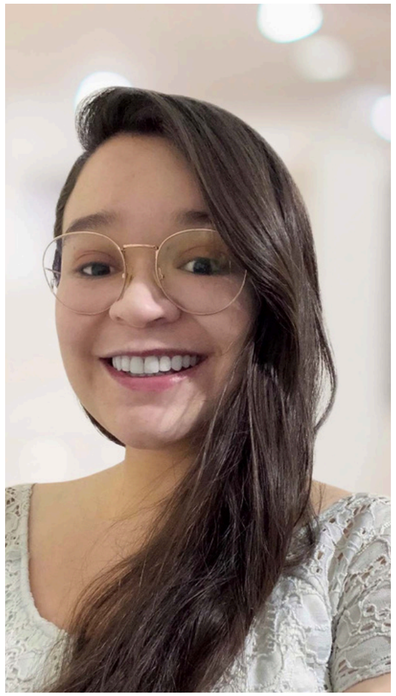
Machine Learning Techniques for Alzheimer's Detection Using Handwriting Data

A Hybrid Feature Selection and Ensemble Learning Approach

Meet Our Team

CAMILO ESTRADA



STEPHANIA NINO

ALICE NGUYEN

PIERO VERA

AGENDA

01

Introduction and Dataset

02

EDA

03

Methods

04

Implementation and performance

05

Conclusions

Introduction and Dataset

Introduction

Project Overview

- **Goal:** Improve Alzheimer's detection using handwriting data with advanced ML techniques.
- **Approach:** Hybrid feature selection and ensemble learning.

Objectives

- **Improve Feature Selection:** Use a hybrid Stacked-genetic feature selection method for selecting a combination feature (filter, wrapper) based on the Ensemble approach.
- **Implement Ensemble Learning:** Apply multiple ensemble learning techniques, to improve classification performance.
- **Evaluate Model Performance:** Compare the performance of the proposed models using standard metrics.
- **Benchmark Against Existing Methods:** Benchmark the proposed models against existing methods to demonstrate improvements.

The DARWIN Dataset

Among the columns, we found the structure of the features, as follows:

Column 1	Column 2	...	Column 18	Column 19	Column 20	...
Feature 1 (Task 1)	Feature 2 (Task 1)	...	Feature 18 (Task 1)	Feature 1 (Task 2)	Feature 2 (Task 2)	...

From where, each 18 set of columns represents the numeric values for one handwriting task, resulting in 25 performed tasks (450 numeric features).

The DARWIN Dataset

- Purpose: Diagnose Alzheimer's disease through handwriting analysis.
- Participants: 174 individuals (Alzheimer’s patients and healthy controls). Balanced on age, level of education, type of work (manual or intellectual), and gender.
- Data: Handwriting samples from various tasks.

Key Features

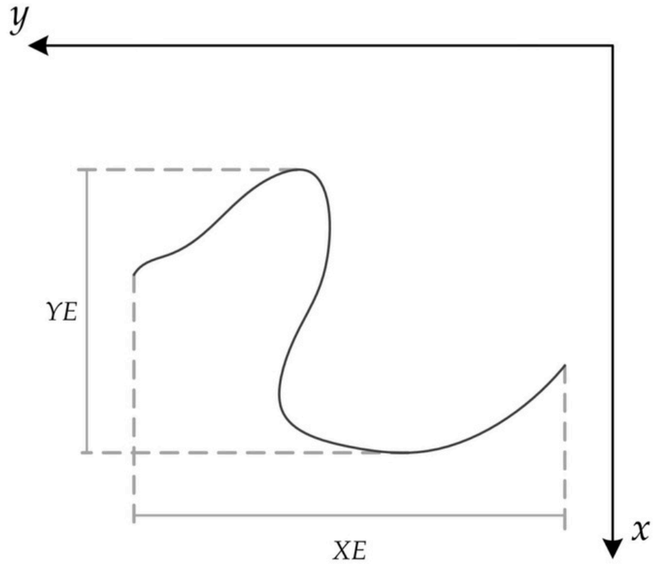
- Pen Pressure as Pressure Mean (PM): Indicates fine motor control issues.
- Speed as Mean Speed on Paper (MSP): Varies between healthy individuals and those with Alzheimer's.
- Pendowns number (PWN): uninterrupted line present 1 as value.

Categories

- Graphic tasks: joining some points and drawing geometrical figures;
- Copy tasks: repeating letters, words, and numbers;
- Memory tasks: tasted the changes against memorized objects
- Dictation tasks: variation when the working memory is used.

