Gokaraju Rangaraju Institute of Engineering and Technology
(Autonomous)

Department of Information Technology

A Major Project on

Explainable Machine Learning Framework for Stroke Prediction Using Risk Factor Analysis and Class Balancing Techniques

Guide
Dr. Y J Nagendra Kumar
Head of the Department

Team Members :
(21241A12F6) J. Krishna Chaitanya
(21241A12D9) C. Abhinav Reddy
(21241A12G4) Vinod Pawar

Contents :

Abstract :

Stroke is a leading cause of mortality and disability worldwide, making early prediction crucial for effective intervention. Traditional stroke prediction methods are often time-consuming and lack accuracy. This study proposes an Explainable Machine Learning Framework that leverages risk factor analysis and class balancing techniques to enhance prediction accuracy and interpretability. Various machine learning models are trained using key clinical features, with techniques like SMOTE applied to address class imbalance. Furthermore, SHAP and LIME are incorporated to improve model transparency, allowing clinicians to understand critical stroke risk factors. The proposed framework demonstrates high predictive accuracy while ensuring explainability, enabling better decision-making in healthcare.

Introduction :

Stroke is a severe medical condition that occurs due to the disruption of blood flow to the brain, leading to neurological impairment. It remains one of the leading causes of death and long-term disability globally, affecting millions each year. Early detection is critical to prevent severe complications, yet traditional prediction methods are often inefficient, resource-intensive, and lack interpretability.

Machine learning (ML) has emerged as a powerful tool for stroke prediction, capable of analyzing vast amounts of clinical data to identify high-risk patients. However, data imbalance and the lack of explainability in ML models hinder their real-world applicability. This study aims to overcome these challenges by integrating:

- Risk factor analysis to identify key predictors of stroke.
- Class balancing techniques like SMOTE to improve model fairness.
- Explainable AI (XAI) methods such as SHAP and LIME to enhance interpretability.

Aim :

- The aim of this project is to develop an Explainable Machine Learning Framework for early stroke prediction by leveraging risk factor analysis and class balancing techniques. The system ensures high accuracy and interpretability, enabling clinicians to make informed decisions for better patient outcomes.

Scope :

- This project focuses on developing an explainable machine learning framework for accurate and transparent stroke prediction. It addresses class imbalance using SMOTE, enhances risk factor analysis, and integrates SHAP and LIME for interpretability. The system aids healthcare professionals in identifying high-risk patients early, improving clinical decision-making and enabling timely interventions. Its scalable design allows integration into real-time healthcare applications for better stroke prevention and management.

Literature Survey :

TITLE	AUTHOR	PROS	CONS	Dataset	Accuracy
Global Stroke Fact Sheet 2022	Worid Stroke Organization (WO)	Comprehensive global stroke data; highlights growing	Does not offer predictive modeling approach	Global population data	Not applicable
Relationship between Social Support and Participation in Stroke	Systematic Review Team	Identifies trends in incidence, death, and DALYs	Does not include prediction or ML techniques	Giobal stroke statistics	Not applicable
Global Burden of Stroke	Epidemiological Research Group	Distinguishes ischemic vs, hemorrhagic stroke with biomarkers	Moderate sensitivity, not AI-based	189 patients	Speci[[□□]-ᴅ]%
Blood Biomarkers for Stroke Classification	Clinical Validation Study		Focused only on Chinese populaton	2010-2012 Census data	
Prevalence and Risk Factors of Stroke in Elderly (China)	National Stroke Screening Survey (China)	Highlights demographic-specific risk	Focused only on Chinese population	Multiple clinical trials	
Hypertension & Diabetes as Stroke Risk Factors	EMBASE and MEDLINE Review	Identifies key comorbidities contributing to siroke	Genetal review, lacks ML integration	Not applicable	

Existing System :

Current stroke prediction methods rely on traditional statistical models and manual assessments, which are often time-consuming and prone to errors. These models struggle with class imbalance issues, leading to poor prediction accuracy, especially for stroke patients. Additionally, existing approaches lack explainability, making it difficult for doctors to understand the key risk factors contributing to a patient's stroke risk. The absence of effective handling of missing data and imbalanced datasets further limits prediction reliability.

Disadvantages:

- Low Accuracy due to class imbalance and missing values.
- Time-Consuming as predictions require manual processing.
- Lack of Explainability, making it hard for doctors to trust model decisions.

