# Healthcare Provider Fraud Detection Analysis

## 1. Abstract

Healthcare fraud detection is an essential task for maintaining the integrity of healthcare systems worldwide. This project focuses on developing machine learning models to identify fraudulent healthcare providers using Medicare claims data. We implemented several models, including Logistic Regression, Random Forest, and MLP Classifier, to predict fraud based on features derived from inpatient, outpatient, and beneficiary data. We conducted extensive experiments using various model configurations and evaluated their performance using accuracy, precision, recall, F1-score, and ROC-AUC. The MLP Classifier achieved the highest accuracy of 91.65%, with substantial improvements in recall compared to other models.

## 2. Introduction

### 2.1 Objectives

Healthcare fraud detection involves identifying providers submitting fraudulent claims, which cost healthcare systems billions annually. The objective of this project is to create an effective machine learning model to classify providers as either fraudulent or non-fraudulent based on claims data. These models help reduce unnecessary claims and improve healthcare efficiency.

### 2.2 Problem Formulation

We framed the task of detecting fraudulent providers as a **binary classification** problem using supervised machine learning. The dataset consists of various claims records labeled as either fraudulent or non-fraudulent. The challenge lies in identifying subtle fraud patterns within a highly imbalanced dataset.

**Input:** Claims data including diagnosis codes, admission/discharge dates, reimbursement amounts, and patient demographics.

**Output:** A binary classification label indicating whether the provider is fraudulent.

# 3. Methodology

## 3.1 Dataset

The dataset contains records from three tables: **inpatient, outpatient, and beneficiary** data. Important features include `AdmissionDates`, `DischargeDates`, `ClaimAmounts`, and patient demographics like age and chronic conditions.

**Preprocessing:**

1. **Handling Missing Data**: Missing values in critical columns like `DOD` (date of death) were filled, and features such as `NumProc`, `NumPhysicians`, and `AdmissionDays` were created.
2. **Feature Engineering**: Derived features like the number of procedures (`NumProc`), number of unique claims (`NumUniqueClaims`), and patient age were calculated.
3. **Target Label Creation**: The target variable, `PotentialFraud`, was mapped to binary labels (1 for fraud, 0 for non-fraud).

## 3.2 Machine Learning Models

### Logistic Regression:

Logistic Regression served as the baseline model. It provides a simple yet interpretable approach, utilizing features like reimbursement amounts and chronic conditions to predict fraud.

### Random Forest Classifier:

Random Forest, a more complex model, was used to handle non-linear relationships and imbalanced data. Hyperparameter tuning was performed using **GridSearchCV** to optimize parameters such as `n_estimators` and `min_samples_split`.

### MLP Classifier:

The MLP (Multilayer Perceptron) Classifier, a neural network-based model, was employed to capture more complex patterns in the data. It used a 2-layer architecture and logistic activation for binary classification.