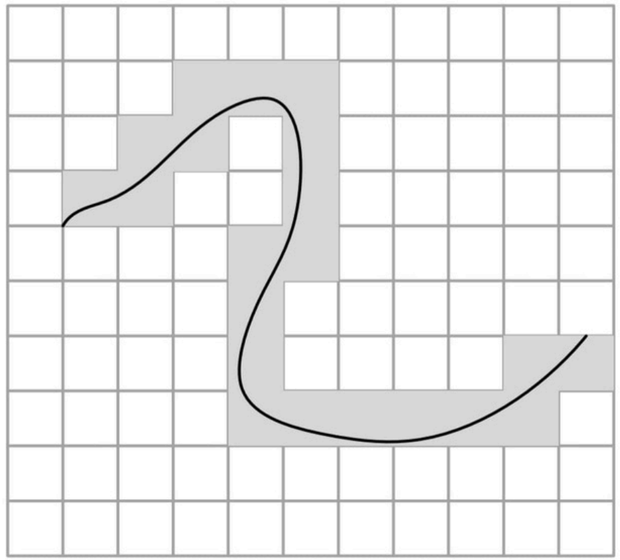
Data collection

Source: Amazon Oficial Wacom website



Max Y Extension feature

Max X Extension feature



Dispersion Index feature

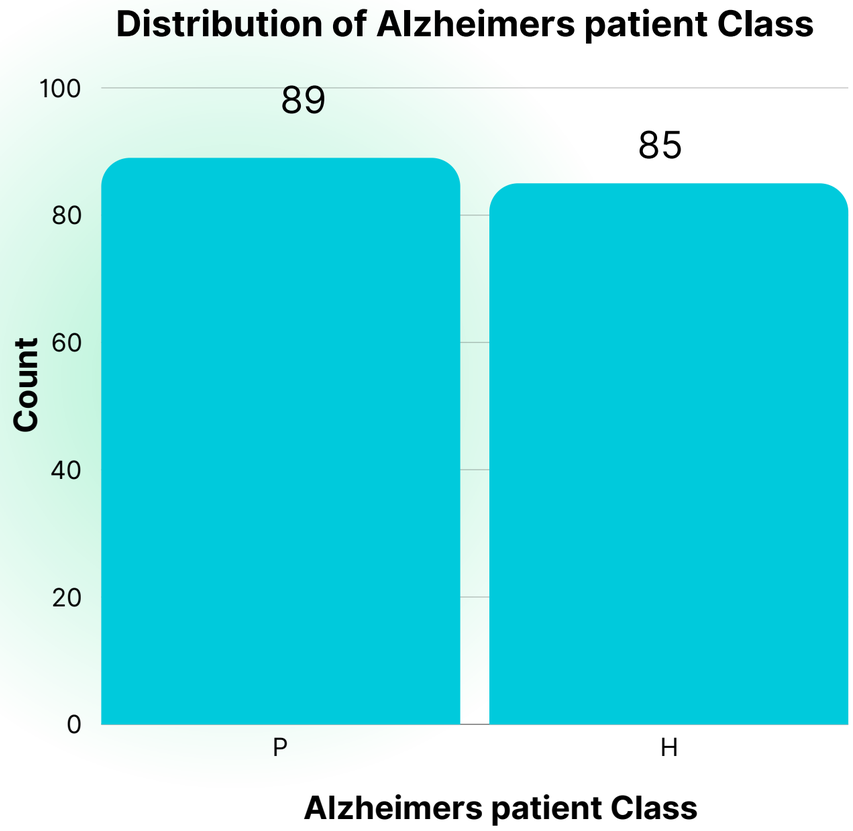
21 squares / 100 squares

= 0.21

EDA

EDA

Target Variable



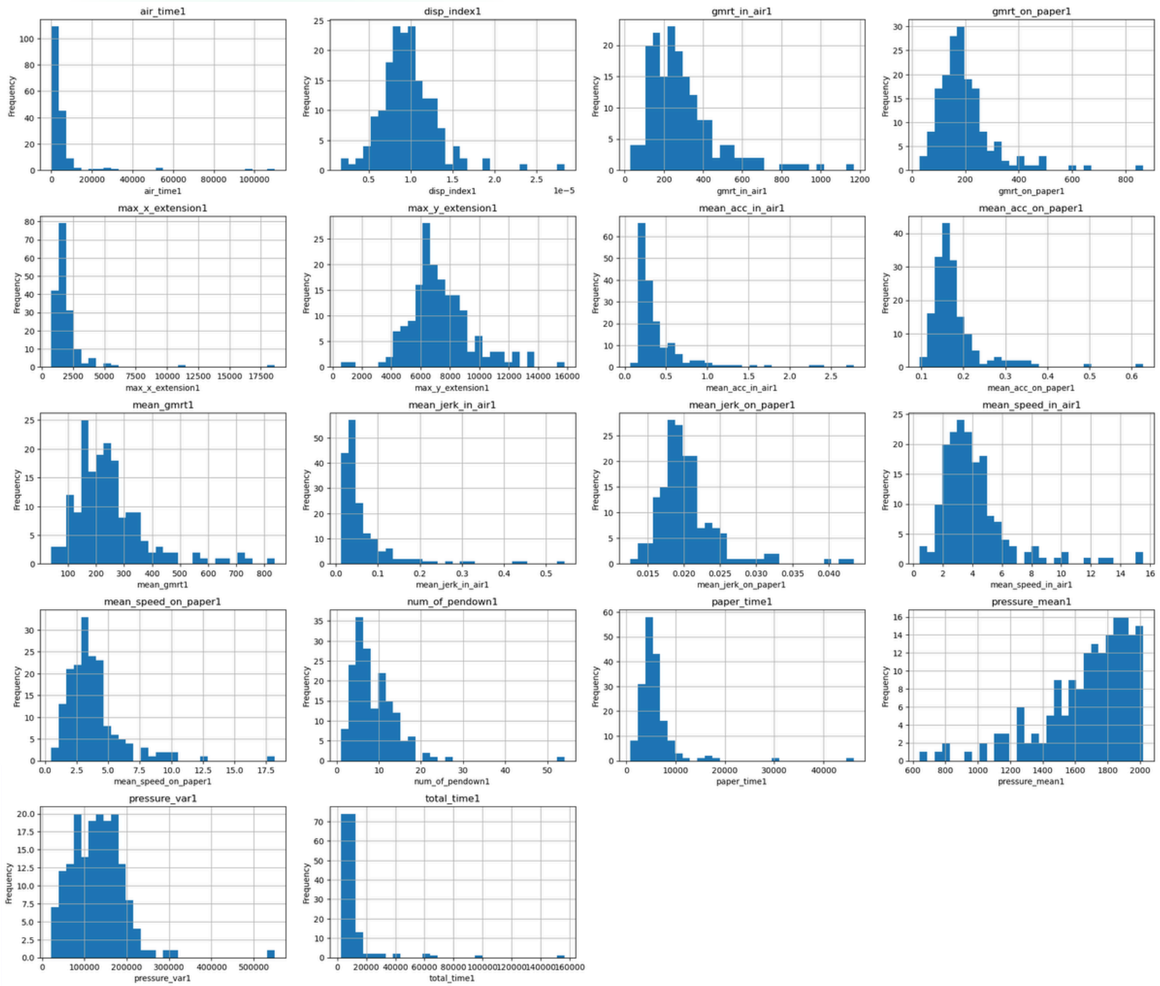
We faced a balanced dataset by the target variable, where:

- **P: Stands for "Patients",** referring to individuals diagnosed with Alzheimer's Disease.
- **H: Stands for "Healthy"**, referring to individuals who are not diagnosed with Alzheimer's Disease and serve as a control group.

