AI for Heart Stroke Prevention : Tackling Imbalance with SMOTE

Siva Sai Gopaal Praturi

Advisor: Prof.Yifan Hu

Northeastern University

1. Background

- Stroke is one of the leading causes of death and long-term disability globally, and its impact can be significantly reduced through early detection and intervention.
- Despite advancements in healthcare, predicting stroke remains a challenge due to the rarity of actual stroke cases in datasets, which leads to a severe class imbalance.
- Traditional machine learning models tend to underperform on such imbalanced data, failing to detect minority class cases like stroke, which are the most critical to catch.
- To address this, this project explores the application of machine learning classifiers enhanced with SMOTE (Synthetic Minority Over-sampling Technique) to improve detection of stroke risks in a healthcare dataset.

2. Objectives

The objective of this project are to:

- Build and evaluate ML models that can effectively predict the risk of stroke based on patient data
- Handle severe class imbalance using SMOTE
- Compare performance of models before and after SMOTE
- Assess how well models generalize to real-world data
- Apply interpretability tools (like feature importance and SHAP) to explain predictions for trust and transparency in healthcare settings

3. Dataset

Source: Public healthcare dataset from Kaggle – Stroke Prediction Dataset
Total Rows: ~5,110 entries
Features:

- Demographics: gender, age, ever_married, work_type, Residence_type
- Health indicators: hypertension, heart_disease, avg_glucose_level, bmi
- Lifestyle: smoking_status
- Target: stroke (1 = stroke, 0 = no stroke)

Class Distribution:

- Stroke (Positive): ~5%
- No Stroke (Negative): ~95%
  → *Heavily imbalanced* → motivates the use of SMOTE

4. Methodology

Preprocessing:

- Filled missing values (bmi) using median imputation
- Winsorized numerical columns to cap outliers
- One-hot encoded categorical variables (e.g., gender, work_type)
- Dropped non-predictive features (id)

Imbalance Handling:

- Applied SMOTE (Synthetic Minority Over-sampling Technique) to the training set to oversample stroke cases
- Tested on the original (imbalanced) test set to simulate real-world prediction as well as SMOTE applied data to get a understanding.

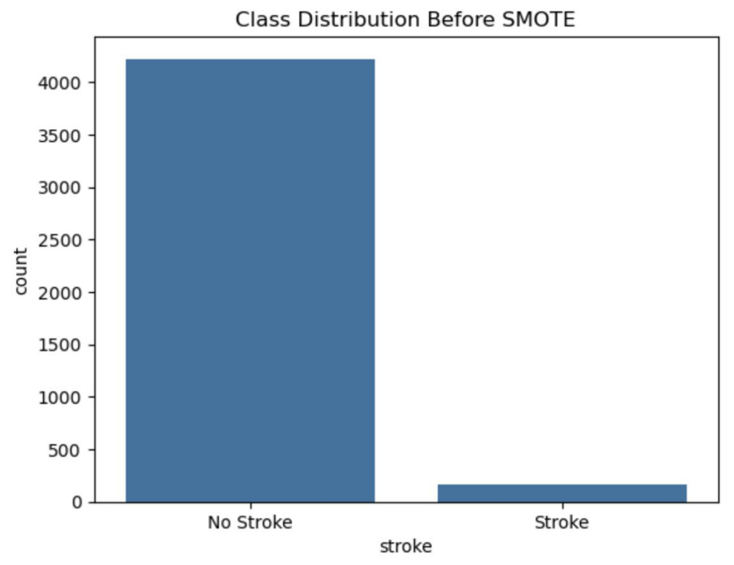

Fig1: Data imbalance before SMOTE application

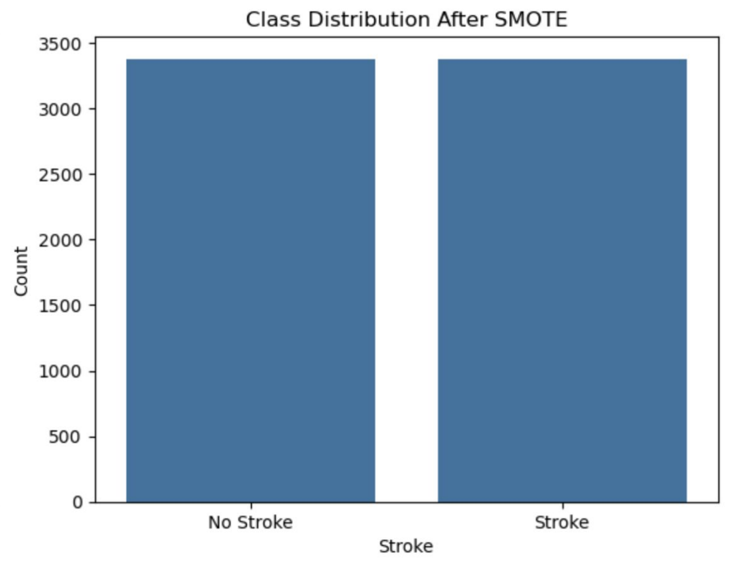

Fig2: Data balanced after SMOTE application

Models Evaluated:

1. Random Forest Classifier
2. Logistic Regression
3. Support Vector Machine (SVM)
4. XGBoost Classifier

Evaluation Metrics:

- Accuracy
- Precision, Recall, F1-Score (especially for stroke class)
- ROC-AUC
- Confusion Matrix
- Train vs Test Comparison
- Threshold tuning (e.g., predict_proba with 0.3 cutoff for better recall)

5. Results and Discussions

- We evaluated four machine learning models (Random Forest, Logistic Regression, SVM, and XGBoost) across three setups to understand the impact of class imbalance and SMOTE-based oversampling.
- 1. Original Data (Without SMOTE): Models trained and tested on the imbalanced data achieved high accuracy but failed to detect stroke cases, with recall near zero. This confirmed the dominance of the majority class and the need for imbalance handling.
- 2. Training SMOTE applied data; Testing original data : Applying SMOTE to the training set improved real-world performance.
  - Logistic Regression showed the most dependable performance, identifying more stroke cases while keeping false alarms low, making it the most trustworthy overall..
  - SVM demonstrated a strong ability to detect strokes but triggered more incorrect alerts
  - Random Forest and XGBoost performed perfectly on the synthetic training data but failed to generalize to real-world cases, missing most actual strokes
- 3. Training and testing SMOTE applied datat: This setup tested model learning on synthetic data.
  - Random Forest and XGBoost achieved 100% accuracy, indicating severe overfitting.
  - Logistic Regression remained stable at 88%, showing healthy learning.
  - SVM performed poorly, suggesting difficulty adapting to synthetic patterns.
- This SHAP plot shows how each feature influences the model's predictions, with color indicating feature value (blue for low, red for high). The spread of dots indicates the distribution of SHAP values for each feature. (Fig 6)
- SHAP analysis identified age, average glucose level, and BMI as the most influential features, supporting their clinical importance.

