE

Apply the Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes in the dataset.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

Feature Selection using CHI2

Perform feature selection using the CHI2 test to determine the most significant features and split the dataset into training and testing sets.

Model Training and Evaluation

Train various machine learning algorithms and evaluate their performance using metrics such as accuracy, precision, recall, and confusion matrix.

Choosing the Best Model

Identify the best-performing model based on evaluation metrics and visualize the performance.

SHAP Explanation of Features

Use SHAP to explain which features contribute most to the model's predictions.

Comparison of Algorithm Performance

Present a summary of all algorithms' performance in a tabular format for easy comparison.

Testing with CATBOOST Algorithm

Read the test data, preprocess it similarly to the training data, and use the CATBOOST algorithm to make predictions.

Extension: CATBOOST Classifier

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

The CATBOOST classifier enhances prediction accuracy by employing a forest of weak classifiers. Each classifier is trained, and the best one is selected based on voting. This ensemble approach improves overall prediction accuracy and robustness.

This workflow involves detailed steps from importing necessary packages and reading the dataset to training multiple models and evaluating their performance. By following these procedures, you can ensure a comprehensive analysis and effective application of machine learning techniques.

EVALUATION:

Precision:

Formula: $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$

Code: precision = precision_score(testY, predict, average='macro') * 100

Recall (Sensitivity):

Formula: $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$

Code: recall = recall_score(testY, predict, average='macro') * 100

F1 Score:

Formula: $F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Code: f1 = f1_score(testY, predict, average='macro') * 100

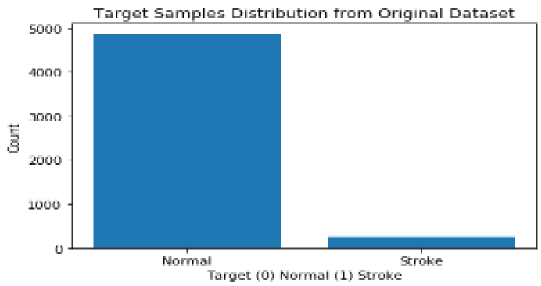
Accuracy:

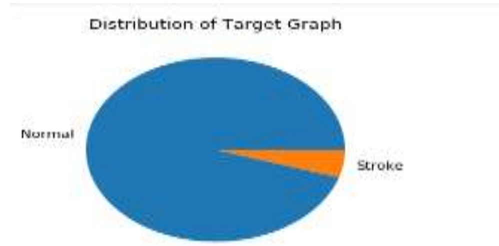
Formula: $\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$

Code: accuracy = accuracy_score(testY, predict) * 100

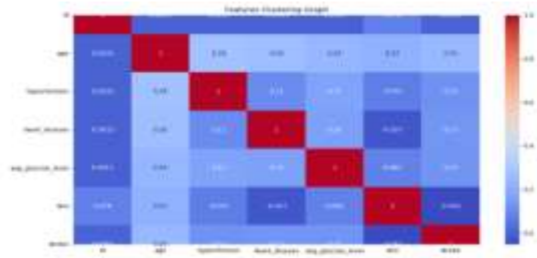
RESULTS:

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

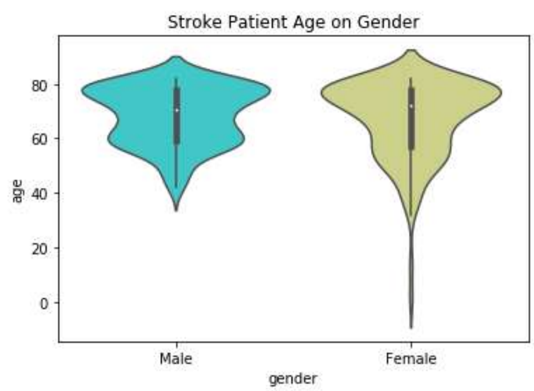




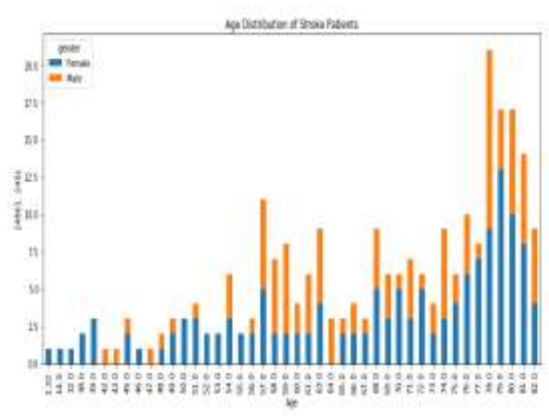
Graph for Normal and Stroke labels



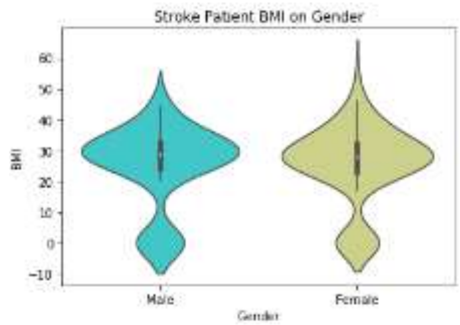
Displaying cluster features correlation graph and all values are not highly correlates. High correlated means features will have score more than 90%



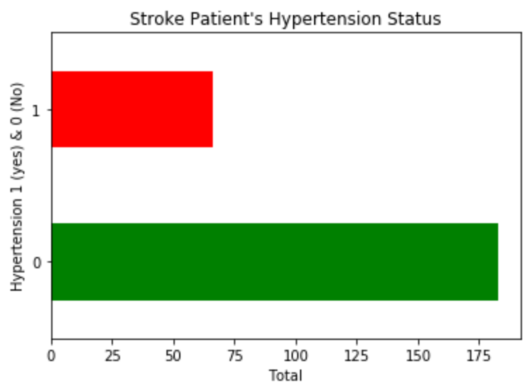
Gender with strokes on different age where x-axis represents Gender and y-axis represent age



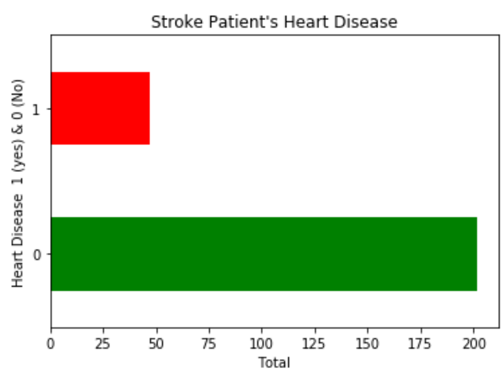
x-axis represents Age and y-axis represents stroke count where blue stack part is for Female and orange for Male



Displaying gender and BMI on stroke patients

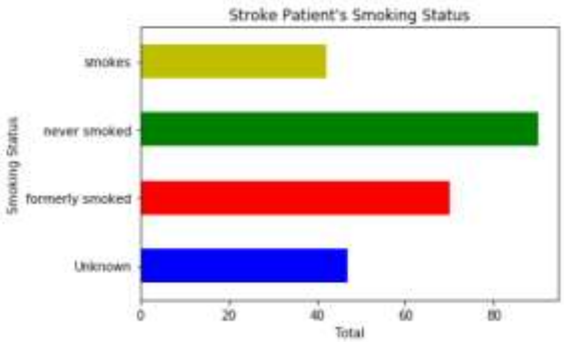


Stroke patients suffering from hypertension

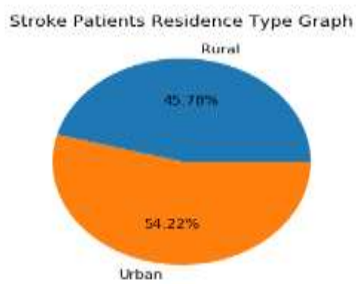


Stroke patients suffering from heart disease

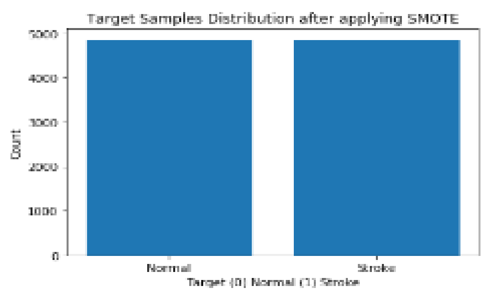
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024



