CAD (Coronary Artery Disease) using AI

Report: (PDF) Minor Project Report G9.pdf

Feature Engineering - (303,54) to (303,23)

- Clean and preprocess data (handle missing values, scaling, encoding).
- Used Min-Max scaling (No normalization required for tree based models)
- Applied 10 ML algorithms on the pre-processed dataset
- **SMOTE (Synthetic Minority Over-sampling Technique) applied if imbalance ratio < 0.6** to balance classes.
- **Synthetic samples generated** for the minority class.
- **Ensured balanced class distribution** before training models.

Results (On full dataset)

Model	Accuracy	Precision	Recall	F1 Score	AUROC	Training Time
XGBoost	0.919540	0.928571	0.906977	0.917647	0.979387	0.115706
CatBoost	0.908046	0.926829	0.883721	0.904762	0.974101	3.897274
SVM	0.908046	0.926829	0.883721	0.904762	0.956131	0.022643
Logistic Regression	0.896552	0.904762	0.883721	0.894118	0.949789	0.013163
Random Forest	0.896552	0.904762	0.883721	0.894118	0.974366	0.235903
Decision Tree	0.885057	0.851064	0.930233	0.888889	0.883192	0.012956
AdaBoost	0.873563	0.900000	0.837209	0.867470	0.958245	0.131567
LightGBM	0.862069	0.897436	0.813953	0.853659	0.969345	0.093369
KNN	0.850575	0.968750	0.720930	0.826667	0.869450	0.003061
Naïve Bayes	0.597701	0.900000	0.209302	0.339623	0.887685	0.003288

- **Stacking ensemble** used multiple base models (RF, XGBoost, LGBM, CatBoost, ExtraTrees, SVM) for diverse learning.
- Random Forest as meta-model to combine predictions from base models.
- Stratified K-Fold cross-validation (5 folds) ensures robust model performance.

Results(Stacking ensemble):

Accuracy	**0.9080**
Precision	**0.9268**
Recall	**0.8837**
F1 Score	**0.9048**
AUROC	**0.9646**
Train Time	**30.9778 seconds**

Feature Selection Methods & Equations

1. **XGBoost Feature Importance**
    - **Uses decision trees; importance is based on split frequency and gain.**
    - **Formula:**
    - $$Importance(f) = \sum_{t \in T_f} \frac{Gain(t)}{|T|}$$

2. Recursive Feature Elimination (RFE) with Random Forest

- **Iteratively removes least important features based on Gini Importance.**
- **Formula:**
- $$Gini(f) = \sum_{t \in T_f} [p_L Gini_L + p_R Gini_R - Gini_{parent}]$$

3. Mutual Information (MI) Scores

- **Measures feature-target dependency.**
- **Formula:**
- $$I(X;Y) = \sum P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

4. Hybrid Feature Selection

- **Final Score:**

$$(0.5 \times XGB\_Importance) + (0.3 \times MI\_Score)$$

Top-ranked features were selected, refining the dataset to 23 final features.

Google Colab Notebooks (Implementations):

Pre-processing Dataset (Min-Max scaling) URL:
https://colab.research.google.com/drive/1YNO92DuR6Cz0BRXBIsrmOn32FeOMJAk8?usp=sharing

Applying all 10 ML algorithms URL:
https://colab.research.google.com/drive/1aiMX3gIvA80vp374TYRldLJCtF8jBel2?usp=sharing

Feature Engineering Hybrid Approach URL:
https://colab.research.google.com/drive/1qoqem6p8KwC0D0OUtn5-qavwg4fMG-SK?usp=sharing

Ensemble (Stacking based) URL:
https://colab.research.google.com/drive/1n3LB8oiLHS1vpg6LSUA9fiIH-FA1kVFF?usp=sharing

Voting (Majority vs Weighted voting):
Voting Ensemble with Weights.ipynb

Dataset Folder Link: Minor Project 2025

Results (On reduced dataset, feature engineering done, reduced to 23 features from 54):