Proposed System :

The proposed system introduces an explainable machine learning framework for accurate stroke prediction. It employs data preprocessing techniques like missing value handling, feature selection (Chi-Square, ANOVA, Mutual Information Score), and class balancing using SMOTE. The system trains six ML models (Random Forest, SVM, KNN, XGBoost, Logistic Regression, Naïve Bayes), with Random Forest achieving the highest accuracy. To improve transparency, SHAP and LIME are used to provide insights into the most critical risk factors, aiding clinicians in better decision-making.

Advantages:

- Higher Accuracy with improved data processing and model selection.
- Faster Predictions, reducing manual effort.
- Explainability using SHAP and LIME, making results interpretable for doctors.
- Scalability for integration into real-time healthcare applications.

System Specifications :

Software :

- Anaconda Navigator/Jupyter Notebook
- Python 3.7 or later

Python libraries :

- TensorFlow
- NumPy
- Matplotlib
- CSV
- Pandas
- SKlearn
- Imblearn
- Shap
- Lime

Hardware :

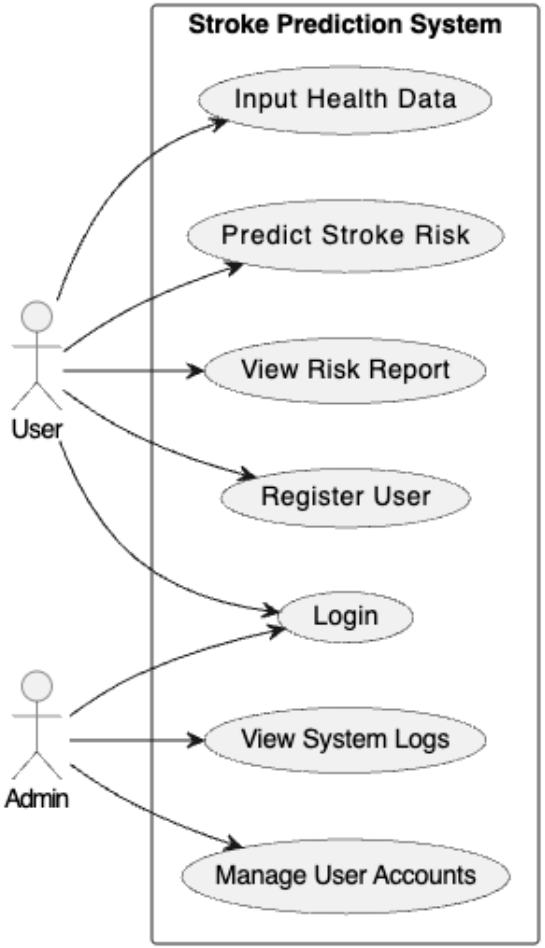
- Processor: Intel Core i5 and above
- Memory: 4GB RAM (Higher specs recommended for better performance)
- Input devices: Keyboard, Mouse
- Internet

Dataset Description :

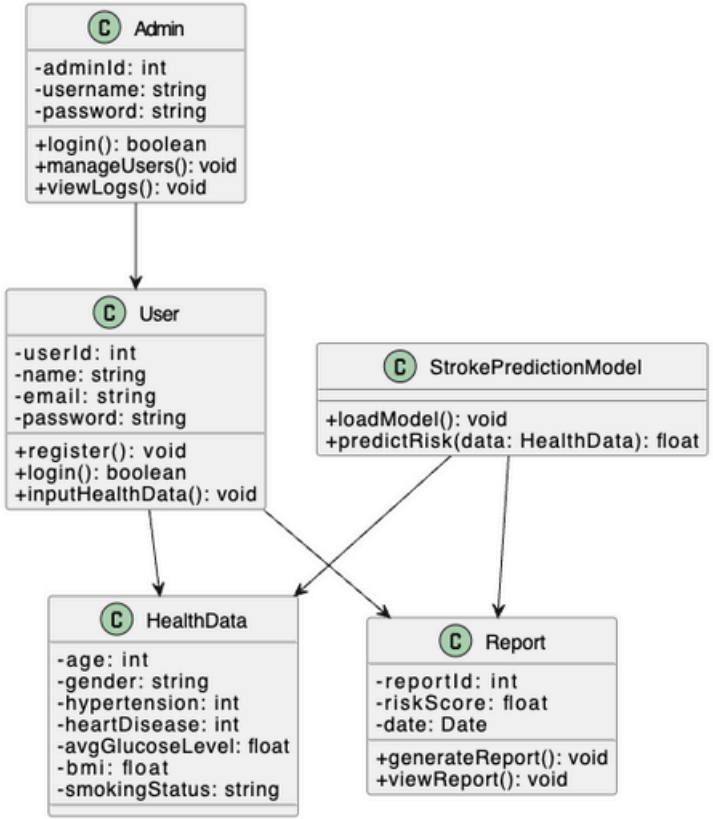
- Total Rows: 5,110
- Total Columns: 12
- Number of Features: Typically 12 attributes