EDA

Numerical Variables

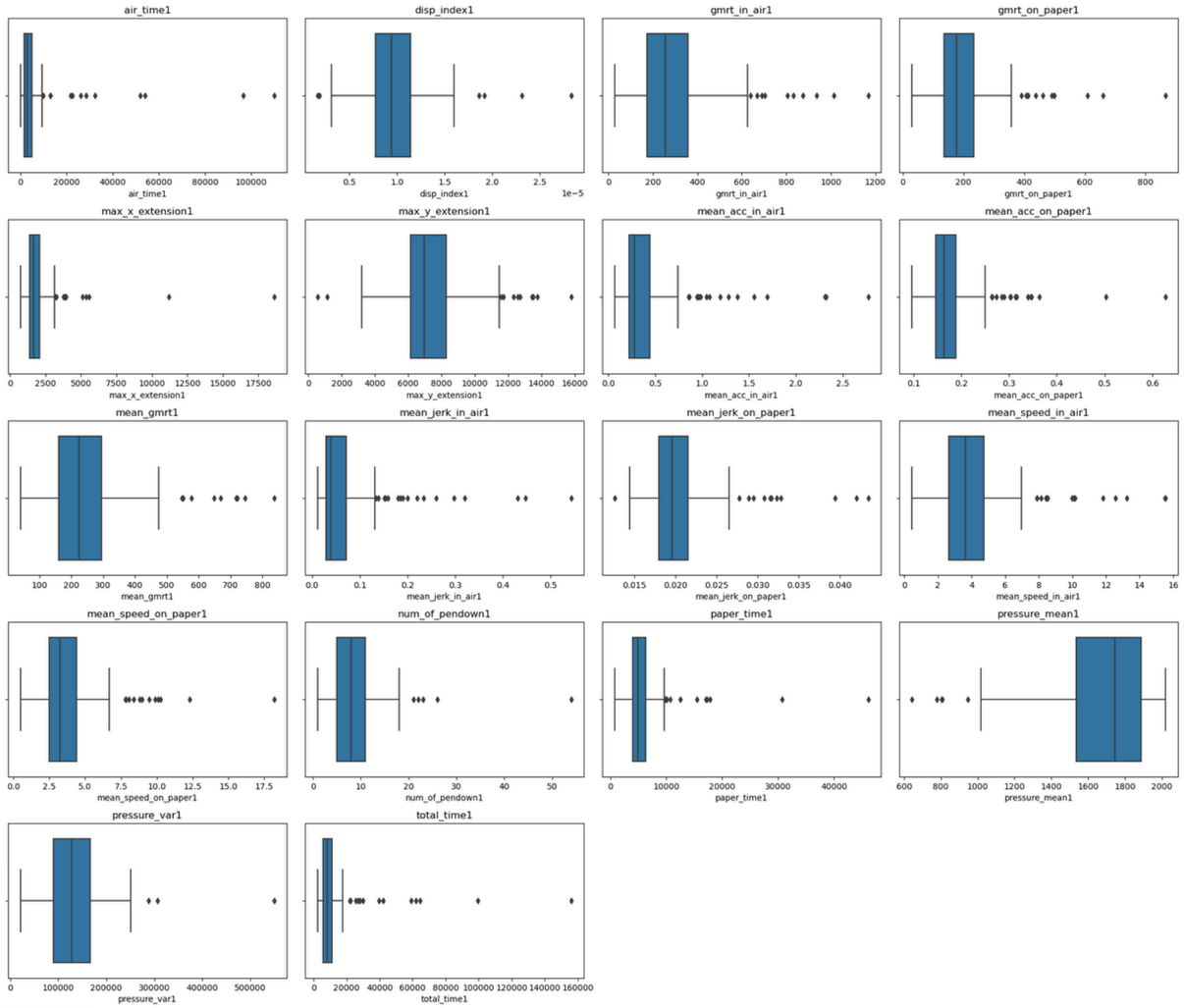
The dataset includes 25 tasks, each with 18 features.
EDA performed on Task 1 for illustration purposes.



- **Right-Skewed Distribution:** Most features show a right-skewed distribution, with lower values for air time, tremor, acceleration, and speed.
- **Variability:** Noticeable variability in maximum extensions, number of pendowns, and pressure variance, indicating diverse writing styles among participants.

EDA

Outlier Analysis

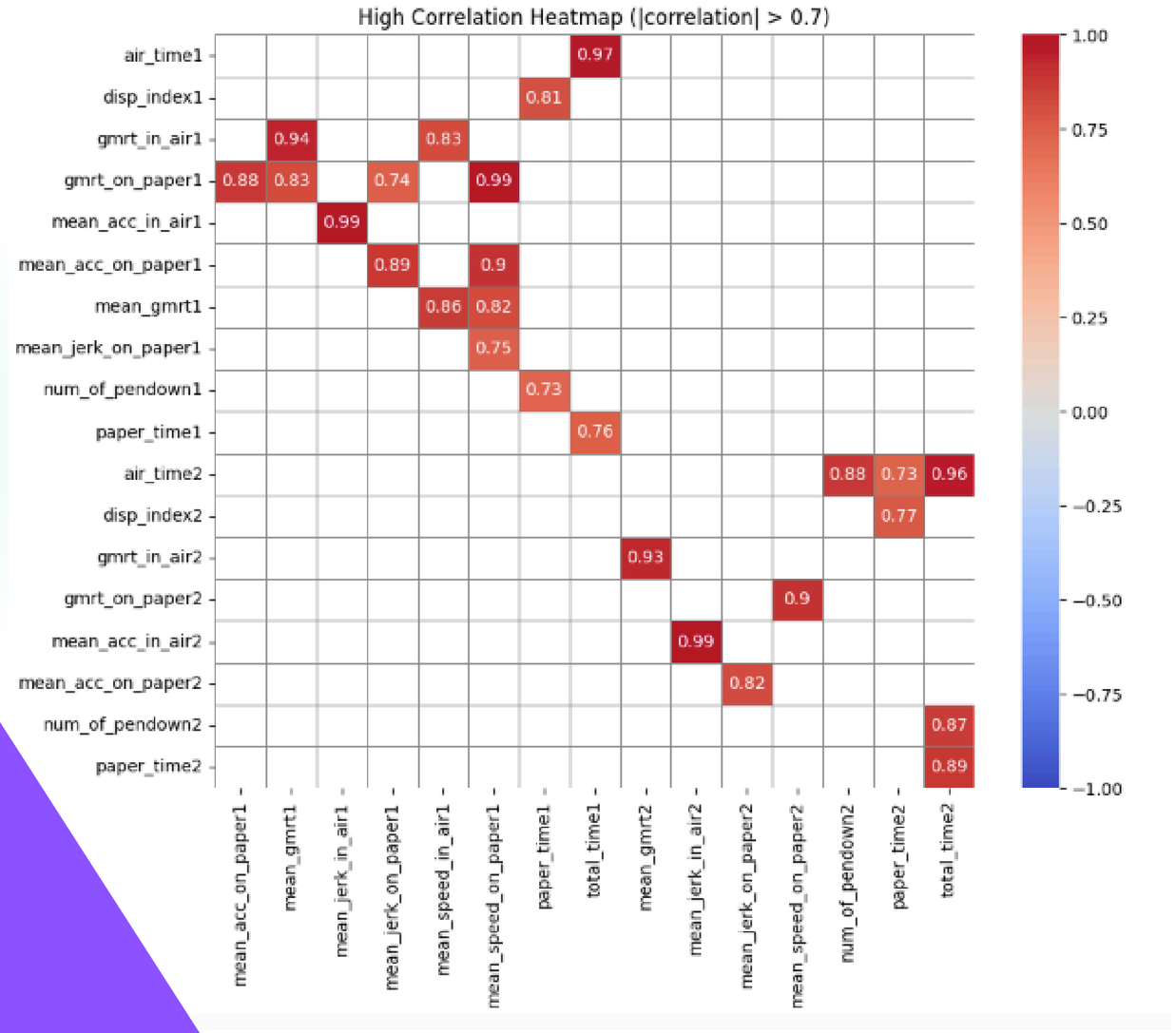


Many features show significant outliers.

