Stroke prediction analysis using machine learning classifiers and feature technique

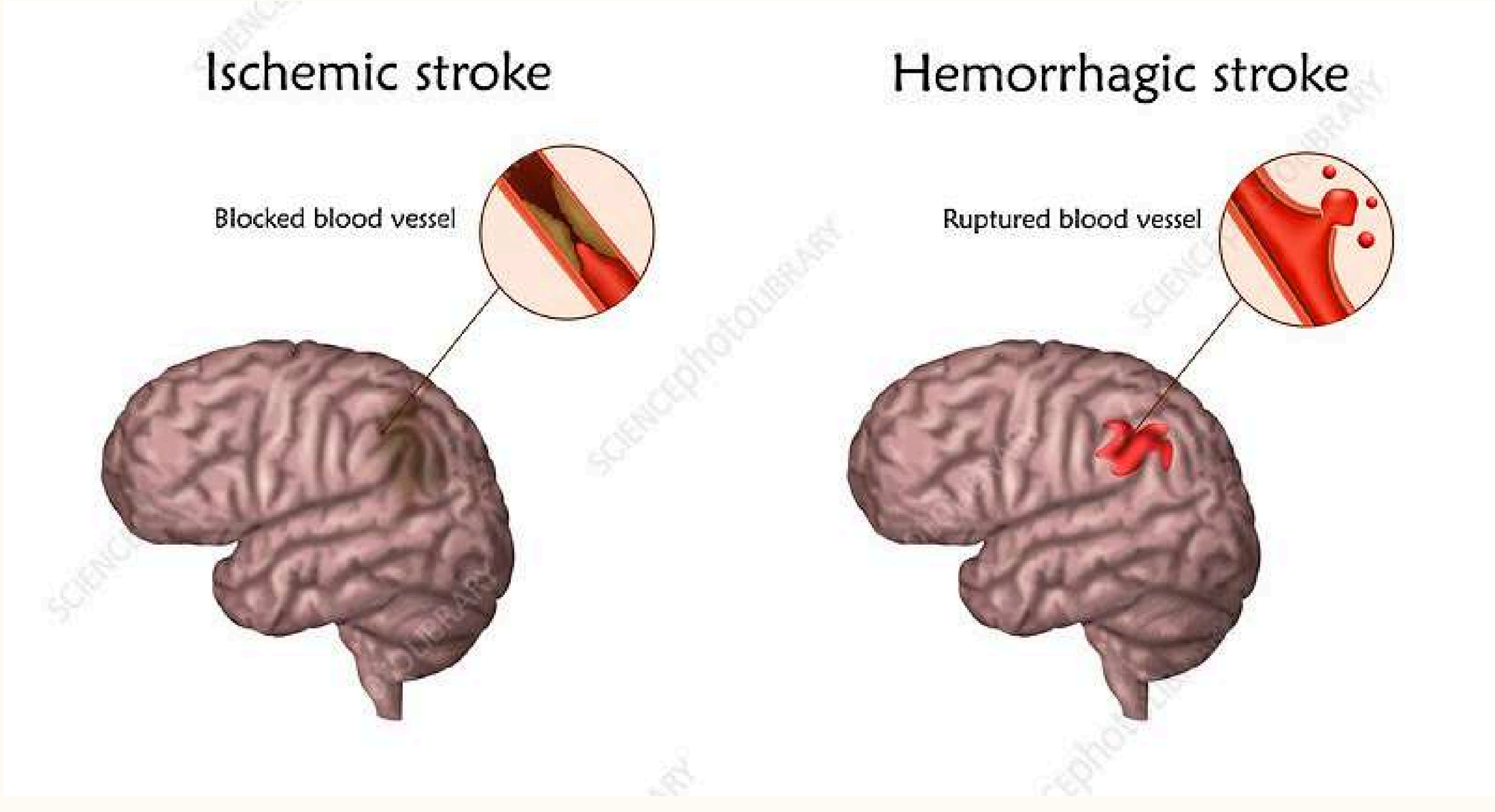
Md. Monirul Islam
Sharmin Akter
Md. Rokunojjaman
Jahid Hasan Rony
Al Amin
Susmita Kar

Introduction

Stroke

- Stroke occurs when blood flow to a part of the brain is interrupted, often due to a blood clot (thrombosis).
- Leads to brain cell death, affecting body functions controlled by that brain region.