Column Name	Description	Data Type	Values (Examples)
id	Unique identifier for each patient	Numerical	1, 2, 3, ..., 5110
gender	Patient's gender	Categorical	Male, Female, Other
age	Patient's age in years	Numerical	45, 32, 67
hypertension	Whether the patient has hypertension	Numerical	0 (No), 1 (Yes)
heart_disease	Whether the patient has heart disease	Numerical	0 (No), 1 (Yes)
ever_married	Marital status of the patient	Categorical	Yes, No
work_type	Type of employment	Categorical	Private, Self-employed, Govt
Residence_type	Type of residence (urban or rural)	Categorical	Urban, Rural
avg_glucose_level	Average blood glucose level	Numerical	85.5, 102.3, 76.8
bmi	Body Mass Index (BMI) of the patient	Numerical	23.4, 29.7, 31.2
smoking_status	Patient's smoking habits	Categorical	Never smoked, Smokes, Formerly smoked
stroke	Stroke occurrence (Target Variable)	Numerical	0 (No Stroke), 1 (Stroke)

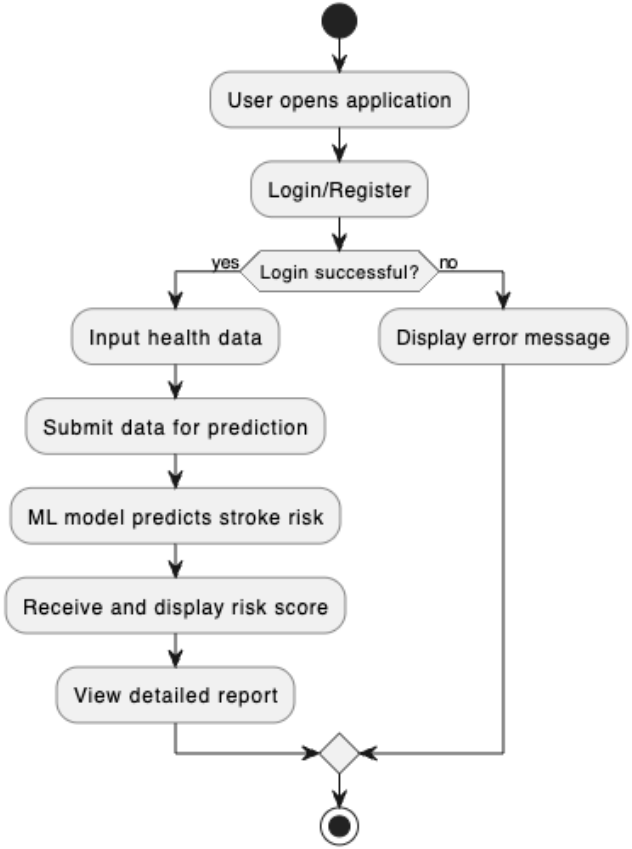
Use case Diagram :



