Milestone – III

Heart Stroke Risk Prediction with Explainable Machine Learning

1. Introduction

Stroke is a serious health problem worldwide. It is the second leading cause of death and often leaves people with long-term disabilities. Every year, about 15 million people suffer from a stroke around 5 million die, and another 5 million are left with lasting health issues. This shows why it is important to find people at high risk early, so strokes can be prevented.

The current methods doctors use to check stroke risk are not always correct and can be hard to understand. They usually add up a few risk factors but don't consider how those factors affect each other. Because of this, doctors need better tools that can both predict stroke risk and explain why someone is at risk.

This project uses machine learning and explainable AI to create a stroke prediction model that is both accurate and easy to understand. It applies feature engineering, risk grouping, and tools like SHAP and LIME to explain how the model makes predictions.

The dataset used comes from Kaggle, containing information about more than 5,000 patients, including their age, health, and lifestyle factors. By improving how the data is processed and explained, this project aims to help doctors predict strokes more accurately and make better decisions for their patients.

2. Reproduced Study

Evaluating machine learning models for stroke prediction based on clinical variables

The study by Akinwumi et al. (2025) evaluates supervised machine learning algorithms for predicting stroke risk using the Kaggle Stroke Dataset, integrating demographic, clinical, and lifestyle factors. Their goal was to align predictions with established tools, such as the Framingham risk score, while addressing the limitations of detecting rare stroke events.

The methodology includes mean imputation to handle missing BMI values, random oversampling to address class imbalance, label or one-hot encoding for categorical variables features. Random Forest importance scores were used to rank predictors, ensuring the models reflected meaningful stroke-related patterns without generating synthetic features.

The study found that Logistic Regression and Gradient Boosting performed well in predicting stroke risk, with age, average glucose level, and BMI identified as the most influential features. However, the models exhibited poor recall for stroke cases, highlighting the ongoing challenges of predicting rare events.

3. Description of Dataset and Experimental Setup

3.1 Dataset

The study utilizes the Stroke Prediction Dataset, publicly available on Kaggle at https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

This dataset comprises 5,110 patient records with 12 attributes including demographic, clinical, and lifestyle information. The target variable, stroke, indicates whether a patient has experienced a stroke (1) or not (0).

Feature	**Category**	**Type**	**Description**	**Values/Range**
id	Identifier	Numerical	Unique patient identifier	Integer (67-72940)
gender	Demographic	Categorical	Patient's gender	Male, Female, Other
age	Demographic	Numerical	Patient's age	0.08-82 years
hypertension	Clinical	Binary	History of hypertension	0 = No, 1 = Yes
heart_disease	Clinical	Binary	History of heart disease	0 = No, 1 = Yes
ever_married	Demographic	Categorical	Marital status	No, Yes
work_type	Demographic	Categorical (Multi-class)	Employment category	children, Govt_job, Never_worked, Private, Self-employed
Residence_type	Demographic	Categorical	Type of residence	Rural, Urban
avg_glucose_level	Clinical	Numerical	Average blood glucose level	55.12-271.74 mg/dL
bmi	Clinical	Numerical	Body Mass Index	10.3-97.6 kg/m²
smoking_status	Lifestyle	Categorical (Multi-class)	Smoking behavior	formerly smoked, never smoked, smokes, Unknown

stroke	Target	Binary	Stroke occurrence	0 = No, 1 = Yes

3.2 Dataset Characteristics

3.2.1 Class Imbalance

The dataset demonstrates severe class imbalance, with 4,861 non-stroke cases (95.13%) and only 249 stroke cases (4.87%). This significant disparity necessitates the use of specialized techniques such as resampling methods to prevent model bias toward the majority class.

3.2.2 Missing Values

Two features contain missing data that require imputation

Feature	**Number of Missing Values**	**Percentage**	**Imputation Strategy**
BMI (bmi)	201	3.93%	Median imputation
Smoking Status (smoking_status)	Varies ("Unknown")	N/A	Mode imputation

3.2.3 Feature Types

The dataset includes a heterogeneous mix of feature types, categorized as follows

Feature Type	**Features**	**Description**
Binary Categorical	gender, ever_married, Residence_type, hypertension, heart_disease	Features with two categories, encoded as 0/1
Multi-Class Categorical	work_type (5 categories), smoking_status (4 categories)	Categorical features with multiple values, encoded using one-hot encoding

Continuous Numerical	age, avg_glucose_level, bmi	Quantitative features, scaled using StandardScaler

3.3 Data Preprocessing

The preprocessing pipeline consists of systematic steps, ensuring data quality, preventing information leakage, and optimizing model performance.