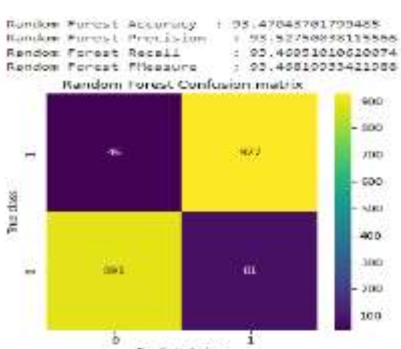
Stroke patients with smoke status



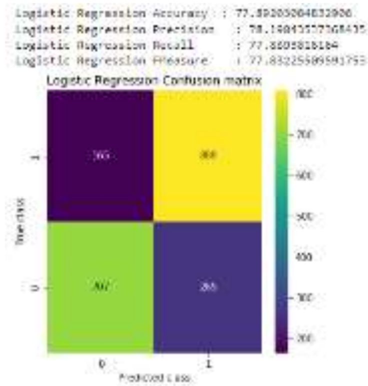
Residence type of stroke patients



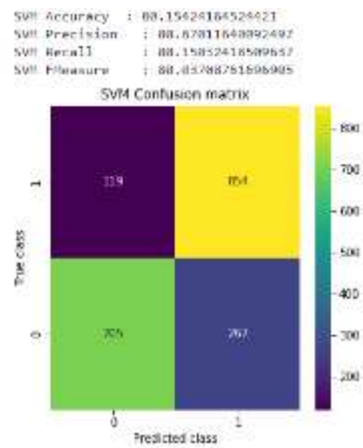
SMOTE we can see both classes has equal number of records



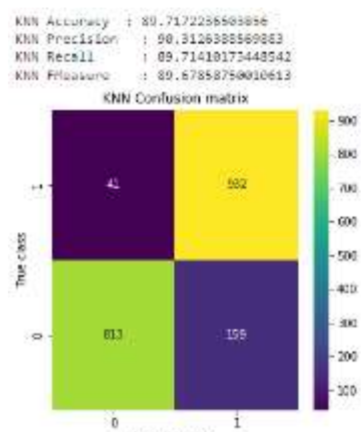
Training Random Forest algorithm got 94%



Training Logistic Regression got 78% accuracy

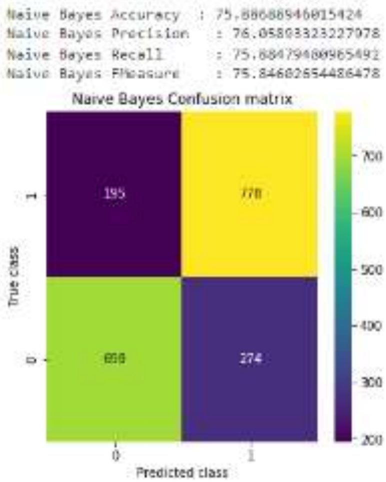


Training SVM got 80% accuracy

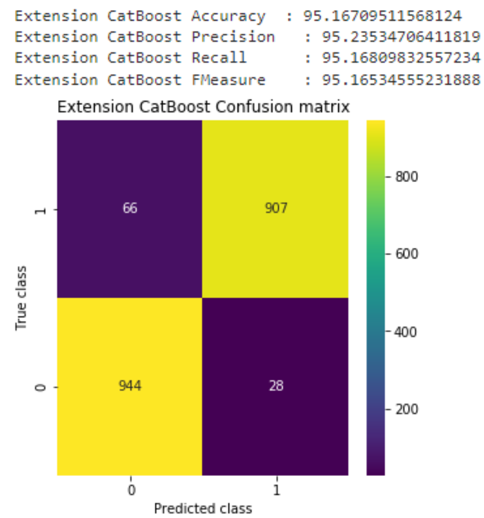


Training KNN got 92% accuracy

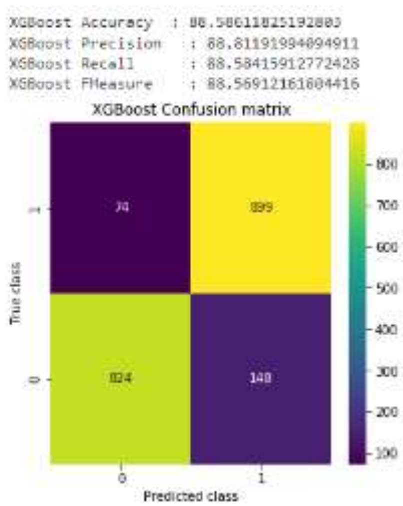
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024