Model	Accuracy	Precision	Recall	F1 Score	AUROC	Train Time
AdaBoost	0.942529	0.975000	0.906977	0.939759	0.966173	0.116817
Random Forest	0.919540	0.928571	0.906977	0.917647	0.970402	0.216211
XGBoost	0.919540	0.950000	0.883721	0.915663	0.975687	0.084674
CatBoost	0.896552	0.925000	0.860465	0.891566	0.968288	2.708236
Naïve Bayes	0.896552	0.925000	0.860465	0.891566	0.933932	0.002209
Logistic	0.885057	0.883721	0.883721	0.883721	0.932347	0.014886

Regression						
LightGBM	0.885057	0.945946	0.813953	0.875000	0.972516	0.067397
SVM	0.862069	0.860465	0.860465	0.860465	0.932347	0.018304
Decision Tree	0.839080	0.891892	0.767442	0.825000	0.838266	0.009412
KNN	0.827586	0.868421	0.767442	0.814815	0.894820	0.001947

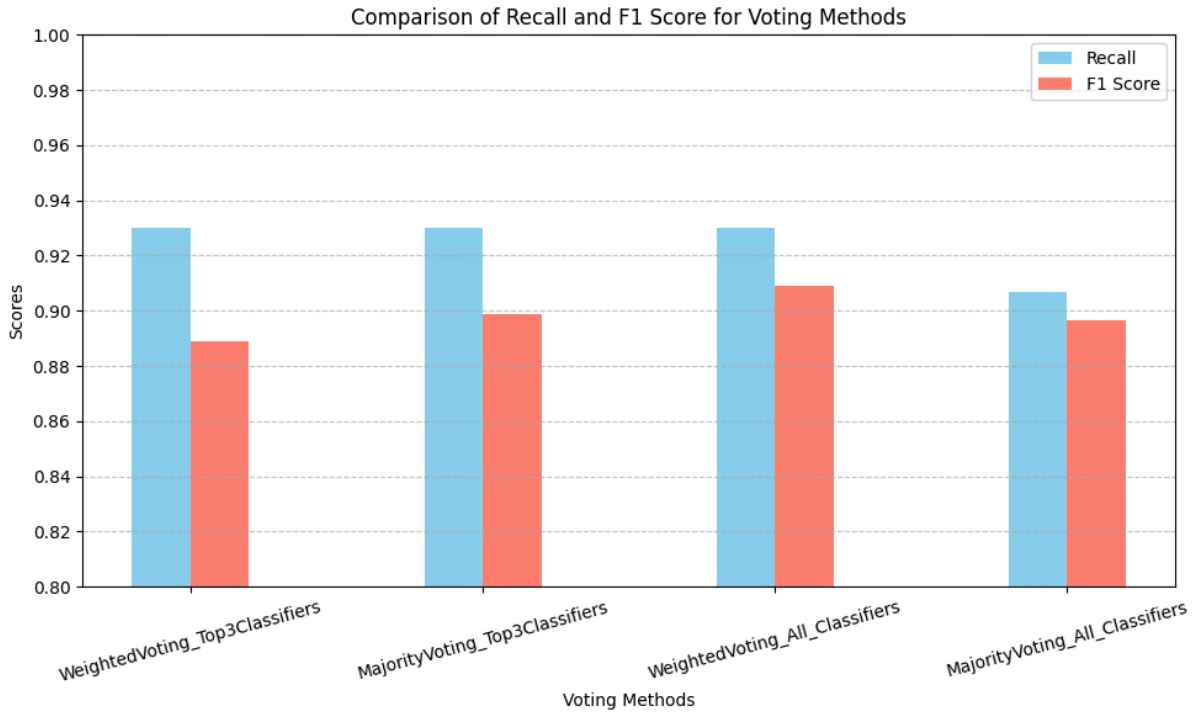
- Feature Engineering on dataset (Hybrid approach)
- Applied 10 ML algorithms
- Ensemble learning (Stacking based classifier)

