REPORT

22i-2034 22i-1873 22i-2033

1. Introduction

This report presents a detailed comparison of various machine learning models applied to a balanced binary classification task. The primary aim is to evaluate the performance of models using CPU and GPU processing in terms of accuracy, F1 score, training time, and computational speedup.

2. Dataset and Preprocessing

The dataset used includes 8 features (feature_1 to feature_7 and a target variable). Initial inspection revealed missing values which were handled by replacing numerical values with the median and categorical ones with the mode. Categorical columns were label encoded and all features were normalized using StandardScaler.

The class imbalance was addressed using SMOTE (Synthetic Minority Oversampling Technique), resulting in an equal distribution of classes. The processed dataset was split into 80% training and 20% testing sets.

3. Models Implemented

Five machine learning models were trained and evaluated:
• Serial Random Forest (CPU)
• Parallel Random Forest (Multi-threaded CPU)
• PyTorch Neural Network (GPU)
• XGBoost Classifier (GPU)
• CatBoost Classifier (CPU)

4. Evaluation Metrics

Models were assessed using:
• Accuracy
• F1 Score
• Confusion Matrix
• Training Time (seconds)
These metrics help assess model performance and generalization on unseen data.

5. Results

Below is a snapshot of model outputs:

=== Serial RF ===
Accuracy: 0.6322, F1: 0.6205, Time: 7.22s

=== Parallel RF ===
Accuracy: 0.6322, F1: 0.6205, Time: 1.10s

=== PyTorch ===