This highlights the individual differences in writing behavior, which can be crucial for distinguishing between Alzheimer's patients and healthy controls.

EDA

Correlation Analysis



- Strong correlations between several features within each task only.
- None of the 18 features in Task 1 are highly correlated with the features in Tasks 2, 3, ... or 25.
- High correlation indicates redundancy, which can be addressed by feature selection methods.
- We validated the high correlation scores performing VIF whereas many coefficients were on the scale of millions.

Methods

Methods

Dimension reduction

- PCA

Feature selection

- Stepwise, Backward and Forward
- Filter & Genetic Algorithm (GA)
- A multi-task learning approach

Classifiers

- Support Vector Classifier (SVC)
- Gaussian Naive Bayes
- Decision Tree Classifier
- Multi-Layer Perceptron Classifier (MLPClassifier)
- KNeighborsClassifier
- RandomForestClassifier
- Logistic Regression
- AdaBoost (Meta-learner)

Data processing

- Remove the column ID.
- The target variable "class" with

  'H' representing **Healthy person** -> 0

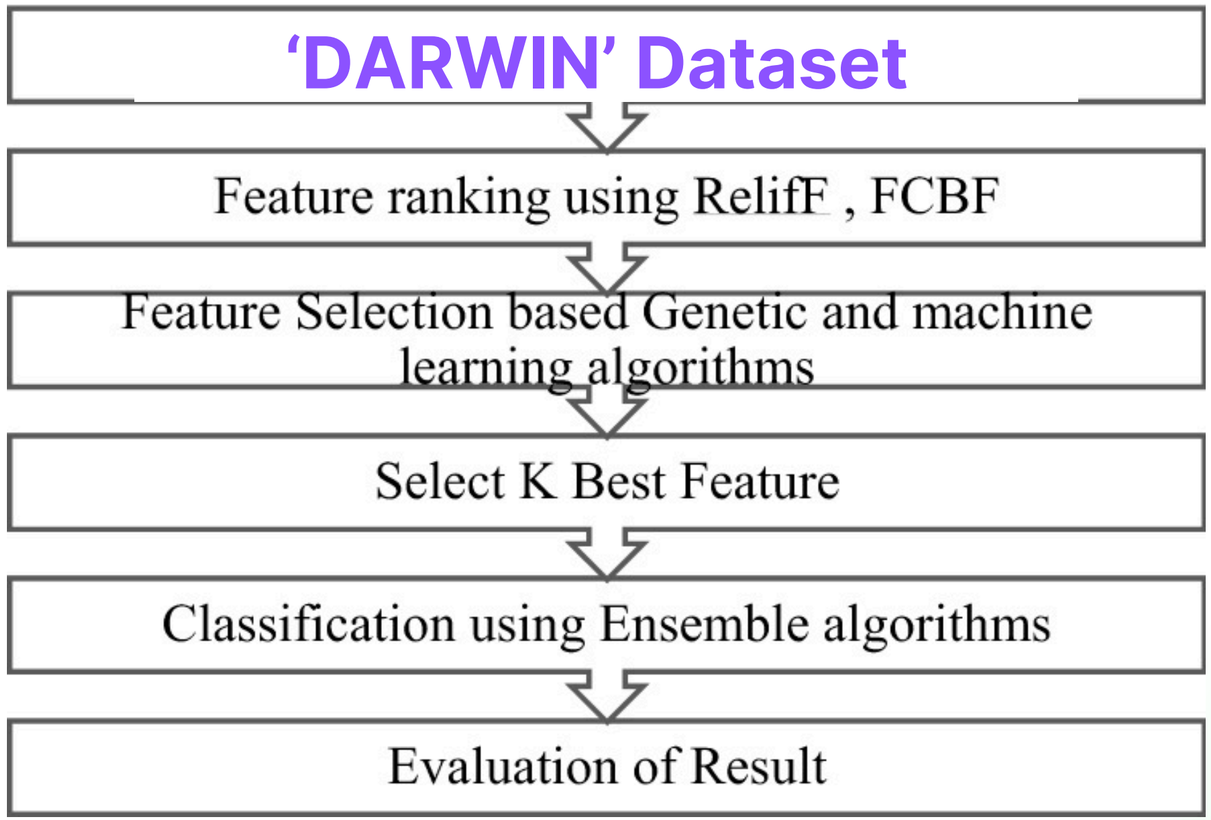
  'P' representing **Patient** -> 1
- Scale the dataset
- Split into training and test sets with a ratio of **80:20**

Filter & Genetic Algorithm (GA)

Goal:

- Reduce and eliminate low-value features with the filter algorithm
- Select high-value features with the genetic algorithm