Random Forest	Logistic Regression	SVM	XGBoost
0.7762	0.7517	0.4646	0.7738

Fig4: ROC AUC values for each model

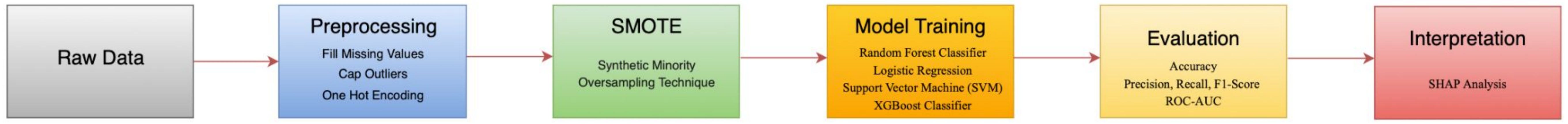

Fig3: Workflow

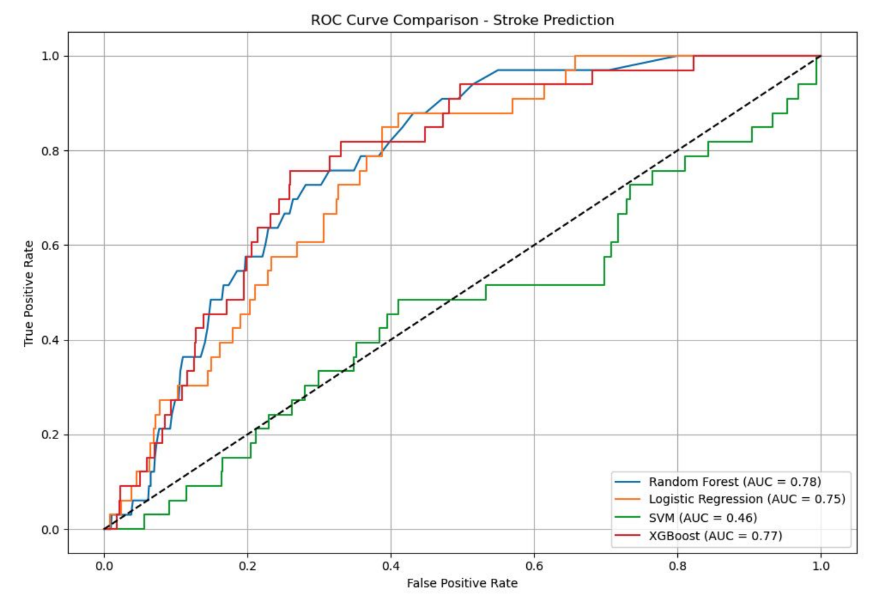

Fig5: ROC AUC (Training on SMOTE data, testing on original data)

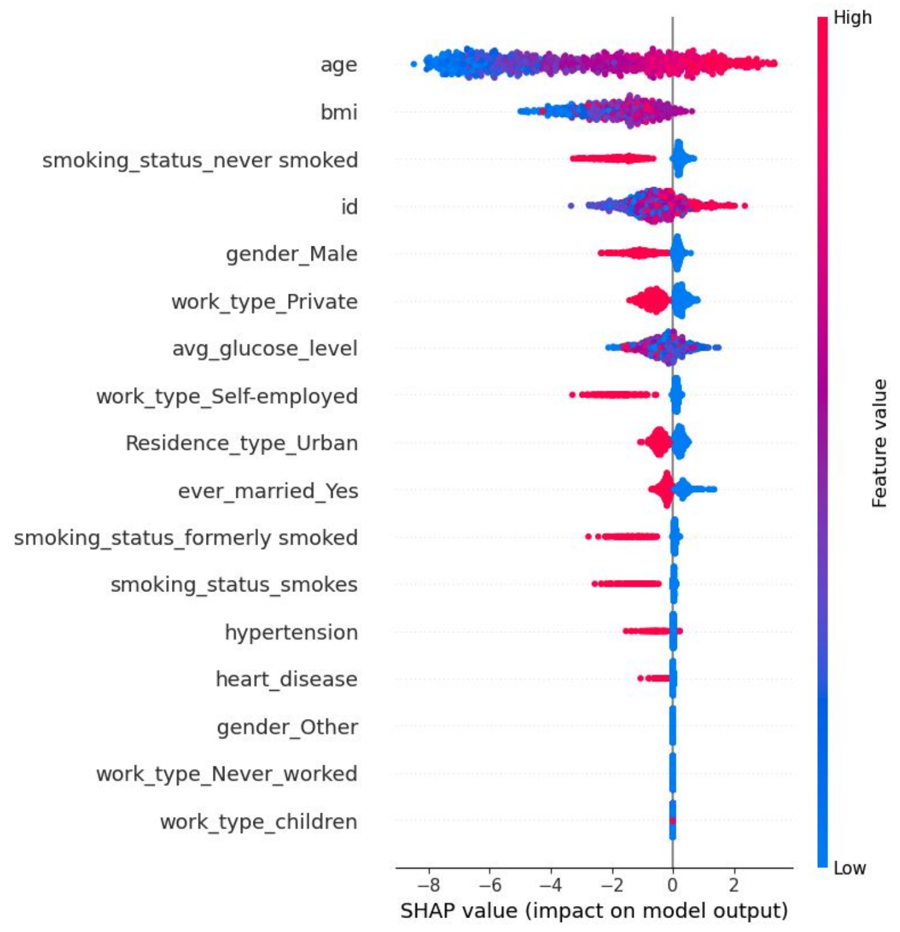

Fig 6: SHAP Analysis

6. Limitations:

- Small positive class (only 33 stroke cases in test set) → limits statistical power
- SMOTE generates synthetic data, which may not fully reflect real-world complexity
- Classifiers like RF and XGB tend to overfit on synthetic data if not tuned carefully
- Dataset lacks richer features like blood pressure, ECG, family history, etc.
- No time-series data — real stroke risk may depend on patient history trends

7. Conclusion, Impact and Future Work

- This project shows that machine learning, paired with SMOTE, can effectively predict stroke risk despite class imbalance. While XGBoost and Random Forest achieved high accuracy, Logistic Regression provided the best balance and generalization. Interpretability tools enables the model to support clinical decision-making when further validated on richer datasets.
- With further validation and real-world data, this approach could support early intervention as a clinical pre-screening tool, potentially helping to save lives.
- Future Work:
  - Use of Real-Time or Wearable Data
  - Validation on Larger & Diverse Datasets
  - Feature Enrichment with Clinical Data