Stroke Prediction Using Machine Learning Algorithms

Mohamed Elramy
Department of Computer Science
at Modern Academy
mmelramy10@gmail.com

Abdelrahman Elneel
Department of Computer Science
at Luxor University
abodyelneel14@gmail.com

Mohamed Tarek
Department of Navigation science and space technology
mmelramy10@gmail.com

Abanoub George
Department of Navigation science and space technology
mmelramy10@gmail.com

Waled Yaser
Department of Computer Science
at Modern Academy
mmelramy10@gmail.com

ABSTRACT

Stroke is a globally leading cause of death and disability. Early predictions of stroke likelihood can significantly aid in prevention and timely medical intervention. This project aims to develop a robust machine learning model to predict stroke occurrence based on various physiological and lifestyle attributes, such as age, hypertension, heart disease, and BMI. By integrating multiple datasets to address class imbalance and employing rigorous preprocessing techniques, we evaluated several algorithms including Logistic Regression, Decision Trees, Random Forest, and XGBoost. The study successfully identified **XGBoost** as the optimal model, achieving a testing accuracy of **99.45%**, demonstrating its potential as a reliable diagnostic support tool.

INTRODUCTION

A stroke occurs when the blood supply to part of the brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Risk factors include physiological conditions like high blood pressure and heart disease, as well as lifestyle choices such as smoking. The complexity of these factors makes manual prediction difficult. This project leverages machine learning techniques to analyze patient data and classify individuals into "Stroke" or "No Stroke" categories. The primary goal is to build a high-accuracy classification model that minimizes false negatives to ensure high-risk patients are identified correctly.

Due to the critical importance of early detection and intervention, developing a system capable of predicting stroke risk has become increasingly necessary. With the growing availability of medical

RELATED WORK

A strong and accurate diagnosis of stroke risk can save millions of lives and deliver a large amount of information on which machine learning (ML) models can be trained. ML may provide helpful inputs in this regard, especially for making diagnostics based on tabular health records. Previous studies have utilized Support Vector Machines (SVM) and Artificial Neural Networks (ANN) to classify stroke risks. However, medical datasets often suffer from class imbalance (few positive stroke cases vs. many negative cases), which hinders model performance. Our study builds upon this by integrating data augmentation techniques and employing advanced ensemble methods like XGBoost to achieve superior accuracy.

METHODOLOGY

The proposed methodology consists of 4 steps. In step 1 data collection is being performed, step 2 gives an overview of preprocessing and handling imbalance, step 3 exploratory analysis is performed to understand the dataset, and the last step, step 4, includes the hyperparameter tuning by grid search **CV.**

A. *Data Collection*

We utilized the "Healthcare Dataset Stroke Data" available publicly in Kaggle.com. Initially, the dataset contained 5110 rows with attributes such as gender, age, hypertension, heart disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, and smoking_status. To improve model robustness, a supplementary dataset was merged, resulting in a comprehensive dataset of approximately 45,293 records.

B. *Data Preprocessing*

The process of converting raw data into a comprehensible format is known as data preprocessing. Real-world medical data often contains noise and missing values.

Handling Missing Values: The bmi column contained missing values, which were filled using the mean of the column to preserve data distribution.

Data Cleaning: from **fig.1** In the smoking_status attribute, "Unknown" values were replaced with "Passive Smoker" to retain valuable information, and the gender column was standardized.

Outlier Removal: We used the Interquartile Range (IQR) method to remove outliers from numerical features like avg_glucose_level and bmi to ensure the model is not biased by extreme values.

Encoding: Label Encoding was applied to convert categorical variables (e.g., Gender, Work Type, Residence Type) into numeric format.

Scaling: StandardScaler was used to normalize numerical features like Age, Glucose Level, and BMI.

Splitting the Dataset: We divided the data into a 70:30 split. This means we use 70% of the data to train the model while keeping the remaining 30% for testing.

C. *Exploratory Data Analysis*

Exploratory data analysis is used to evaluate the dataset with the aim of summarizing key characteristics. We generated correlation matrices and heatmaps to visualize the relationships between features like Age, Hypertension, and the target variable 'Stroke'. Histograms and count plots were used to analyze the distribution of categorical data.