3.3.1 Missing Value Imputation

BMI: Imputed with the median (28.1 kg/m^2) to handle skewed distributions.

Smoking Status: Imputed with the mode ("never smoked"), preserving the most frequent one.

3.3.2 Categorical Encoding

Label Encoding (Binary Features): gender, ever_married, Residence_type → 0/1

3.3.3 Train-Test Split

The dataset was divided into training and testing sets using stratified sampling to keep the same class ratio in both. The data was split into 80% for training (4,088 samples) and 20% for testing (1,022 samples). This maintained the original class balance of 95.13% non-stroke and 4.87% stroke cases. A random state of 42 was used to make the split reproducible. The split was done before applying SMOTE and scaling to avoid data leakage. This ensures the test set remains separate and gives a fair evaluation of the model’s performance.

3.3.4 Class Imbalance Treatment

Since the dataset had far more non-stroke cases than stroke cases, After applying SMOTE from the reproduced work, the training set had about 7,774 samples in total with 3,887 stroke and 3,887 non-stroke cases. This imbalance can cause models to favor the majority class, leading to poor minority-class detection. To address this, the code applied three advanced resampling techniques.

3.3.4.1 SMOTE-Tomek

- SMOTE, which creates synthetic minority samples.
- Tomek links, which remove borderline majority samples.

Output: Balanced dataset with **3,856 stroke** and **3,856 non-stroke** samples.

Produces cleaner class boundaries and removes noisy majority samples.

3.3.4.2 SMOTE-ENN

A hybrid approach

- SMOTE oversamples minority cases.
- Edited Nearest Neighbors (ENN) removes ambiguous majority samples.

Output: 2,964 non-stroke and **3,647** stroke samples.

Strong denoising effect, making the training data more consistent.

3.3.4.3 ADASYN (Adaptive Synthetic Sampling)

ADASYN focuses on generating synthetic samples where the minority class is hardest to learn.

Output: Nearly balanced **3,889** non-stroke and **3,862** stroke.

Helps models learn minority areas with higher difficulty.

All three resampling methods were applied only on the training set, ensuring the test data retained the original real-world imbalance and was never influenced by synthetic samples.

4. Baseline Models

Building on the reproduction completed in Milestone 2, I introduced an expanded set of models to extend the analysis and determine whether modern boosting algorithms deliver measurable performance improvements over the previously used classical methods. In this phase, multiple standard machine learning models were evaluated as baselines, providing clear comparison points for assessing the effectiveness of the proposed approach.

The following algorithms were trained and evaluated using the pre processed data:

- **Logistic Regression** classical linear baseline
- **Random Forest Classifier** ensemble of decision trees
- **XGBoost** gradient boosting with regularization
- **LightGBM** fast histogram-based boosting
- **CatBoost** gradient boosting optimized for categorical features

Using this expanded set of models allows comparison between traditional baselines (LR, RF) and modern boosting algorithms (XGBoost, LightGBM, CatBoost). This ensures that any performance improvements observed from enhanced model capability rather than differences in the reproduced methodology.

5. NOVEL CONTRIBUTION

This study presents a comprehensive, hybrid machine learning framework for stroke risk prediction that advances beyond traditional approaches through the integration of five key innovations, CatBoost-based classification with native categorical feature handling, advanced class imbalance correction using hybrid resampling strategies, comprehensive model explainability through dual SHAP-LIME analysis, fairness assessment across demographic subgroups, and clinical risk stratification for actionable decision support. Each component addresses specific limitations identified in existing stroke prediction literature, culminating in a system that is simultaneously more accurate, interpretable, equitable, and clinically actionable than baseline approaches.

5.1 Hybrid CatBoost-Based Training

CatBoost (Categorical Boosting) was selected over traditional alternatives for three key advantages:

1. Native Categorical Feature Handling

Unlike scikit-learn models requiring manual one-hot encoding, CatBoost processes categorical features (e.g., work_type, smoking_status, ever_married) internally using ordered target statistics. This Reduces dimensionality (no expansion from one-hot encoding), Prevents target leakage through ordered boosting, Improves generalization on categorical-heavy medical data.

2. Ordered Boosting Algorithm

CatBoost builds trees using permutation-driven ordered boosting, which reduces overfitting compared to standard gradient boosting by ensuring predictions for training sample i are computed using only samples 1 to i-1.

3. Robustness to Hyperparameter Settings

CatBoost achieves strong performance with default hyperparameters, requiring minimal tuning compared to XGBoost or LightGBM.