### 3.3 Model Training and Evaluation

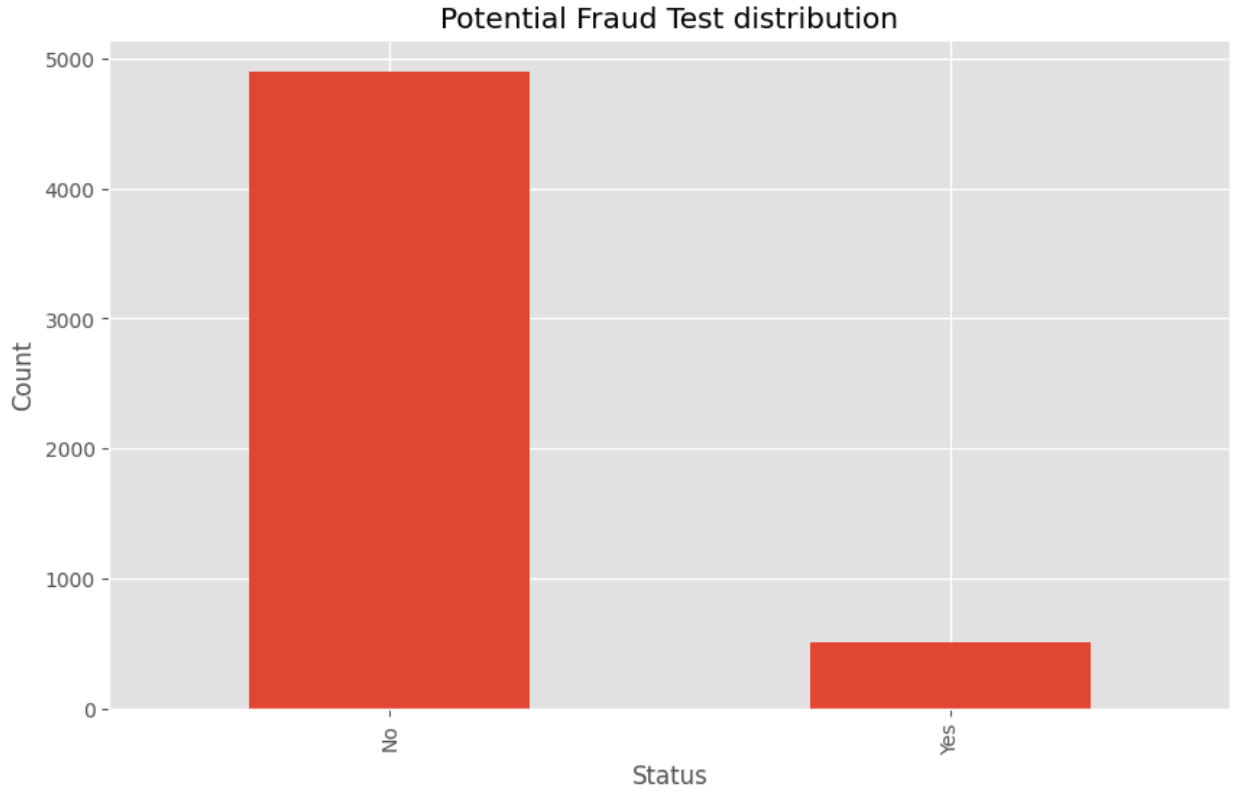
The dataset was split into **70% training** and **30% validation**. Each model was trained on the training data and evaluated on the validation set. The following evaluation metrics were considered:

- **Accuracy**: Overall correctness of the model.
- **Precision, Recall, F1-Score**: Important for classifying fraud accurately.
- **ROC-AUC**: Measures the area under the ROC curve, focusing on model discrimination between classes.

# 4. Results

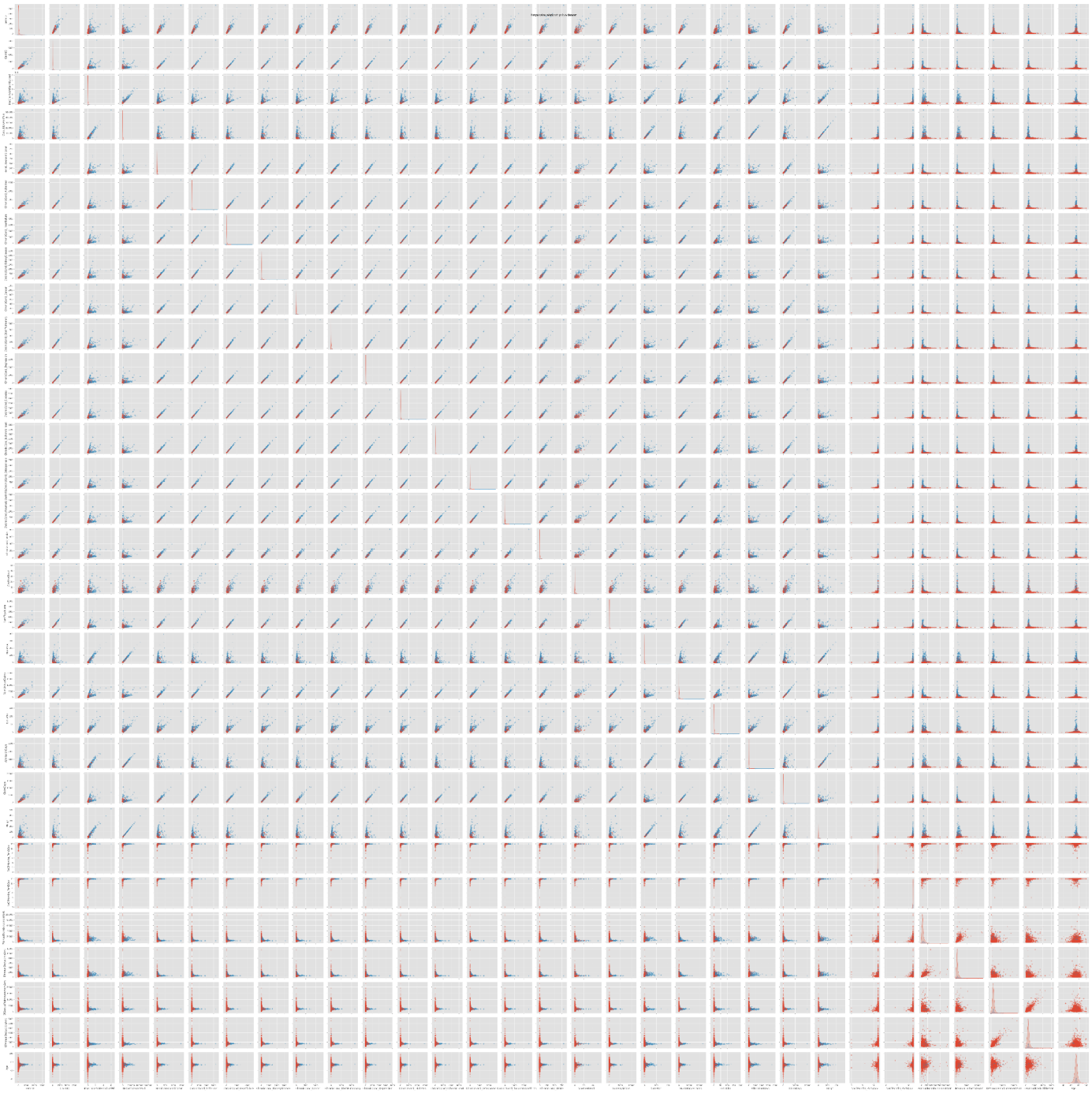
### 4.1 Fraud Distribution Among Providers