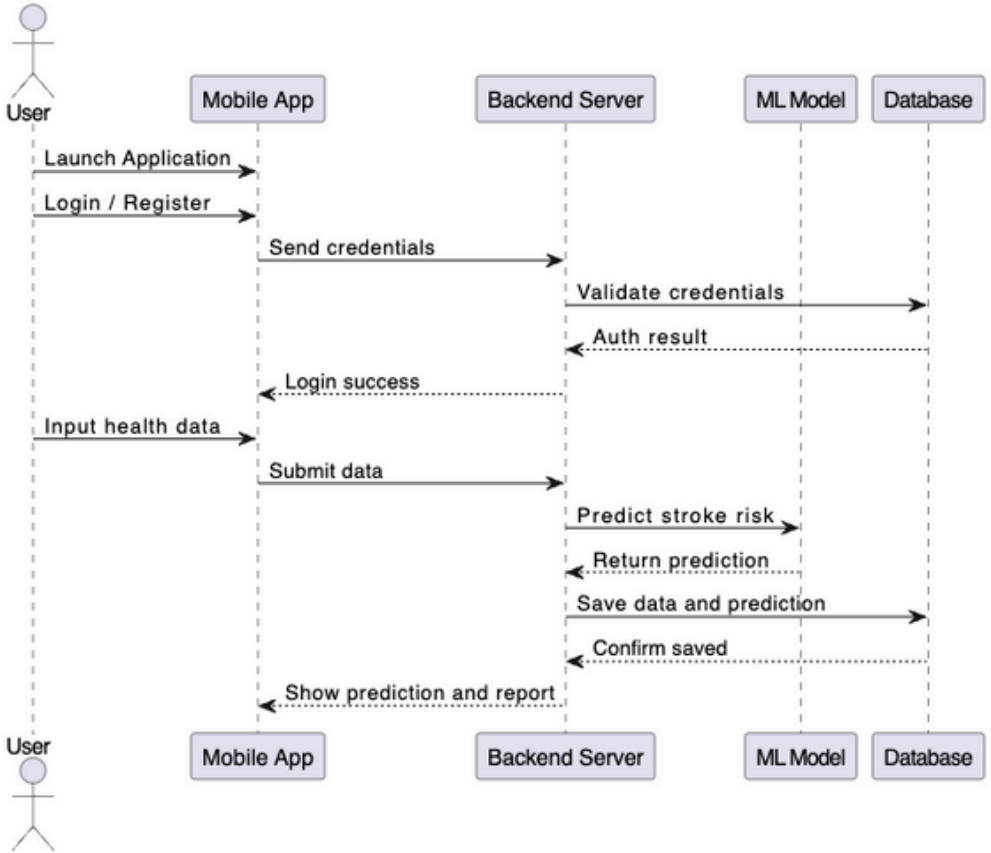
Class diagram:



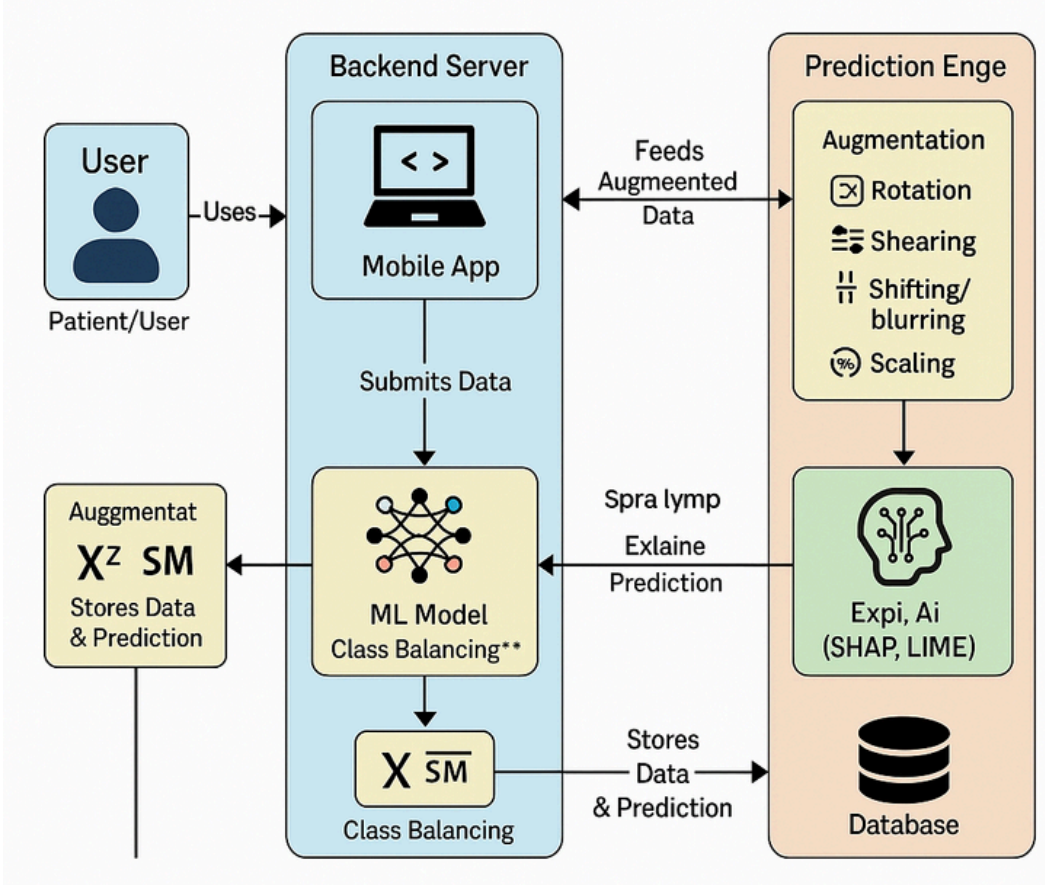
Activity diagram :