Two types of Stroke



Impact of Stroke in the United States

- Major cause of death and disability.
- Early prediction and prevention are crucial.

Predictive Indicators for Stroke

- Risk factors include obesity, physical inactivity, diabetes, age, sex, and race.
- Predictive models can help identify high-risk patients.

Machine Learning in Stroke Prediction

- ML can process large-scale data to forecast stroke risk, offering a tool for early intervention.
- Particularly useful in under-resourced areas where traditional diagnostic tools are lacking.

Role of Mobile Technology

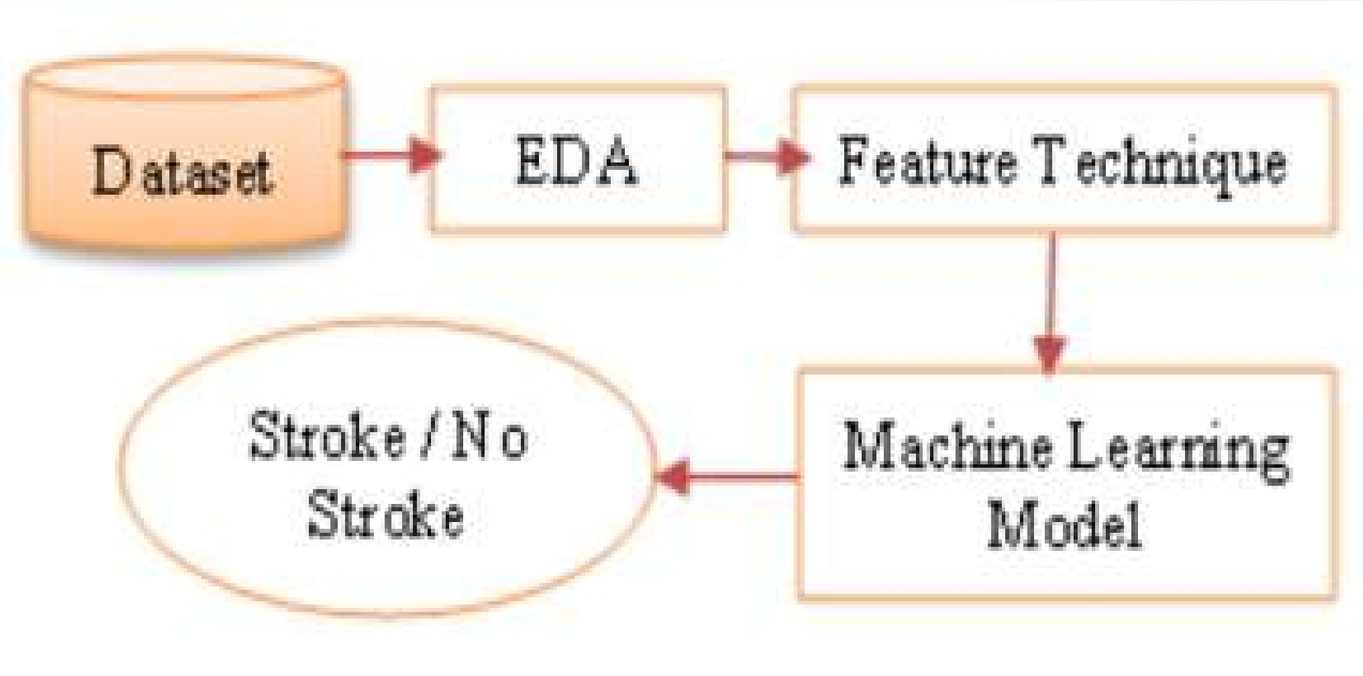
- With over 3.2 billion smartphone users globally, mobile apps can be effective for stroke awareness and prediction.
- Apps provide a user-friendly platform for reaching a broad audience.

Research Motivation

- Aim to improve stroke prediction and prevention through accessible, technology-driven solutions.

Methodology

Block Diagram of the Proposed Methodology



Methodology

Dataset Description

- Dataset consists of 5,110 observations with 12 attributes.
- Key attributes include:
  - gender, age, hypertension, heart_disease, ever_married, work_type, Residence type, average glucose_level, BMI, smoking_status, and stroke
- Stroke is the dependent variable; others are independent.