This bar chart shows the distribution of fraudulent vs. non-fraudulent providers. It clearly illustrates the **class imbalance**, where the majority of providers are non-fraudulent. Handling this imbalance was a key challenge in training machine learning models.



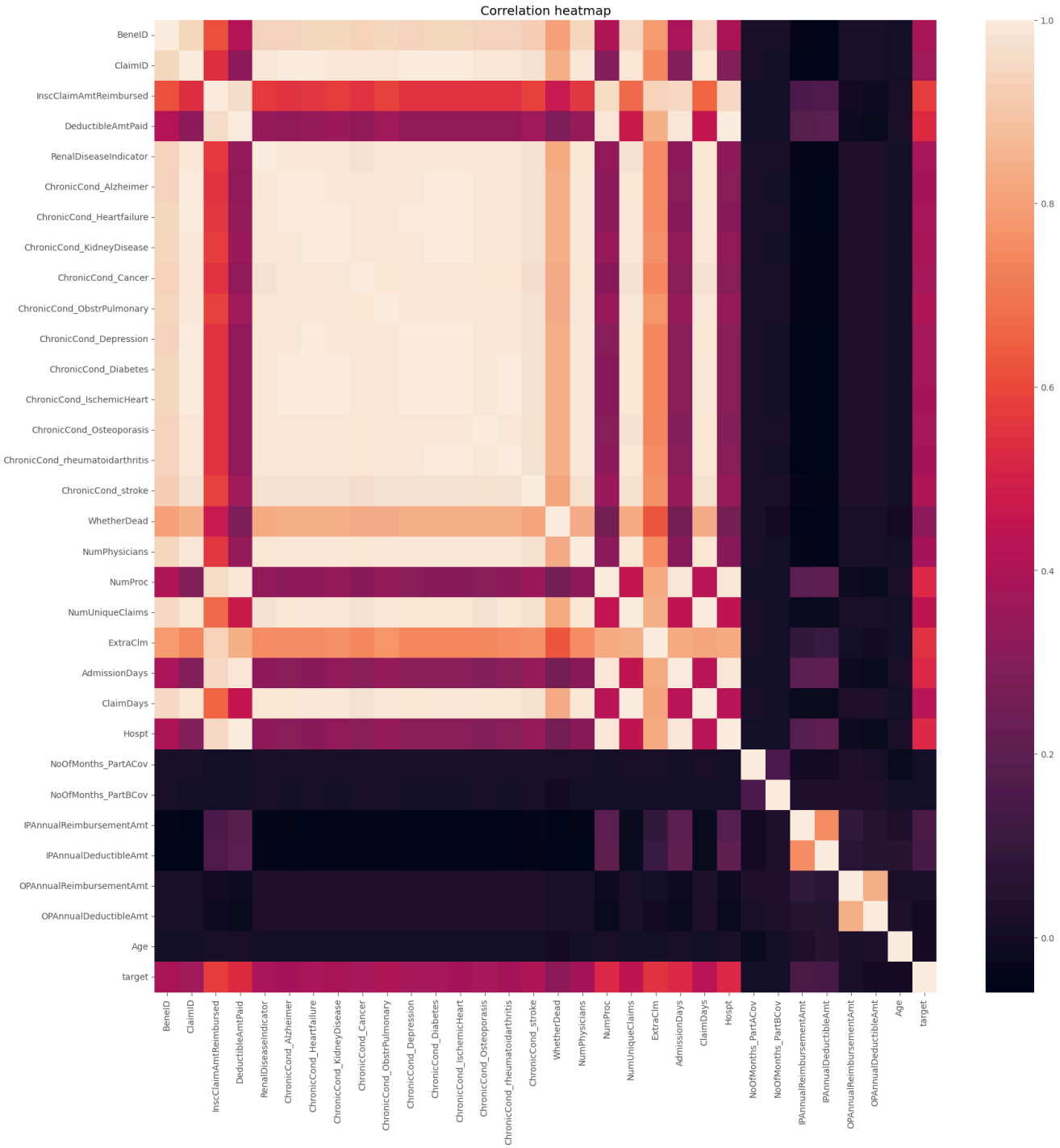
## 4.2 Pairwise Relationships of Features (Seaborn Pairplot)