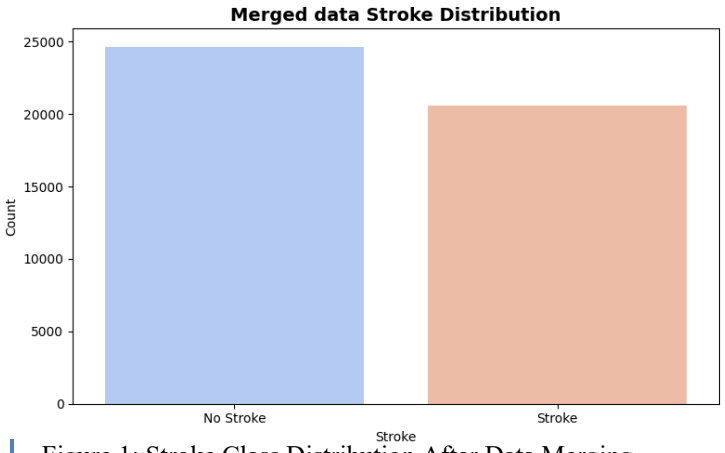


Figure 1: Stroke Class Distribution After Data Merging

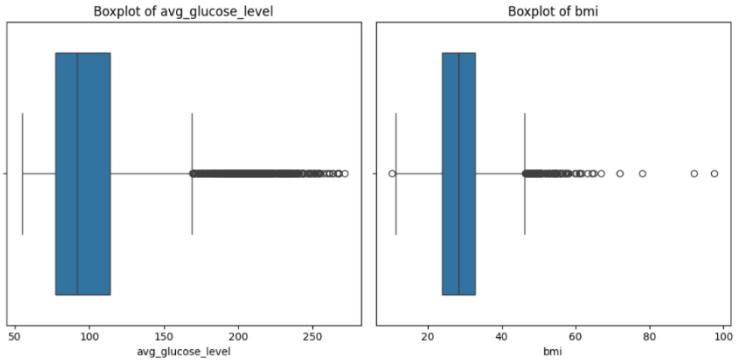


Figure 2: Outlier analysis using boxplots revealed notable outliers in BMI and glucose levels

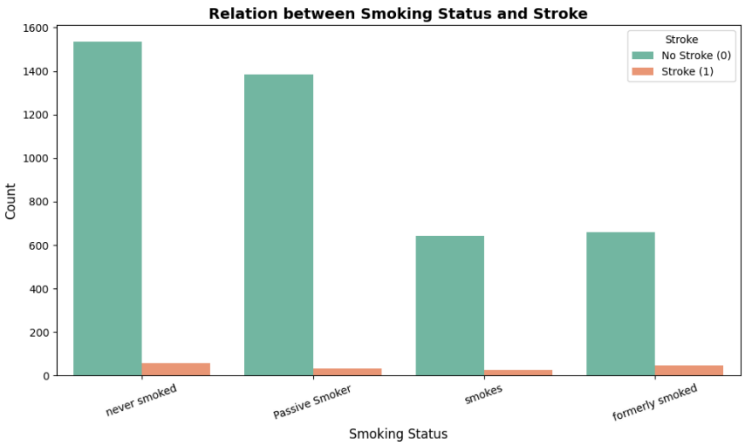


Figure 3: Relation between Smoking Status and Stroke showing the prevalence of stroke among smokers, former smokers, and non-smokers."

D. Hyperparameter tuning by Grid Search CV

Its main goal is to discover the optimal parameters where the model's efficiency is the best. We used GridSearchCV to find the best parameters (e.g., n_estimators, C value) for Random Forest, Logistic Regression, and XGBoost.

IMPLEMENTATION

With the growth of computer technology, predictive modeling is changing. In our project, we use various classification algorithms to predict stroke and use Grid Search to find the most advanced solution. The algorithms employed include:

A. Logistic Regression A statistical model used for binary classification. It models the probability of a certain class or event existing such as pass/fail or stroke/no-stroke.

B. Random Forest This classifier is a meta-estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

C. XGBoost (Extreme Gradient Boosting) An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It uses an ensemble of decision trees and trains them sequentially to correct errors made by previous trees.

RESULT AND DISCUSSION

To evaluate the effectiveness of the Machine Learning algorithms applied in this experiment, we adopted Accuracy, Confusion Matrix, Precision, and Recall.

Model	**Best Score (GridSearch)**
Logistic Regression	~68.9%
Decision Tree	~71.4%
Random Forest	~99.1%
XGBoost	**~99.2%**

Table 5: shows the Confusion Matrix for the final XGBoost model. The model achieved a Test Accuracy of **99.45%**. The confusion matrix revealed:

- **True Negatives:** 7413
- **False Positives:** 44
- **False Negatives:** 31
- **True Positives:** 6100

The results show that the **XGBoost** classifier is the best machine learning algorithm for this dataset. In comparison to Logistic Regression and Decision Trees, XGBoost and Random Forest provided significantly superior accuracy and reliability. The low number of False Negatives (31) indicates the model is highly effective at detecting positive stroke cases, which is critical in a medical context.