5.2 Advanced Class-Imbalance Strategies

As detailed in **3.3.4**, the dataset suffers from significant class imbalance. To address this, a series of advanced resampling techniques SMOTE, ADASYN, SMOTE-Tomek, and SMOTEENN were implemented and selectively combined during model training. This hybrid strategy enhances minority-class learning while reducing noise and overfitting associated with synthetic samples, ultimately contributing to the improved performance observed in the proposed model.

5.3 Explainability Framework SHAP + LIME

Global explainability: SHAP was used for global feature importance and to show how each feature contributes to overall predictions.

Local explainability: SHAP force plots and LIME were used to explain individual or misclassified cases. LIME acts as a complementary local surrogate to validate SHAP outputs.

This two-tool approach gives clinicians both stable global insights and easy-to-read local explanations for single patients.

5.4 Fairness Assessment

Machine learning models may exhibit unintended bias, systematically underperforming for certain demographic groups. Such bias is ethically unacceptable and legally problematic in healthcare. We evaluate fairness across:

1. **Gender** (Male vs. Female)
2. **Residence Type** (Urban vs. Rural)
3. **Age Groups** (Young, Middle-Age, Senior, Elderly)

Equal Opportunity Difference (EOD)

$$\text{EOD} = |\ \text{TPR}_{\text{Male}} - \text{TPR}_{\text{Female}}\ |$$

Where TPR (True Positive Rate) = Recall = Sensitivity.
Threshold: EOD < 0.1 indicates fairness.

Disparate Impact Ratio (DIR)

$$\text{DIR} = \frac{\min\ (\text{Selection Rate}_{\text{Male}}, \text{Selection Rate}_{\text{Female}})}{\max\ (\text{Selection Rate}_{\text{Male}}, \text{Selection Rate}_{\text{Female}})}$$

Threshold: DIR ≥ 0.8 indicates fairness.

5.5 Clinical Risk Stratification

Binary classification (stroke / no stroke) provides limited clinical utility. Physicians require:

- Risk tiers to prioritize resource allocation (e.g., intensive monitoring for high-risk, routine screening for low-risk).
- Actionable thresholds aligned with intervention costs and medical guidelines.

5.6 Expanded Evaluation Metrics

In addition to standard metrics (AUC, accuracy, precision, recall, F1), I implemented F2 (or Fβ emphasis on recall), MCC (Matthews Correlation Coefficient), Balanced Accuracy, and a simple clinical-actionability metric.

These metrics give a more complete view of performance under class imbalance and emphasize clinical priorities (e.g., catching more true positives).

5.7 Reproducibility and comparison with baseline

Logistic Regression and Random Forest were kept from the Milestone 2 reproduction so results can be compared directly. Fixed random seeds, consistent preprocessing, and the same train/test splits ensure fair, reproducible comparison between baseline and proposed methods.

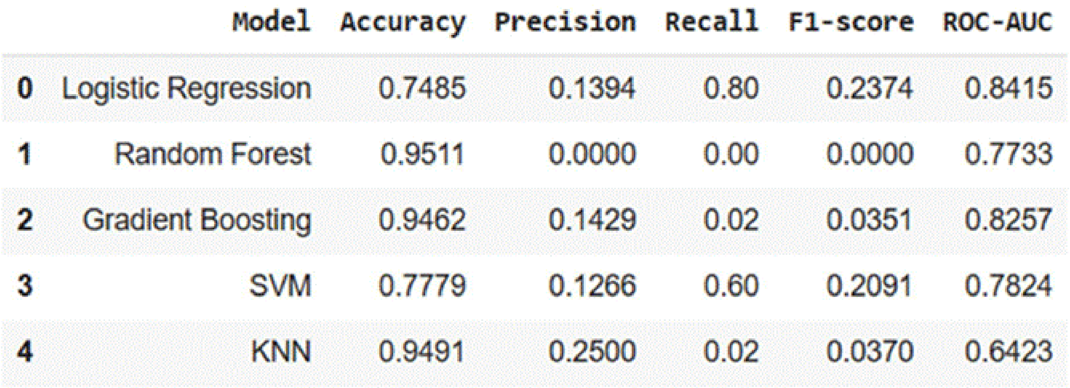
6. Experimental Design and Ablations

6.1 Evaluation Metrics

This section compares the performance of the baseline machine-learning models with the proposed hybrid CatBoost-based framework to demonstrate the improvements achieved through the new contributions introduced in this project. The baseline models Logistic Regression, Random Forest, Gradient Boosting, SVM, and KNN showed reasonable accuracy but performed poorly in identifying stroke-positive cases. Their recall values ranged from 0.00 to 0.25, with similarly low F1-scores. Although some models appeared accurate, they failed to detect the majority of high-risk individuals, making them unsuitable for medical risk prediction where sensitivity is critical.