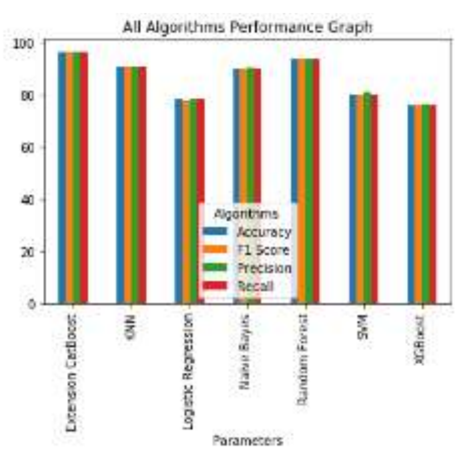
Training Naive Bayes got 77% accuracy



Training CATBOOST got 95% accuracy



Training XGBOOST got 89% accuracy



In all algorithms CATBOOST got high accuracy

Prediction:

```
Test Data = [17739 'Male' 57 0 0 'Yes' 'Private' 'Rural' 84.96 36.7 'Unknown'] Predicted As ====> Normal

Test Data = [12095 'Female' 61 0 1 'Yes' 'Govt_job' 'Rural' 120.46 36.8 'smokes'] Predicted As ====> Stroke

Test Data = [12175 'Female' 54 0 0 'Yes' 'Private' 'Urban' 104.51 27.3 'smokes'] Predicted As ====> Stroke

Test Data = [8213 'Male' 78 0 1 'Yes' 'Private' 'Urban' 219.84 0.0 'Unknown'] Predicted As ====> Stroke

Test Data = [27419 'Female' 59 0 0 'Yes' 'Private' 'Rural' 76.15 0.0 'Unknown'] Predicted As ====> Normal

Test Data = [60491 'Female' 78 0 0 'Yes' 'Private' 'Urban' 58.57 24.2 'Unknown'] Predicted As ====> Normal

Test Data = [12109 'Female' 81 1 0 'Yes' 'Private' 'Rural' 80.43 29.7 'never smoked'] Predicted As ====> Stroke

Test Data = [5317 'Female' 79 0 1 'Yes' 'Private' 'Urban' 214.09 28.2 'never smoked'] Predicted As ====> Stroke

Test Data = [58202 'Female' 50 1 0 'Yes' 'Self-employed' 'Rural' 167.41 30.9
 'never smoked'] Predicted As ====> Stroke
```

Predicted data as 'Normal or Stroke'

CONCLUSION

This project created an automated stroke prediction system integrated with a web application for early intervention. It utilized various preprocessing methods, including missing value imputation, data balancing with SMOTE, and feature selection using the CHI2 algorithm. Six machine learning algorithms were tested, with Random Forest showing the highest accuracy. To further boost performance, CATBOOST was introduced, achieving 95% accuracy. Explainable AI techniques like SHAP highlighted critical predictive features such as smoking, age, and BMI, offering a transparent model that helps doctors focus on key factors for faster stroke recovery and improved care.

REFERENCES:

[1] Learn About Stroke. Accessed: May 25, 2022. [Online]. Available: https://www.world-stroke.org/world-stroke-day-campaign/why-strokematters/learn-about-stroke

[2] T. Elloker and A. J. Rhoda, ''The relationship between social support and participation in stroke: A systematic review,'' Afr. J. Disability, vol. 7, pp. 1–9, Oct. 2018.

[3] M. Katan and A. Luft, ''Global burden of stroke,'' Seminar Neurol., vol. 38, no. 2, pp. 208–211, Apr. 2018.

[4] A. Bustamante, A. Penalba, C. Orset, L. Azurmendi, V. Llombart, A. Simats, E. Pecharroman, O. Ventura, M. Ribó, D. Vivien, J. C. Sanchez, and J. Montaner, ''Blood biomarkers to differentiate ischemic and hemorrhagic strokes,'' Neurology, vol. 96, no. 15, pp. e1928–e1939, Apr. 2021.