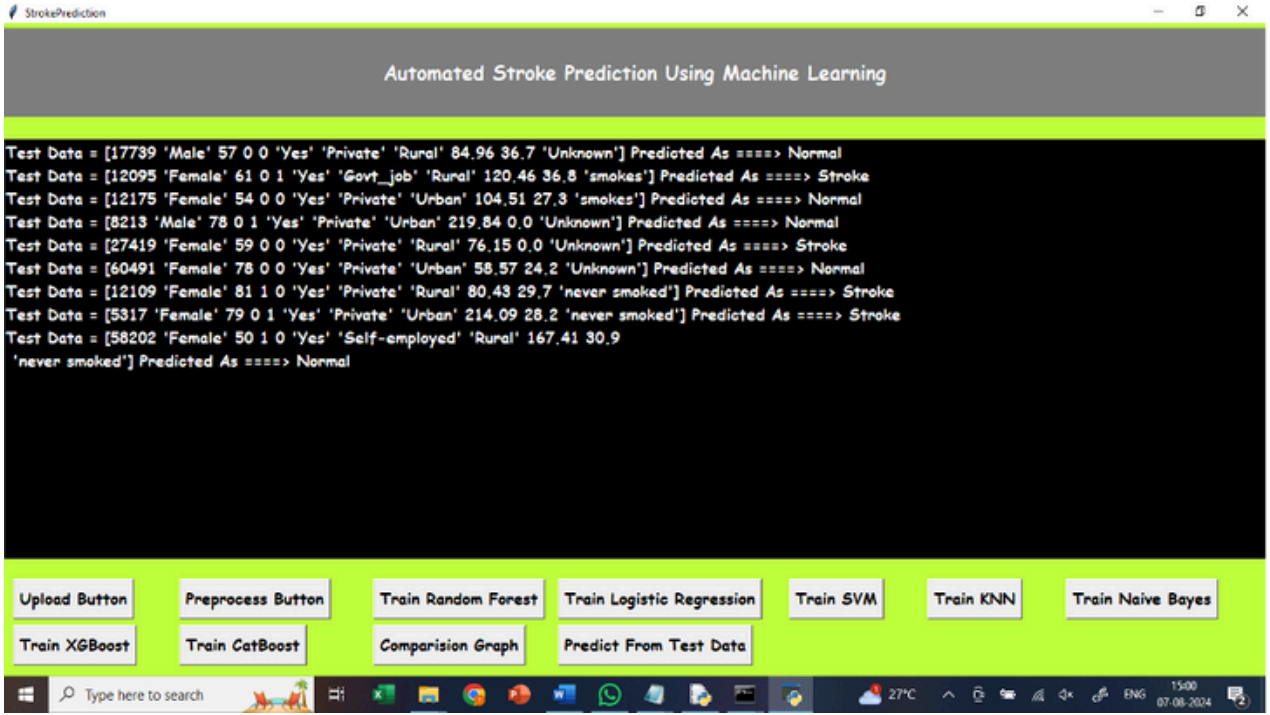
Sequence Diagram :



System Architecture :



Final Output :



In above screen, after uploading test data, it classified data as normal or stroke.

References :

- [1] Learn about Stroke. Available online: **https://www.worldstroke. org/world-stroke-day-campaign/why-strokematters/** 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.

Thank You!