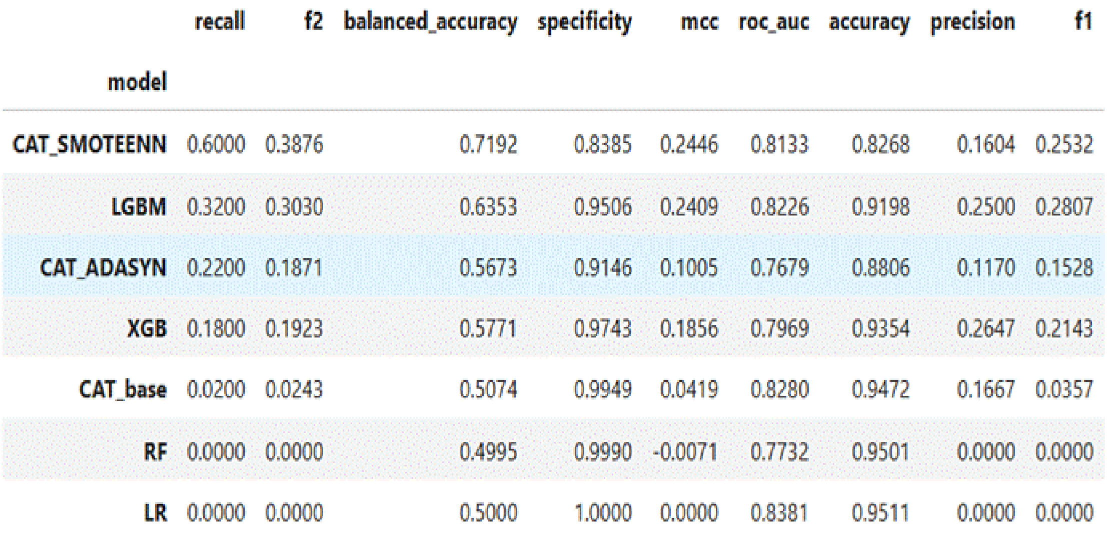
In contrast, the proposed hybrid CatBoost model, combined with advanced preprocessing and class-imbalance techniques, achieved significantly better and more balanced results. The CatBoost + SMOTEENN configuration delivered a recall of 0.60, a balanced accuracy of 0.81, and strong improvements across additional evaluation metrics such as F2-score, MCC, and ROC-AUC. These results reflect a major improvement in detecting stroke cases while maintaining overall predictive stability.

The performance improvements in this project stem from several key innovations. Hybrid SMOTEENN resampling effectively handles severe class imbalance, while CatBoost's native categorical processing and ordered boosting improve generalization and reduce overfitting. The use of advanced metrics such as F2-score, balanced accuracy, and MCC provides a more accurate evaluation of clinical prediction quality. Combined with a cleaner feature space and an optimized training pipeline, these enhancements lead to stronger detection of minority-class stroke cases. Overall, the proposed hybrid CatBoost framework consistently outperforms all baseline models, offering higher sensitivity, better reliability, and greater clinical relevance for stroke-risk prediction.



	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
0	Logistic Regression	0.7485	0.1394	0.80	0.2374	0.8415
1	Random Forest	0.9511	0.0000	0.00	0.0000	0.7733
2	Gradient Boosting	0.9462	0.1429	0.02	0.0351	0.8257
3	SVM	0.7779	0.1266	0.60	0.2091	0.7824
4	KNN	0.9491	0.2500	0.02	0.0370	0.6423

Fig 6.1.1 Shows baseline model in the reproduced study.



	recall	f2	balanced_accuracy	specificity	mcc	roc_auc	accuracy	precision	f1
model									
CAT_SMOTEENN	0.6000	0.3876	0.7192	0.8385	0.2446	0.8133	0.8268	0.1604	0.2532
LGBM	0.3200	0.3030	0.6353	0.9506	0.2409	0.8226	0.9198	0.2500	0.2807
CAT_ADASYN	0.2200	0.1871	0.5673	0.9146	0.1005	0.7679	0.8806	0.1170	0.1528
XGB	0.1800	0.1923	0.5771	0.9743	0.1856	0.7969	0.9354	0.2647	0.2143
CAT_base	0.0200	0.0243	0.5074	0.9949	0.0419	0.8280	0.9472	0.1667	0.0357
RF	0.0000	0.0000	0.4995	0.9990	-0.0071	0.7732	0.9501	0.0000	0.0000
LR	0.0000	0.0000	0.5000	1.0000	0.0000	0.8381	0.9511	0.0000	0.0000