We used Seaborn's pairplot to examine the relationships between different features for providers labeled as fraudulent (`PotentialFraud = 1`) and non-fraudulent (`PotentialFraud = 0`). This plot highlights key pairwise interactions between features, such as `TotalCharges`, `NumClaims`, `AdmissionDays`, etc.



## 4.3 Correlation Heatmap of Features

The correlation heatmap provides insight into how different features are correlated with one another. Features such as `ReimbursementAmounts`, `NumProc`, and `AdmissionDays` show strong correlations, which could be useful in predicting fraudulent claims.



## 4.4 Logistic Regression Results

- **Accuracy: 83.74%**
- **Precision for fraud: 34%**
- **Recall for fraud: 92%**
- **F1-score: 50%**

While Logistic Regression achieved a relatively high recall for fraud, it had a poor precision (only 34%). This indicates that while many fraudulent providers were detected, the model also falsely flagged many non-fraudulent providers.

**Confusion Matrix:**

| | **Predicted Fraud** | **Predicted Non-Fraud** |
|---|---|---|
| **Fraud** | 110 | 10 |
| **Non-Fraud** | 210 | 1023 |

## 4.5 Random Forest Classifier Results

- **Accuracy: 78.42%**
- **Precision for fraud: 28%**
- **Recall for fraud: 93%**
- **F1-score: 43%**

Random Forest improved recall to 93%, but it still had lower accuracy due to many false positives. After tuning the hyperparameters (`n_estimators = 60` and `min_samples_split = 0.25`), the model performance slightly improved.

**Confusion Matrix:**

| | **Predicted Fraud** | **Predicted Non-Fraud** |
|---|---|---|
| **Fraud** | 111 | 9 |
| **Non-Fraud** | 283 | 950 |

### 4.6 MLP Classifier Results

- **Accuracy**: **91.65%**
- **Precision for fraud**: **52%**
- **Recall for fraud**: **68%**
- **F1-score**: **59%**

The MLP Classifier achieved the best performance, balancing both precision and recall. Its higher accuracy, F1-score, and reduced false positives made it the most effective model for this dataset.

**Confusion Matrix:**

| | **Predicted Fraud** | **Predicted Non-Fraud** |
|---|---|---|
| **Fraud** | 82 | 38 |
| **Non-Fraud** | 75 | 1158 |