Federated Learning Output: (Model used - Logistic Regression

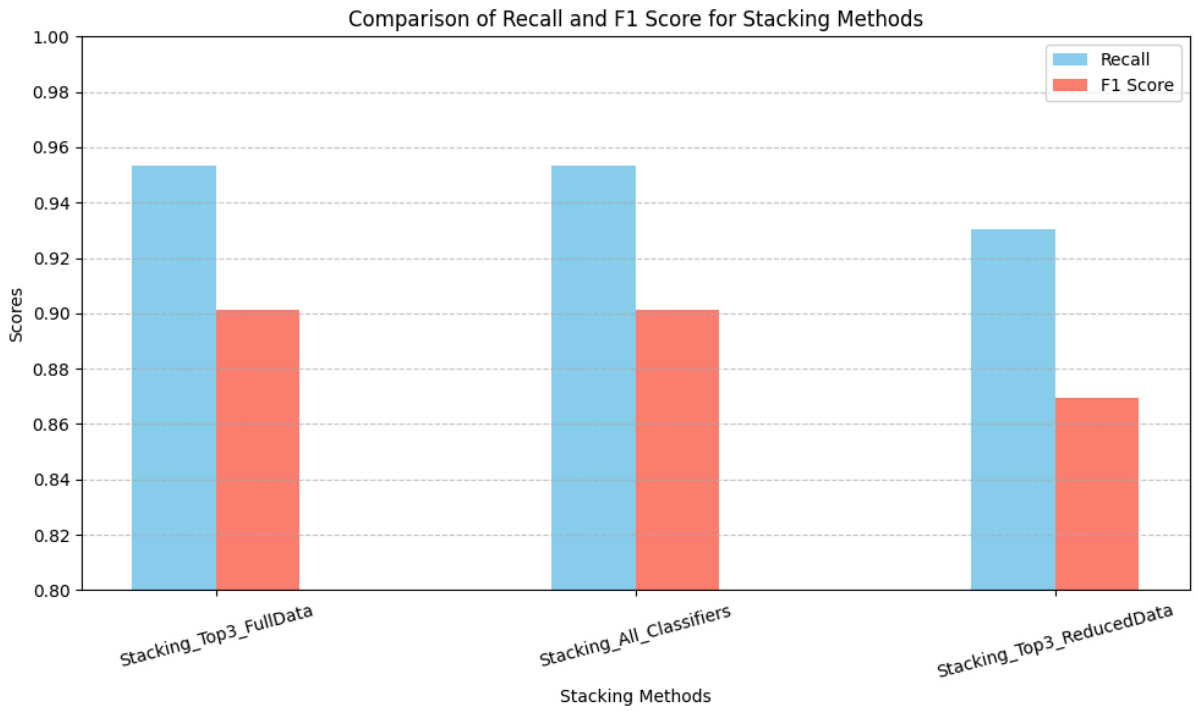
```
Global Model Evaluation
------------------------------
Clients Aggregated: 2
Clients Used: client1, client2

Final Global Weights
------------------------------
Average Coefficients:
[[ 8.28325257e-01  1.05489762e-01 -1.69968208e-01  5.23333610e-01
  -9.62412954e-02  5.59155942e-01  2.38823035e-01  5.49745240e-02
   1.44156039e-01  6.79793534e-01  3.25701795e-01  2.07700304e-02
   6.04516987e-03  1.13552748e-01 -1.05642047e-01  3.24676019e-04
  -1.57113782e-01  4.87602256e-01  2.95568303e-01  9.67766842e-02
   1.34383978e-01  1.11762628e-01 -9.87468281e-03 -2.59162247e-01
   1.32402267e+00 -5.21378838e-01  7.46066797e-02 -7.87949669e-01
  -3.79656061e-01  3.57034126e-02  2.12513148e-01  5.07027753e-01
   4.95296456e-01  6.63468088e-01  2.38772830e-01  2.49177100e-01
   5.45546497e-01  3.89469408e-01  1.01921826e+00 -3.63694409e-01
   4.99383657e-02 -6.67470996e-02  4.63380189e-01 -2.05355654e-01
   1.26367816e-01 -1.17781570e-01 -1.05052689e-01 -2.22653374e-01
   2.02400394e-01 -6.67185702e-02 -3.26066416e-01  1.07239849e+00
  -7.81219934e-01]]

Average Intercept:
[1.14182847]

Evaluation on Full Dataset
------------------------------
Accuracy: 0.8125
```