Article: A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation



Filter Method

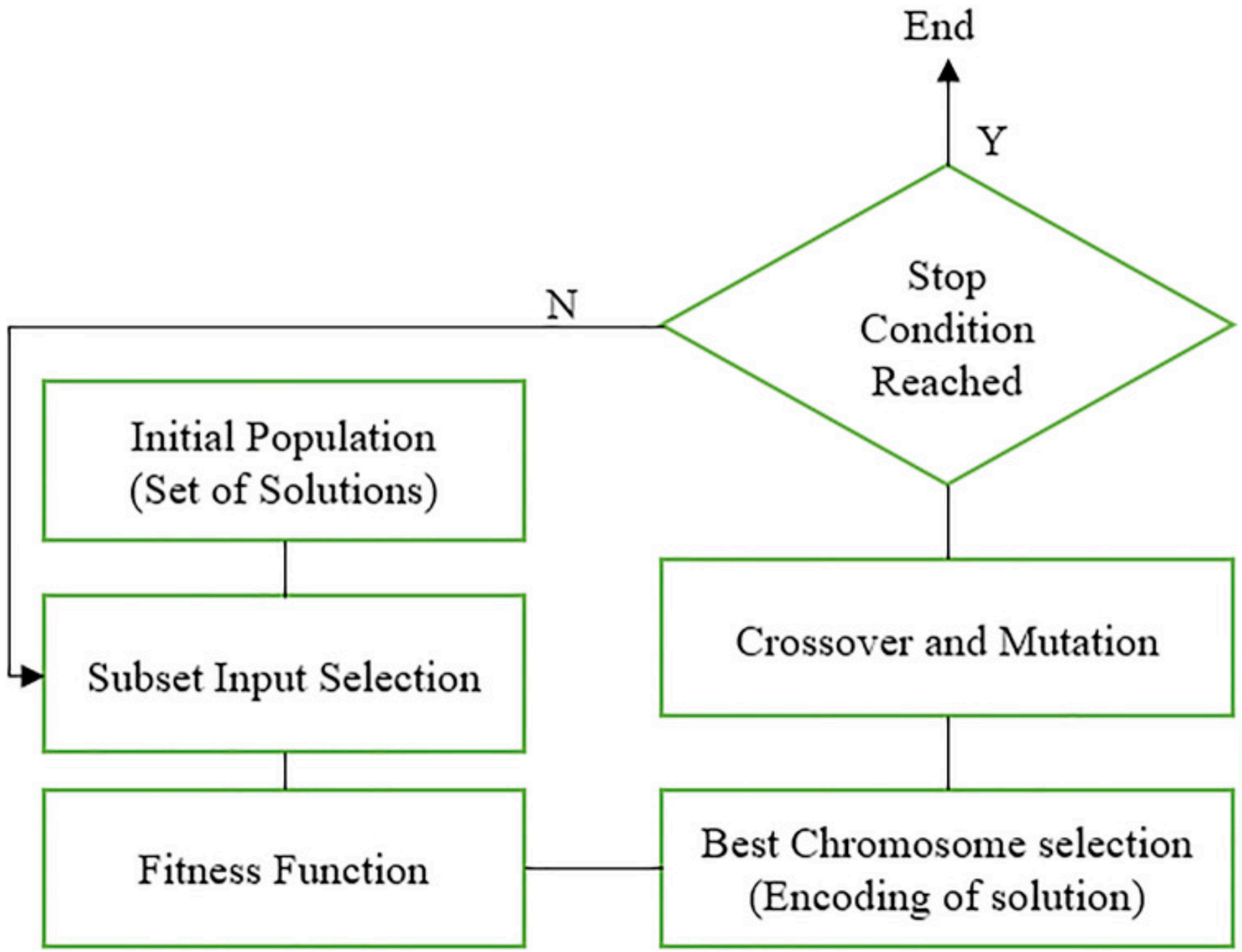
Ranking features to identify the most relevant and non-redundant ones.

ReliefF		FCBF
Evaluates feature importance based on differences between nearest neighbor instances.	∩	Ranks features by consistency and redundancy.

- Select the top 30% of features from both ReliefF and FCBF.
- Ensured only the most relevant were retained for further analysis, improving the efficiency and effectiveness of subsequent modeling steps.

Genetic Algorithm (GA)

Overview of genetic algorithm routing with Iterative Steps



A multi-task learning approach

Article: Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking

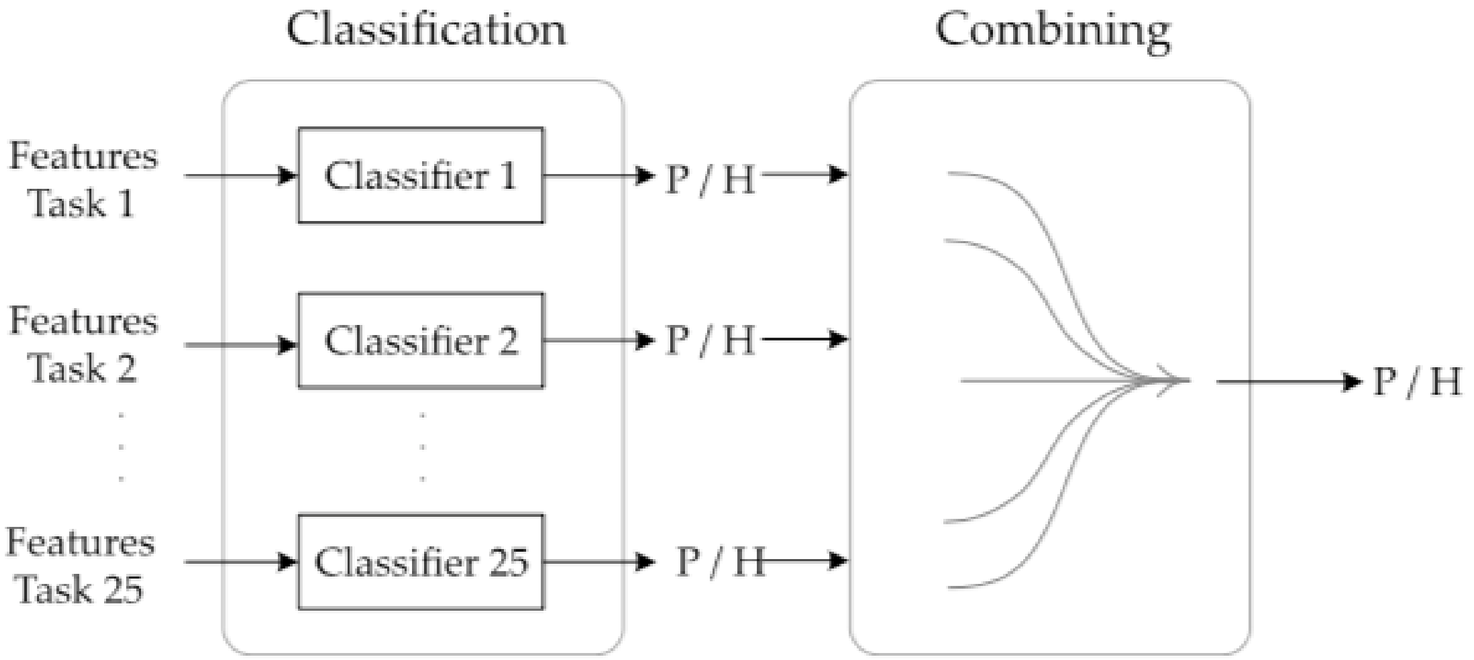


Fig. 6. Combined classifications for each handwriting task.