### ROC-AUC Scores:

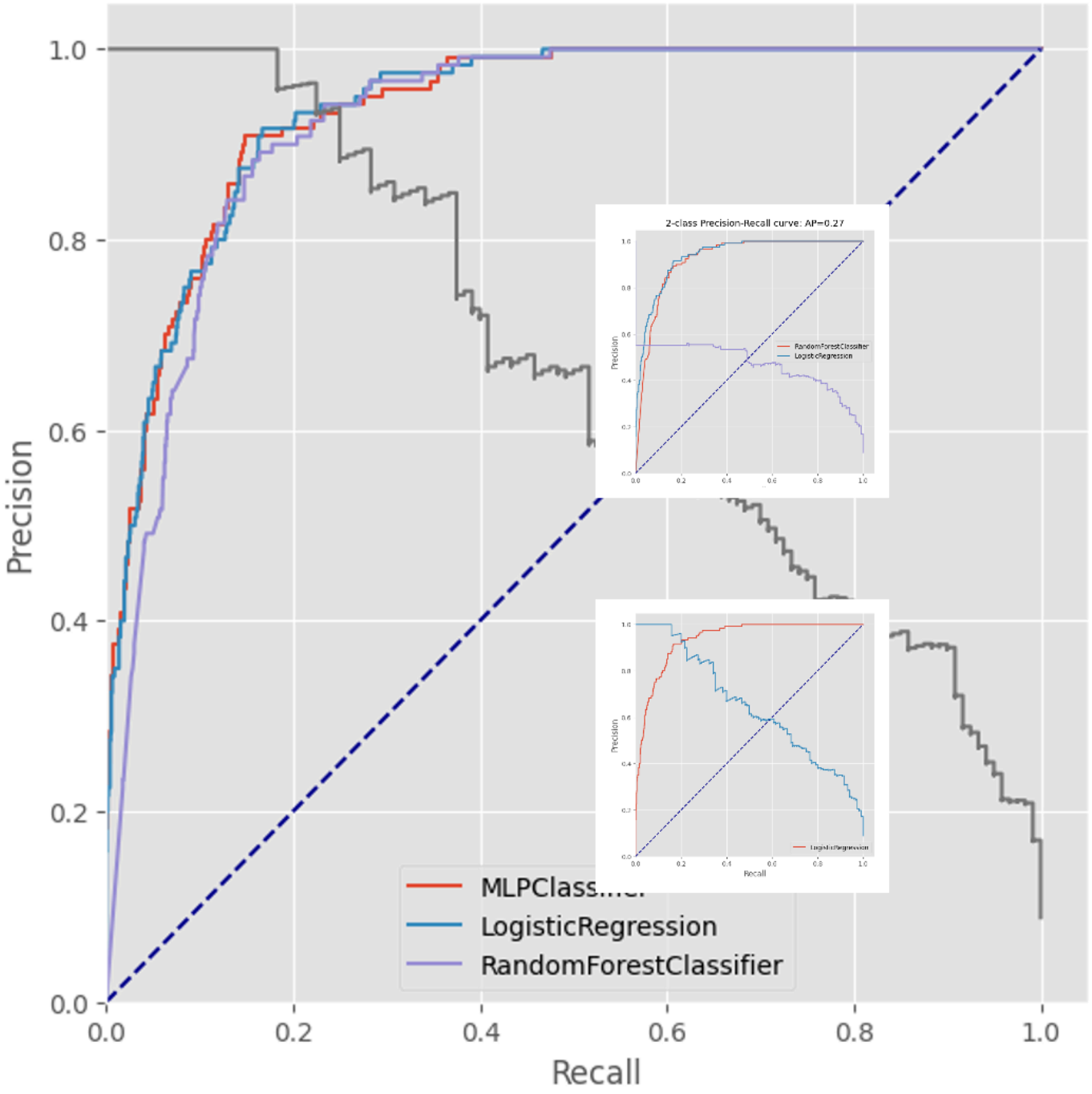
- **Logistic Regression**: 93.64%
- **Random Forest**: 92.21%
- **MLP Classifier**: 93.62%

# 5. Discussion

## 5.1 Comparison of Models

- **Logistic Regression**: Provided high recall for fraud but suffered from poor precision. This made it overly sensitive to fraud, resulting in many false positives.
- **Random Forest**: While better at identifying fraud, it also struggled with precision. However, hyperparameter tuning improved its ability to capture the class imbalance, leading to better recall.
- **MLP Classifier**: Outperformed both Logistic Regression and Random Forest, offering a balance between precision and recall. Its deep learning architecture allowed it to capture more complex patterns in the data.

## Figures: ROC Curve Comparison



# 6. Conclusion

This project explored different machine learning models to detect fraudulent healthcare providers. The **MLP Classifier** emerged as the best model, achieving an accuracy of **91.65%** and a ROC score of **93.62%**, indicating a strong capability for distinguishing between fraudulent and non-fraudulent cases. The model effectively balanced precision (52%) and recall (68%), although further improvement in precision is necessary for practical applications.