Ensemble Learning Visualizations

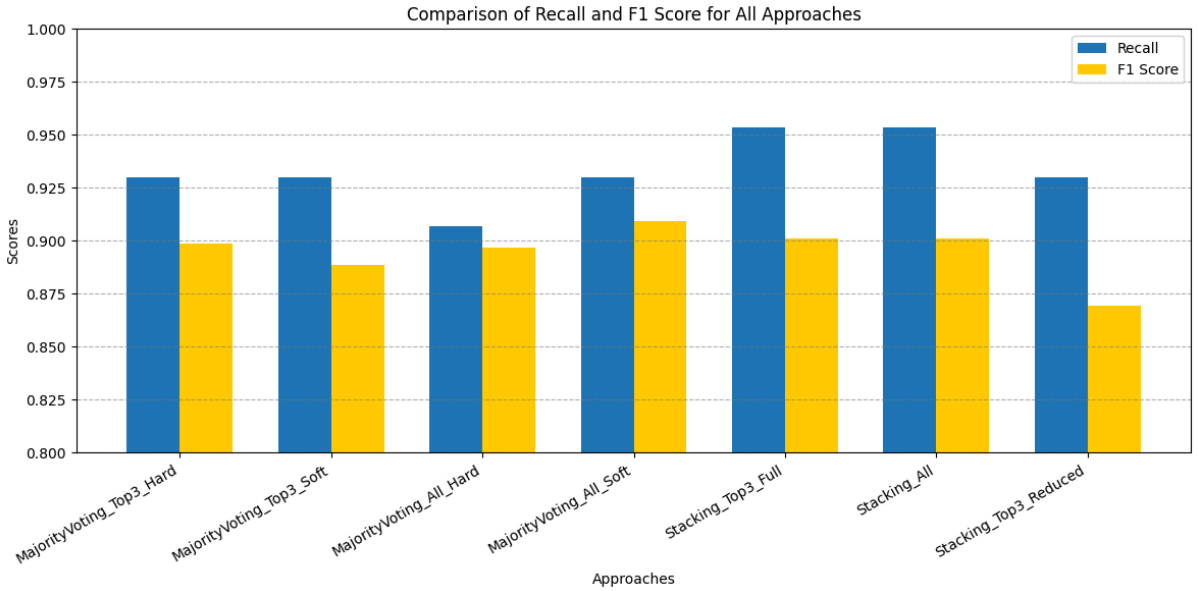
1. Voting (Weighted voting vs Majority voting)



2. Stacking (All vs Top3)



Combined comparison (Voting + Stacking):



Report for All Classifiers (Weighted Voting vs Majority Voting)

1. Report for All Classifiers

Metric	Weighted Voting (F1 Score) on all Classifiers	Majority Voting on all Classifiers
Accuracy	0.8689	0.8525
Precision	0.8889	0.8864
Recall	0.9302	0.9070
F1 Score	0.9091	0.8966

2. Report for Top 3 Classifiers

Metric	Weighted Voting (F1 Score) on Top 3 Classifiers	Majority Voting on Top 3 Classifiers
Accuracy	0.8361	0.8524
Precision	0.8511	0.8695
Recall	0.9302	0.9302
F1 Score	0.8889	0.8988

Next steps (To differentiate our project):

1. **Multi-modal AI**: To use images for detection of CAD
2. **Federated learning** implementation using 4,5 systems