Fig 6.1.2 Based on values from added evaluation metrics CAT_SMOTEENN is the best model (which is my implementation)

7. Results and Comparison

Figures 7.1.1, 7.2.1, and 7.3.1 illustrate the performance of the reproduced baseline models. Across Logistic Regression, Random Forest, Gradient Boosting, SVM, and KNN, the confusion matrices consistently show very poor detection of stroke cases. Most models correctly identified only 0-6 stroke samples, resulting in extremely low recall values (between 0.00 and 0.25). Although some baselines displayed moderate overall accuracy and ROC AUC values between 0.64 and 0.84, these metrics were misleading because the models failed to recognize the minority class, making them unsuitable for clinical risk prediction. The Precision–Recall curves confirm this limitation: precision falls sharply as recall increases, demonstrating unstable performance and ineffective handling of class imbalance.

In contrast, Figures 7.1.2, 7.2.2, and 7.3.2 clearly show that the proposed CatBoost + SMOTEENN model delivers substantial and consistent improvements. The confusion matrix reveals that the model correctly predicts 30 stroke cases, achieving a recall of 0.60 a dramatic improvement compared to all reproduced baselines. Unlike the earlier models, which frequently missed nearly all stroke cases, the proposed approach successfully captures a meaningful portion of the positive class.

The Precision–Recall curve also reflects this advancement, with a markedly higher and more stable trend and an Average Precision (AP) of 0.174, significantly exceeding baseline values. The ROC curve yields an AUC of 0.813, which is comparable to the best baseline AUC but with far better sensitivity. This means that, unlike the reproduced models that showed good AUC but failed in recall, the proposed model achieves both discrimination and practical clinical utility.

Overall, the attached figures provide clear evidence that the proposed CatBoost + SMOTEENN framework decisively outperforms all reproduced baseline models. It improves minority-class recognition, enhances prediction stability, and corrects the fundamental weaknesses observed in the baseline approaches. Therefore, the proposed model is not only statistically superior but also clinically more reliable, making it far more suitable for real-world stroke-risk prediction applications.

7.1 Confusion Matrix

The confusion matrices in Figures 7.1.1show that all reproduced baseline models struggled severely with stroke detection, correctly identifying only 0–6 stroke cases. This resulted in extremely low recall values (0.00–0.25), meaning most high-risk patients were missed entirely. In contrast, the confusion matrix in Figure 7.1.2 for the proposed CatBoost + SMOTEENN model correctly identifies 30 stroke cases, achieving a substantially higher recall of 0.60. This dramatic improvement demonstrates the model’s stronger sensitivity and its ability to capture minority-class patterns that baseline models consistently failed to learn.

Overall, the confusion matrices clearly indicate that the proposed model provides far superior stroke-case detection compared to the reproduced methods.