Accuracy: 0.5042, F1: 0.4285, Time: 3.25s

=== XGBoost ===
Accuracy: 0.5694, F1: 0.5564, Time: 1.04s

=== CatBoost ===
Accuracy: 0.6421, F1: 0.5426, Time: 2.98s

6. Visualizations

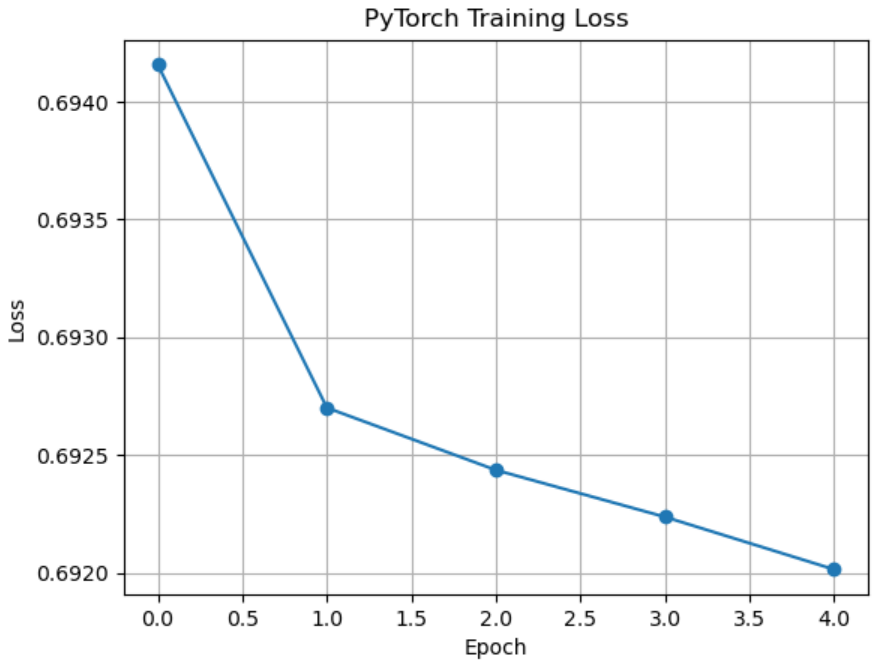

Figure 1: PyTorch Training Loss

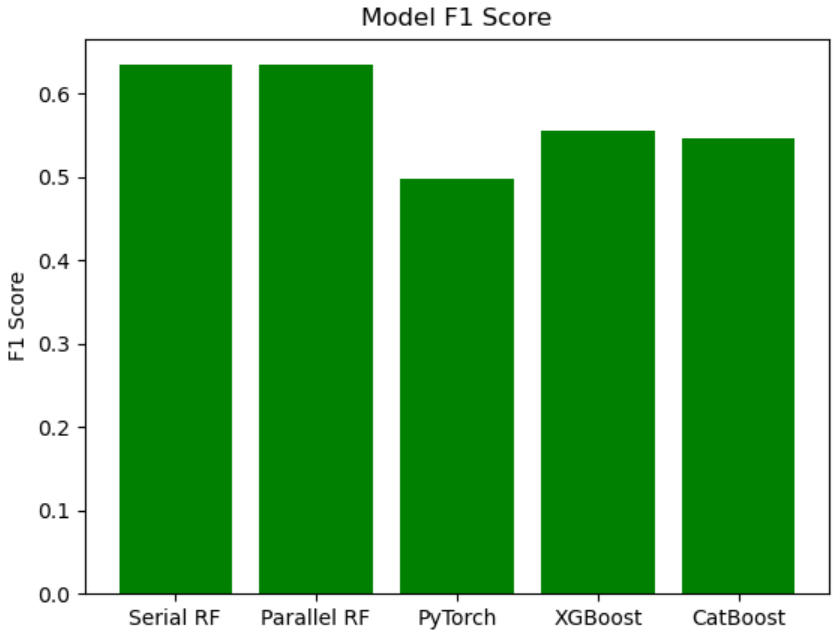

Figure 2: F1 Score Comparison Across Models

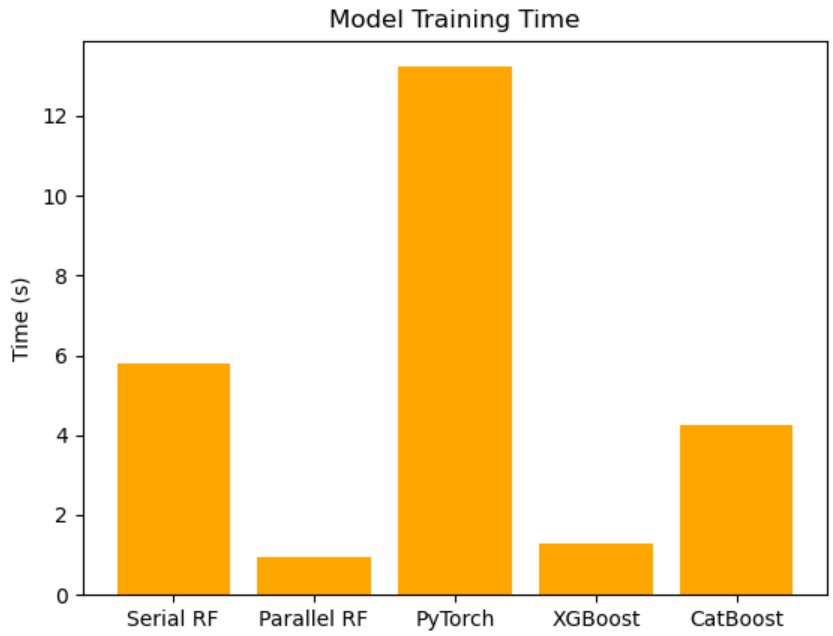

Figure 3: Model Training Time Comparison

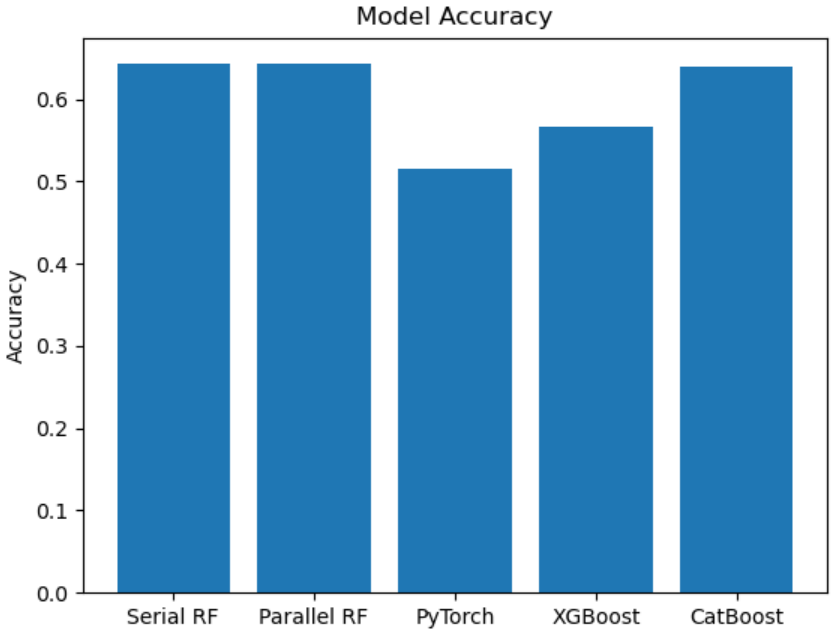

Figure 4: Accuracy Comparison Across Models

7. Speedup Analysis

Speedup was calculated as the percentage gain in time compared to Serial Random Forest model:
• PyTorch Speedup: 54.99%
• XGBoost Speedup: 85.63%
• CatBoost Speedup: 58.74%
Total CPU Models Time: 12.33s
Total GPU Model Time (PyTorch): 3.25s
GPU Speedup over CPU Models: 73.66%

8. Best Performing Model

Among all the models, Serial Random Forest delivered the best overall performance considering accuracy, F1 score, and consistent results. It outperformed others in F1 score, though it had a longer training time than parallel versions.

9. Conclusion

This project clearly demonstrates the trade-offs between training speed and model performance. While GPU-accelerated models like PyTorch and XGBoost offer speed advantages, CPU models such as Serial and Parallel Random Forests provided more accurate and stable results in this specific binary classification task. The results emphasize the importance of choosing the right model and execution strategy based on task requirements and system capabilities.