Implementation and performance

PCA

Dataset	Number of features
Original	450
PCs	139

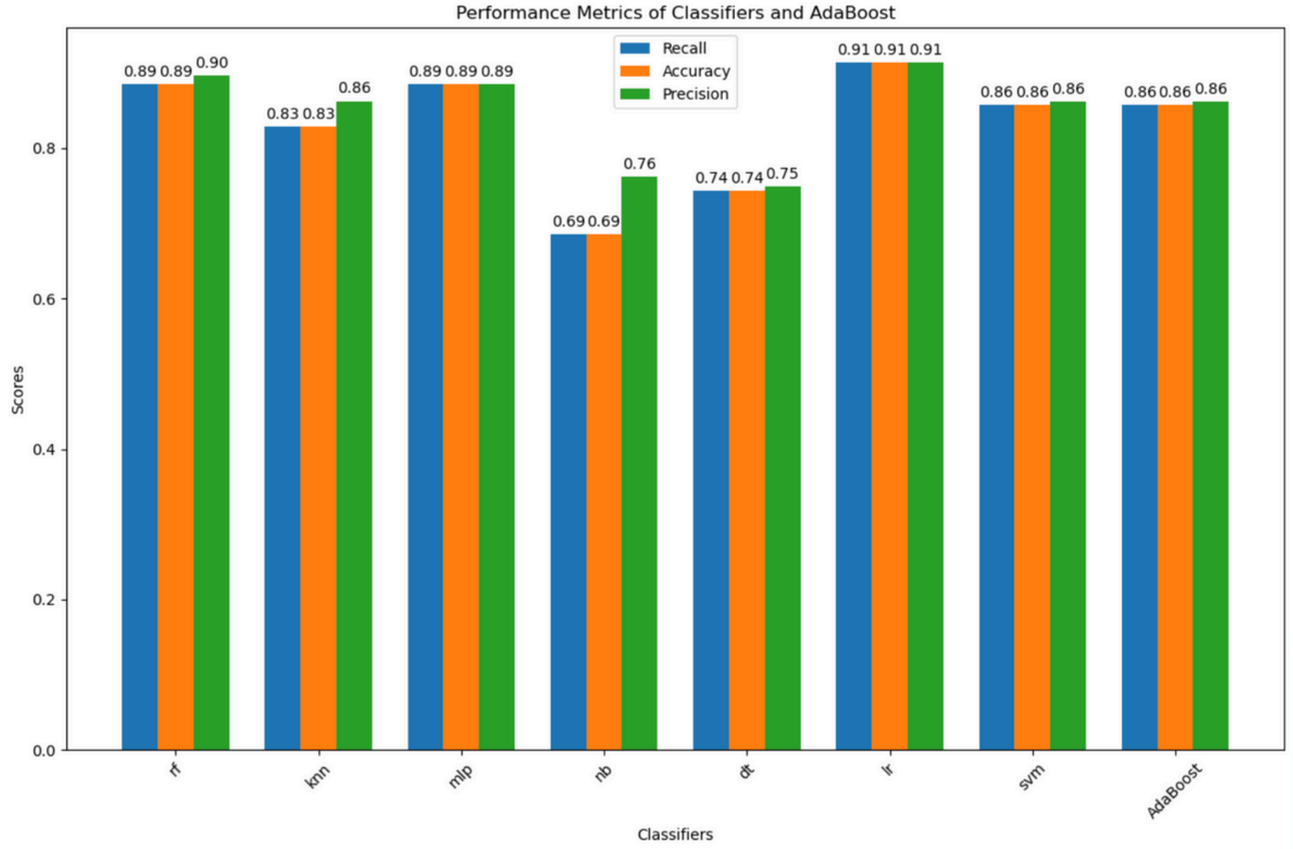
Covering 76-86 % of variance

- ☑ Followed the Kaiser criteria (eigen value greaters than 1).
- ☑ For each group of features corresponding to one determined task, we reduced its feature dimension.

PCA - Ensemble Learning and Performance Metrics

Ensemble Learning Techniques:

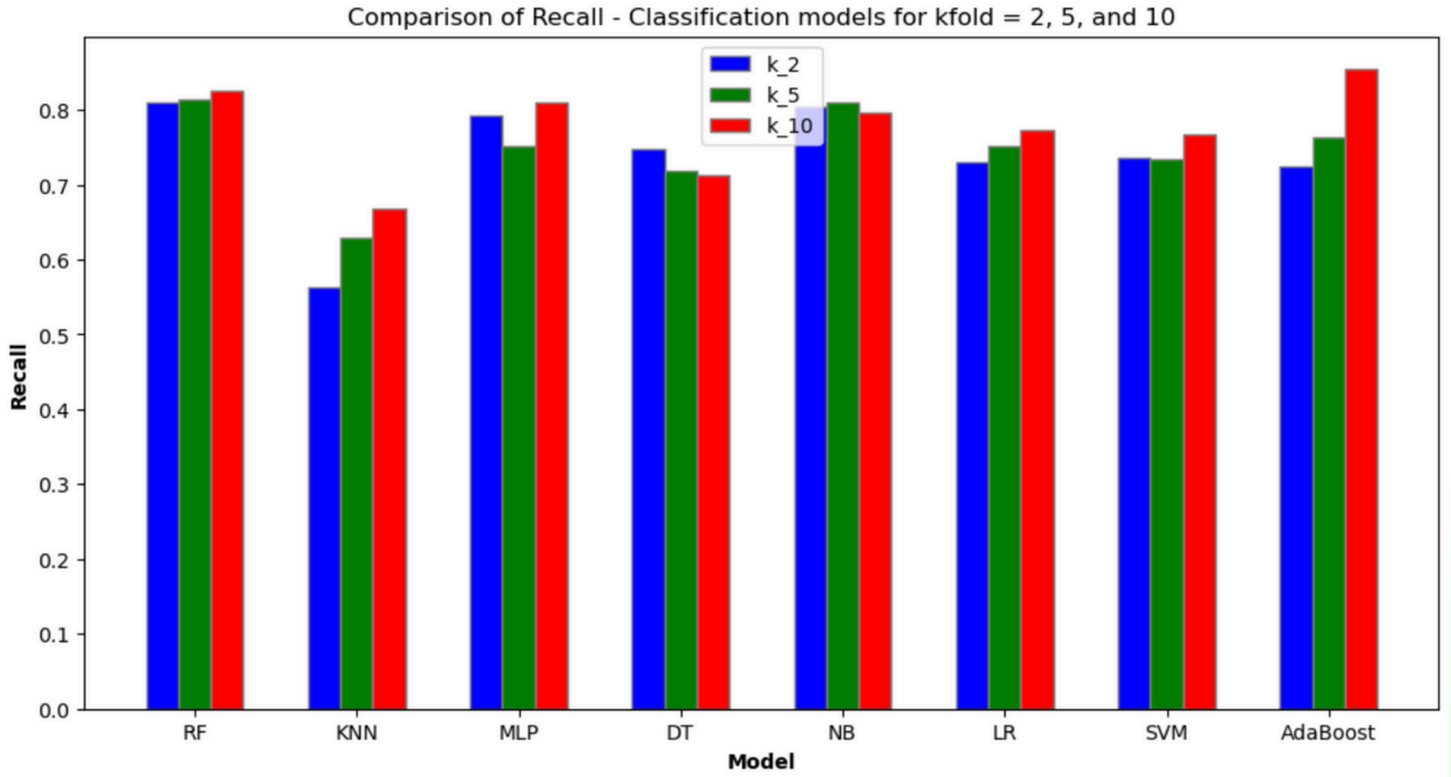
- Basic Learners: SVM, Naive Bayes, Decision Tree, MLP, KNN, RFC, LR
- Meta-Learner: AdaBoost