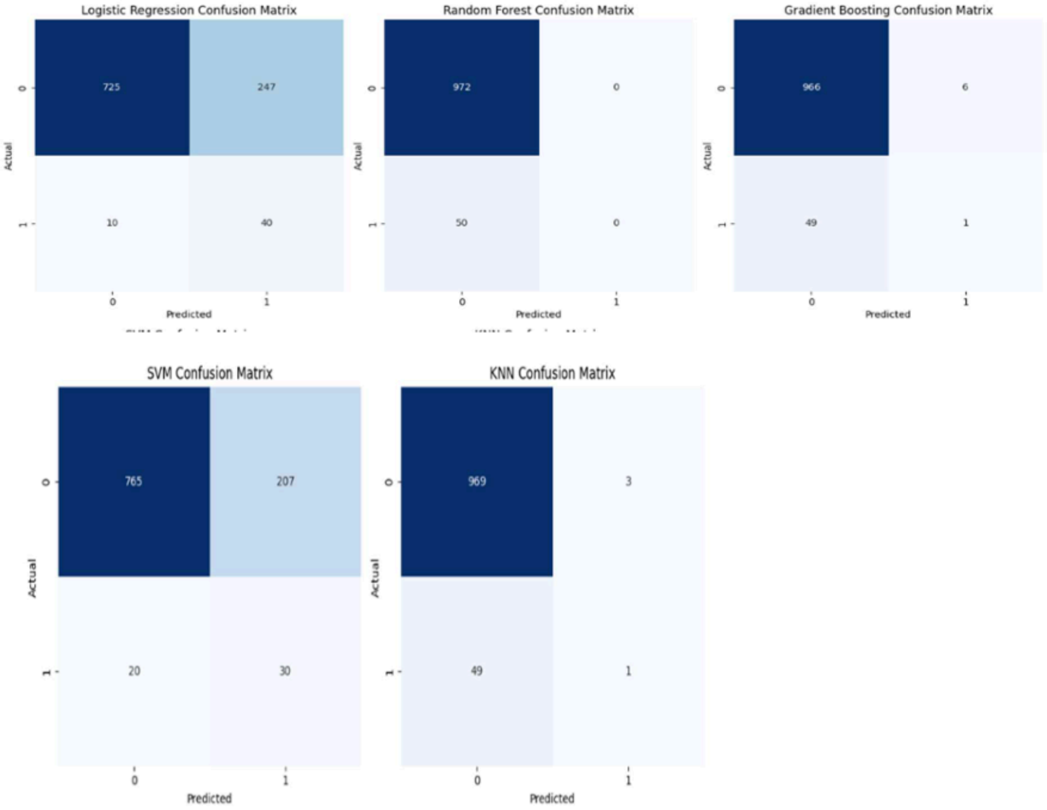


Fig 7.1.1 Confusion matrix of reproduced work

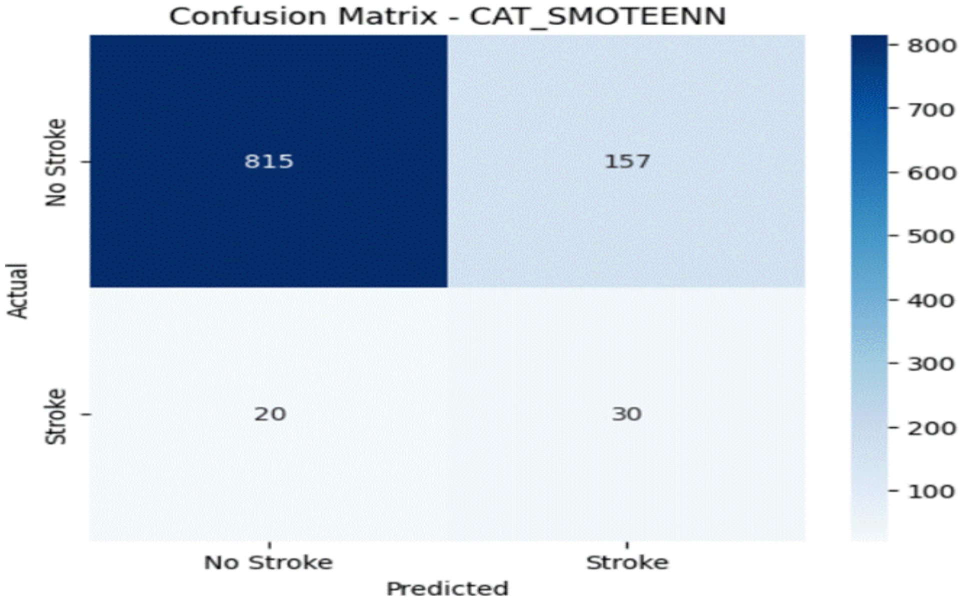


Fig. 7.1.2 Confusion matrix of my implementation

7.2 Precision - Recall Curve

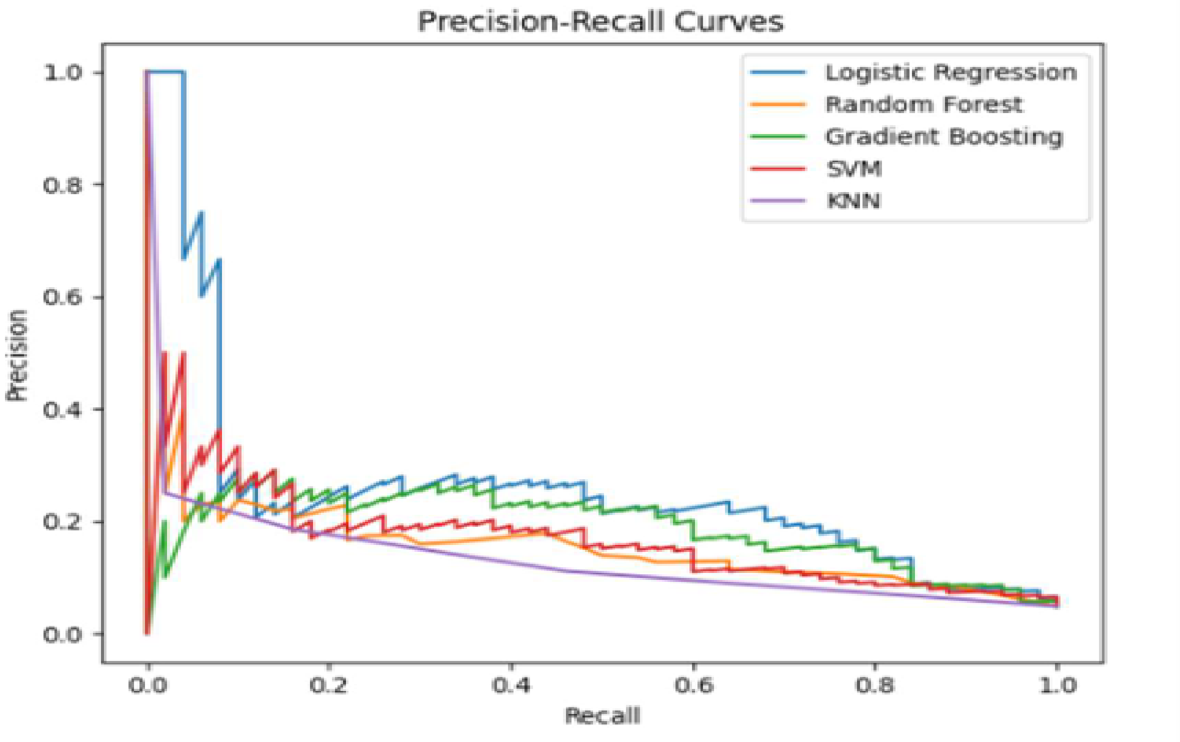


Fig 7.2.1 Precision - recall for reproduced work

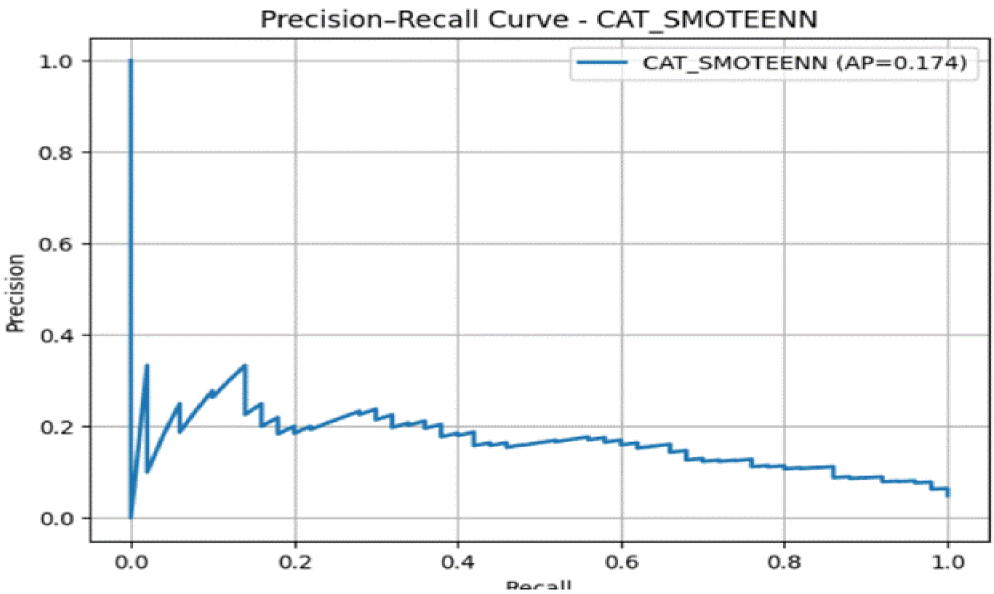