[5] X. Xia, W. Yue, B. Chao, M. Li, L. Cao, L. Wang, Y. Shen, and X. Li, ''Prevalence and risk factors of stroke in the elderly in northern China: Data from the national stroke screening survey,'' J. Neurol., vol. 266, no. 6, pp. 1449–1458, Jun. 2019.

[6] A. Alloubani, A. Saleh, and I. Abdelhafiz, ''Hypertension and diabetes mellitus as a predictive risk factors for stroke,'' Diabetes Metabolic Syndrome, Clin. Res. Rev., vol. 12, no. 4, pp. 577–584, Jul. 2018.

[7] A. K. Boehme, C. Esenwa, and M. S. V. Elkind, ''Stroke risk factors, genetics, and prevention,'' Circ. Res., vol. 120, no. 3, pp. 472–495, Feb. 2018.

[8] I. Mosley, M. Nicol, G. Donnan, I. Patrick, and H. Dewey, ‘‘Stroke symptoms and the decision to call for an ambulance,’’ Stroke, vol. 38, no. 2, pp. 361–366, Feb. 2007.

[9] J. Lecouturier, M. J. Murtagh, R. G. Thomson, G. A. Ford, M. White, M. Eccles, and H. Rodgers, ‘‘Response to symptoms of stroke in the UK: A systematic review,’’ BMC Health Services Res., vol. 10, no. 1, pp. 1–9, Dec. 2010.

[10] L. Gibson and W. Whiteley, ‘‘The differential diagnosis of suspected stroke: A systematic review,’’ J. Roy. College Physicians Edinburgh, vol. 43, no. 2, pp. 114–118, Jun. 2013.

[11] N. M. Murray, M. Unberath, G. D. Hager, and F. K. Hui, ‘‘Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: A systematic review,’’ J. NeuroInterventional Surgery, vol. 12, no. 2, pp. 156–164, Feb. 2020.

[12] Y. Zhao, S. Fu, S. J. Bielinski, P. A. Decker, A. M. Chamberlain, V. L. Roger, H. Liu, and N. B. Larson, ‘‘Natural language processing and machine learning for identifying incident stroke from electronic health records: Algorithm development and validation,’’ J. Med. Internet Res., vol. 23, no. 3, Mar. 2021, Art. no. e22951.

[13] B. McDermott, A. Elahi, A. Santorelli, M. O’Halloran, J. Avery, and E. Porter, ‘‘Multi-frequency symmetry difference electrical impedance tomography with machine learning for human stroke diagnosis,’’ Physiological Meas., vol. 41, no. 7, Aug. 2020, Art. no. 075010.

[14] A. Bivard, L. Churilov, and M. Parsons, ‘‘Artificial intelligence for decision support in acute stroke—Current roles and potential,’’ Nature Rev. Neurol., vol. 16, no. 10, pp. 575–585, Oct. 2020.

[15] W. Wang, M. Kiik, N. Peek, V. Curcin, I. J. Marshall, A. G. Rudd, Y. Wang, A. Douiri, C. D. Wolfe, and B. Bray, ‘‘A systematic review of machine learning models for predicting outcomes of stroke with structured data,’’ PLoS ONE, vol. 15, no. 6, Jun. 2020, Art. no. e0234722.

[16] M. S. Sirsat, E. Fermé, and J. Câmara, ‘‘Machine learning for brain stroke: A review,’’ J. Stroke Cerebrovascular Diseases, vol. 29, no. 10, Oct. 2020, Art. no. 105162.

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

[18] M. S. Islam, I. Hussain, M. M. Rahman, S. J. Park, and M. A. Hossain, ‘‘Explainable artificial intelligence model for stroke prediction using EEG signal,’’ Sensors, vol. 22, no. 24, p. 9859, Dec. 2022.

[19] E. Dritsas and M. Trigka, ‘‘Stroke risk prediction with machine learning techniques,’’ Sensors, vol. 22, no. 13, p. 4670, Jun. 2022.

[20] C. Kokkotis, G. Giarmatzis, E. Giannakou, S. Moustakidis, T. Tsatalas, D. Tsiptsios, K. Vadikolias, and N. Aggelousis, ‘‘An explainable machine learning pipeline for stroke prediction on imbalanced data,’’ Diagnostics, vol. 12, no. 10, p. 2392, Oct. 2022.