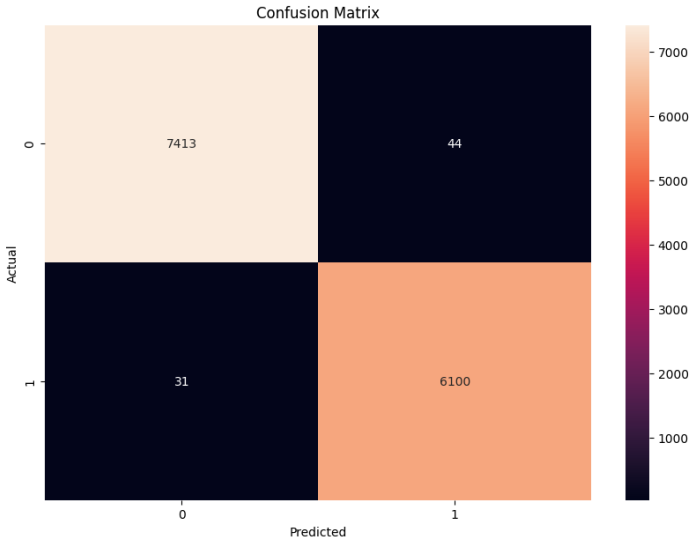
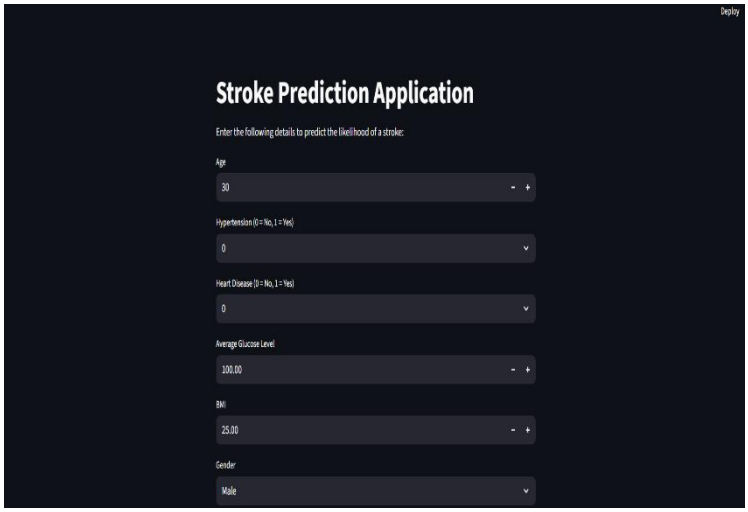


Figure 4: shows the Confusion Matrix for the final XGBoost mode

```
# model prediction
y_pred_final=xgb_final.predict(x_test)
print(f"Accuracy = {accuracy_score(y_test,y_pred_final)}")
print(f"Confusion matrix = {confusion_matrix(y_test,y_pred_final)}")
print(f"Xgb :{classification_report(y_test, y_pred_final)}")
✓ [169] 109ms

Accuracy = 0.994259640859582
Confusion matrix = [[7371   48]
 [  30 6139]]
Xgb :              precision    recall  f1-score   support

           0       1.00      0.99      0.99      7419
           1       0.99      1.00      0.99      6169

    accuracy                           0.99     13588
   macro avg       0.99      0.99      0.99     13588
```

Figure 5: shows the evaluation matrix

GUI



A web-based graphical user interface was implemented using Streamlit to provide an accessible front end for the stroke-prediction model: the sidebar accepts patient attributes (age, gender, hypertension, heart disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, smoking_status) with validation and tooltips, supports single-case entry and CSV batch upload, and exposes a “Predict” action that returns the predicted class (Stroke / No Stroke) together with the prediction probability; the main view presents model summary metrics (accuracy, precision, recall), a confusion matrix, and explainability panels (feature-importance chart and SHAP/partial-dependence visualizations) so clinicians and researchers can explore drivers of each prediction, and the app can be launched locally via streamlit run app.py to allow easy interactive exploration and batch inferenc

VI. CONCLUSION

This work developed a stroke prediction model using supervised machine learning algorithms. Comparative analysis showed that **XGBoost achieved the best performance**, with an accuracy of **99.45%**.
The model can support medical professionals by analyzing physiological risk factors and identifying individuals at high risk of stroke.