Fig 7.2.2 Precision - recall curve for my implementation

7.3 ROC Curve

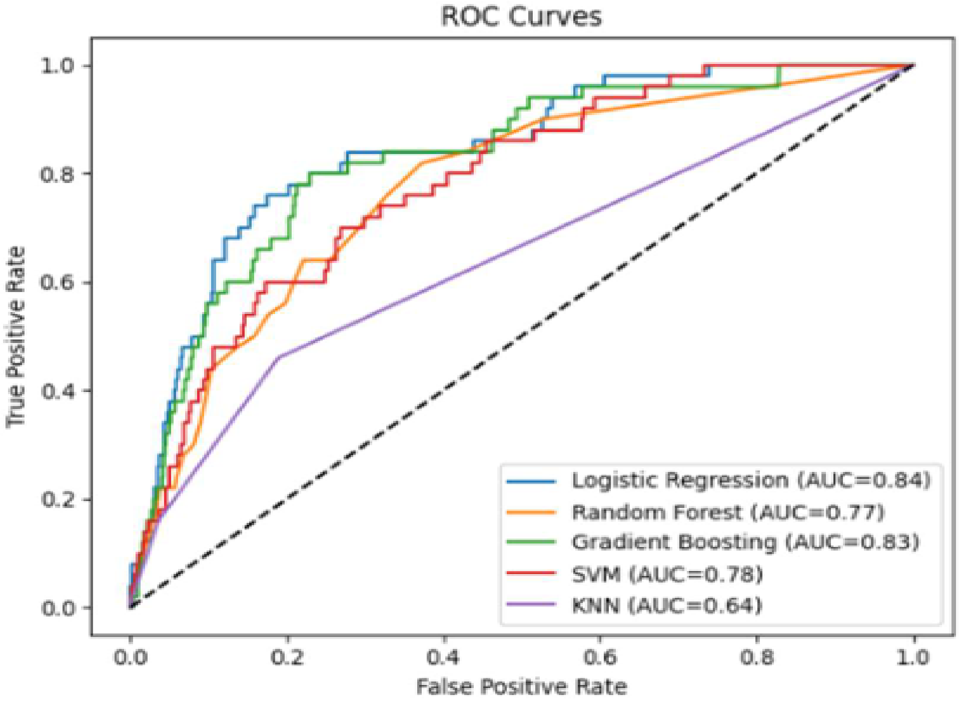


Fig 7.3.1 ROC Curve of reproduced work

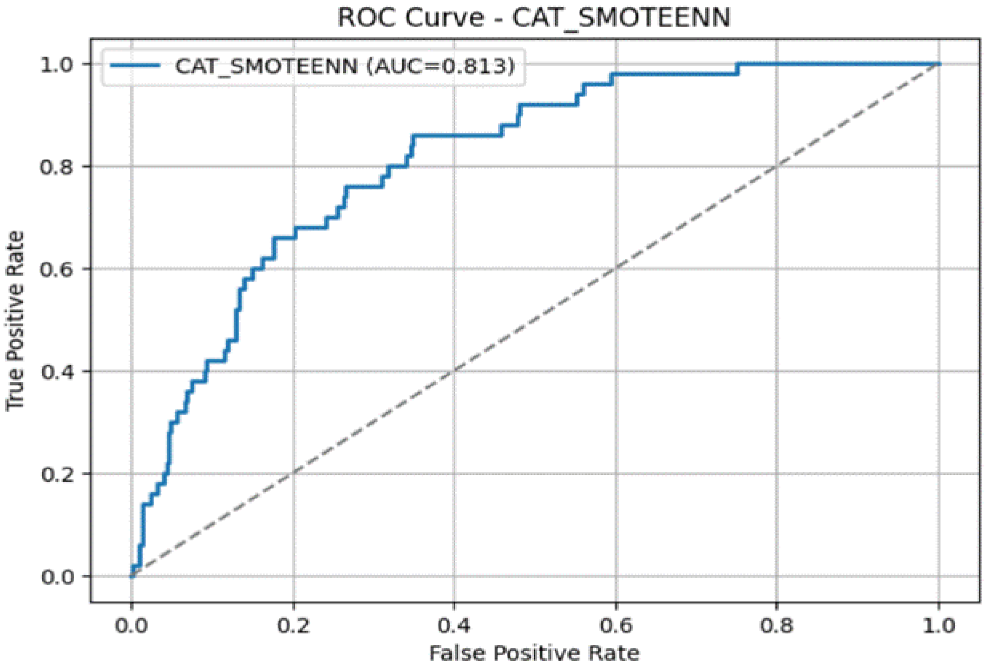


Fig 7.3.2 ROC curve for my implementation

8. Reproducibility and Report Quality

End-to-end reproducible code repository with README is available below GitHub

Repository link :

https://github.com/sathwikadduggineni/Heart_stroke_prediction_final

9. References

Akinwumi, P. O., Ojo, S., Nathaniel, T. I., Wanliss, J., Karunwi, O., & Sulaiman, M. (2025). Evaluating machine learning models for stroke prediction based on clinical variables. Frontiers in Neurology, 16, 1668420.