Methodology

Exploratory Data Analysis

- EDA utilizes data visualization to uncover patterns and spot anomalies.
- Key Steps:
  - Defined missing values.
  - Dropped unnecessary columns (e.g., id).
  - Explored each variable to identify trends and outliers.

- Techniques used include:
  - SMOTE (Synthetic Minority Over-sampling Technique) to handle imbalanced classes.
  - Target variable breakdown: 201 stroke occurrences vs 4,908 non-occurrence.

Machine Learning Models Used

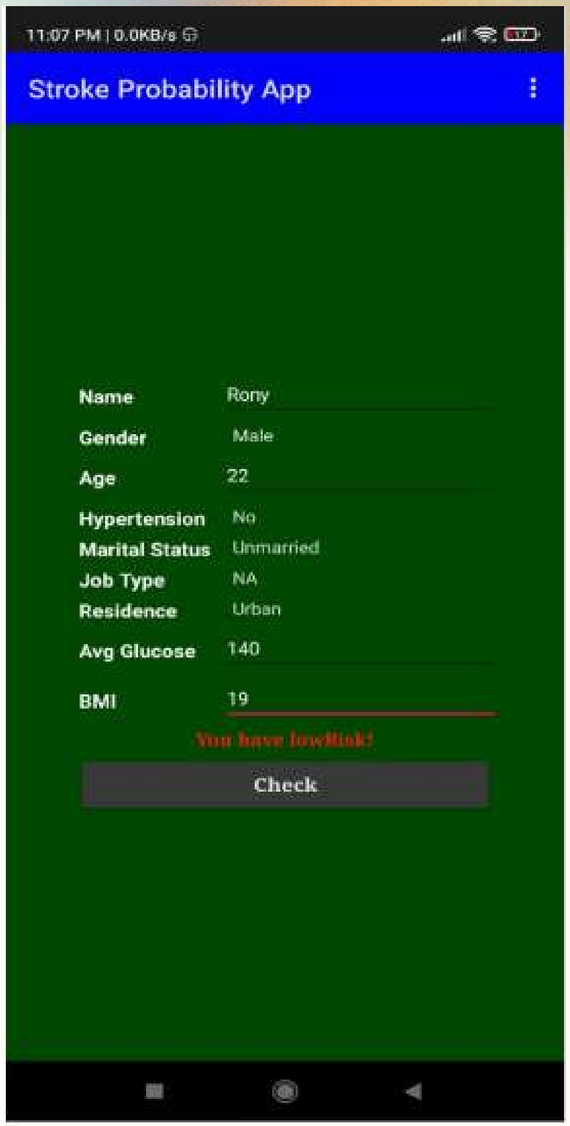
- Applied multiple models to predict stroke:
  - Logistic Regression
  - Decision Tree Classifier
  - KNN
  - Random Forest
- Random Forest performed the best in terms of accuracy.

Random Forest Algorithm

- Supervised Learning Algorithm that constructs a forest of decision trees.
- Uses a bagging method: combines predictions from multiple trees.
- Solves both classification and regression tasks.
- Increased accuracy through randomness in selecting features for node splitting.
- Key Benefit: Reduces overfitting and improves generalization.

User Interface Overview

- Users input data via mobile app
- Data collected: gender, age, work type, heart disease, hypertension, marital status, residence type, BMI, average glucose level, and smoking status
- Data stored in Firestore Cloud Database
- Results processed and shown on the user end



Results and Discussion

- Python used for model implementation and data analysis
- 80% of data used for training, 20% for testing
- Performance metrics: Precision, Recall, F1-Score

Results and Discussion

- Random Forest: Accuracy 96%
- Decision Tree: Accuracy 93%
- K-Nearest Neighbors: Accuracy 90%
- Logistic Regression: Accuracy 87%

ML Model	Accuracies (%)		
	Precision	*Recall*	*F1-Score*
Logistic Regression [23]	87	87	87
DTC [24]	93	93	93
K-NN [25]	90	91	90
Random Forest (proposed)	96	96	96

Conclusion and Recommendation

- Random Forest model showed the highest performance
- SMOTE feature engineering helped handle imbalanced datasets
- Future work: explore deep learning models to enhance accuracy

Thank you!