PCA - Cross-Validation

Cross-Validation Approach:

- Importance of Cross-Validation
- Results for k=2, 5, 10
- Table of Cross-Validation Results

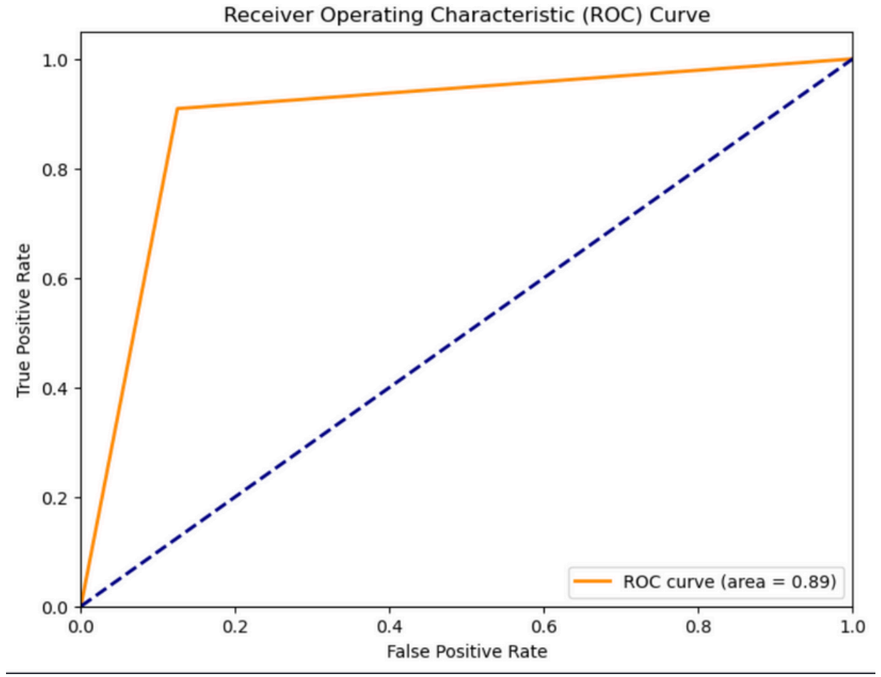


PCA - Best clasifier: Random Forest

Validating model performance using cross-validation, it highlights Random Forest as the best classifier based on **performance metrics** and **cross-validation stability**.

- Final Model Accuracy: 0.89
- Confusion Matrix

	Predicted Patient (P)	Predicted Healthy (H)
True Patient (P)	10	1
True Healthy (H)	3	21



Variable Selection

Stepwise

Method	Number of Features
Forward Selection	79
Backward Elimination	450
Stepwise Selection	**9**

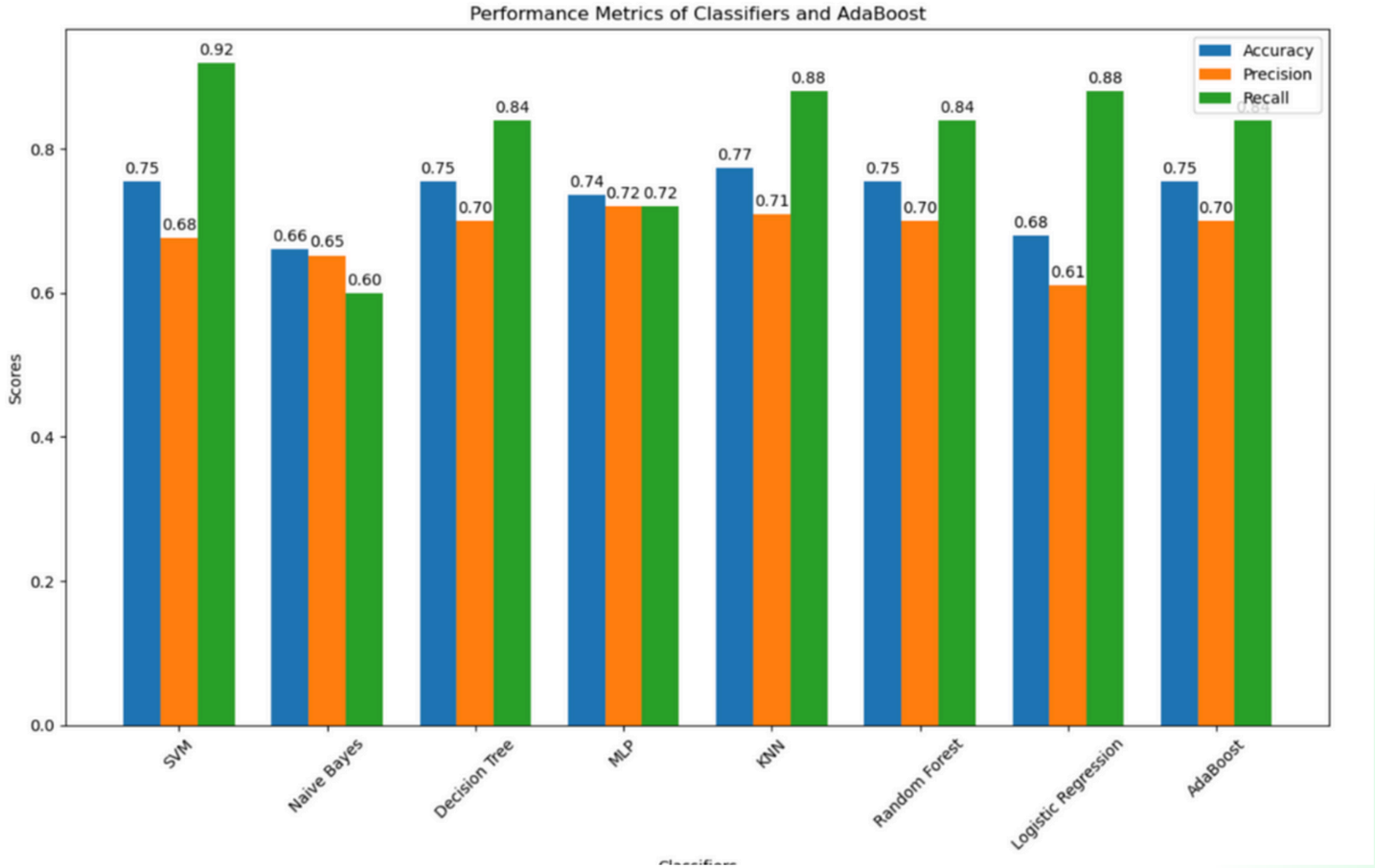
Stepwise Selected Features:

- total_time11, max_x_extension22, gmrt_in_air23, mean_acc_on_paper9, num_of_pendown19, pressure_mean12, total_time17, max_y_extension23, total_time23

Stepwise - Ensemble Learning and Performance Metrics

Ensemble Learning Techniques:

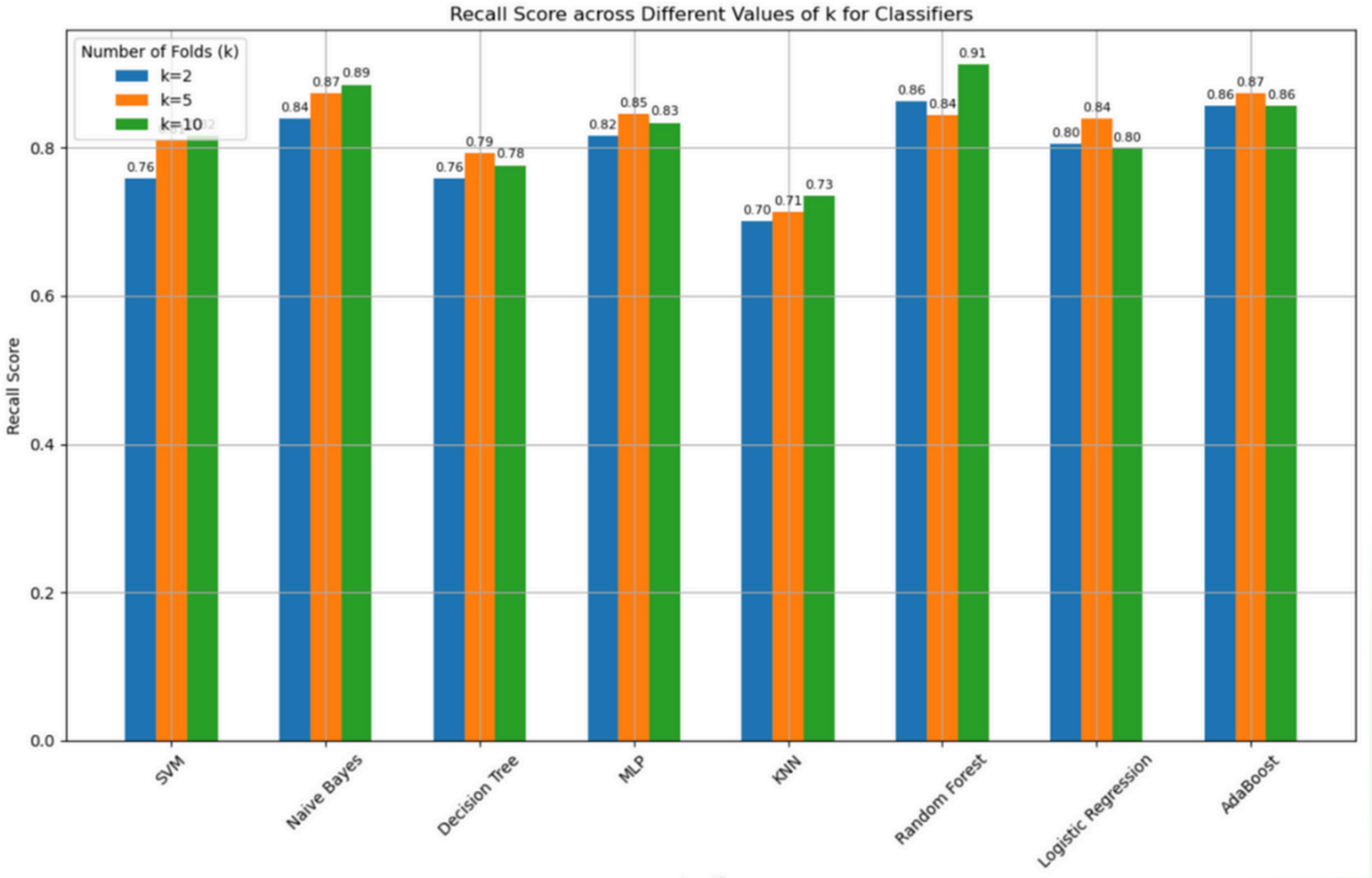
- Basic Learners: SVM, Naive Bayes, Decision Tree, MLP, KNN, RFC, LR
- Meta-Learner: AdaBoost



Stepwise - Cross-Validation

Cross-Validation Approach:

- Importance of Cross-Validation
- Results for k=2, 5, 10
- Table of Cross-Validation Results

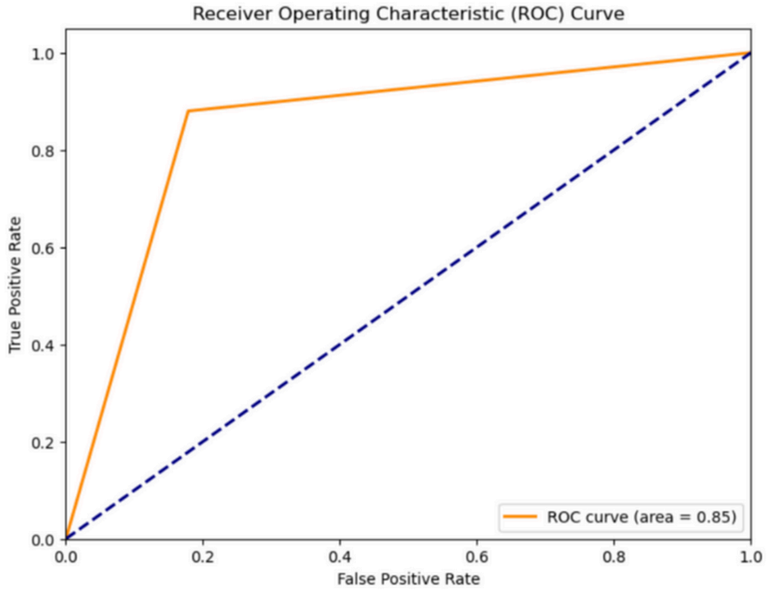


Stepwise - Best clasifier: Random Forest

Validating model performance using cross-validation, it highlights Random Forest as the best classifier based on **performance metrics** and **cross-validation stability**.

- Final Model Accuracy: 0.75
- Confusion Matrix

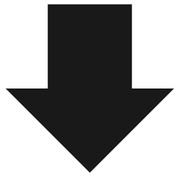
	Predicted Patient (P)	Predicted Healthy (H)
True Patient (P)	23	5
True Healthy (H)	3	22



Filter & GA Performace

Filter method (ReliefF ranking)

450 features



50 features

GA - Classifier Performance and Fitness Scores

GA Parameters:

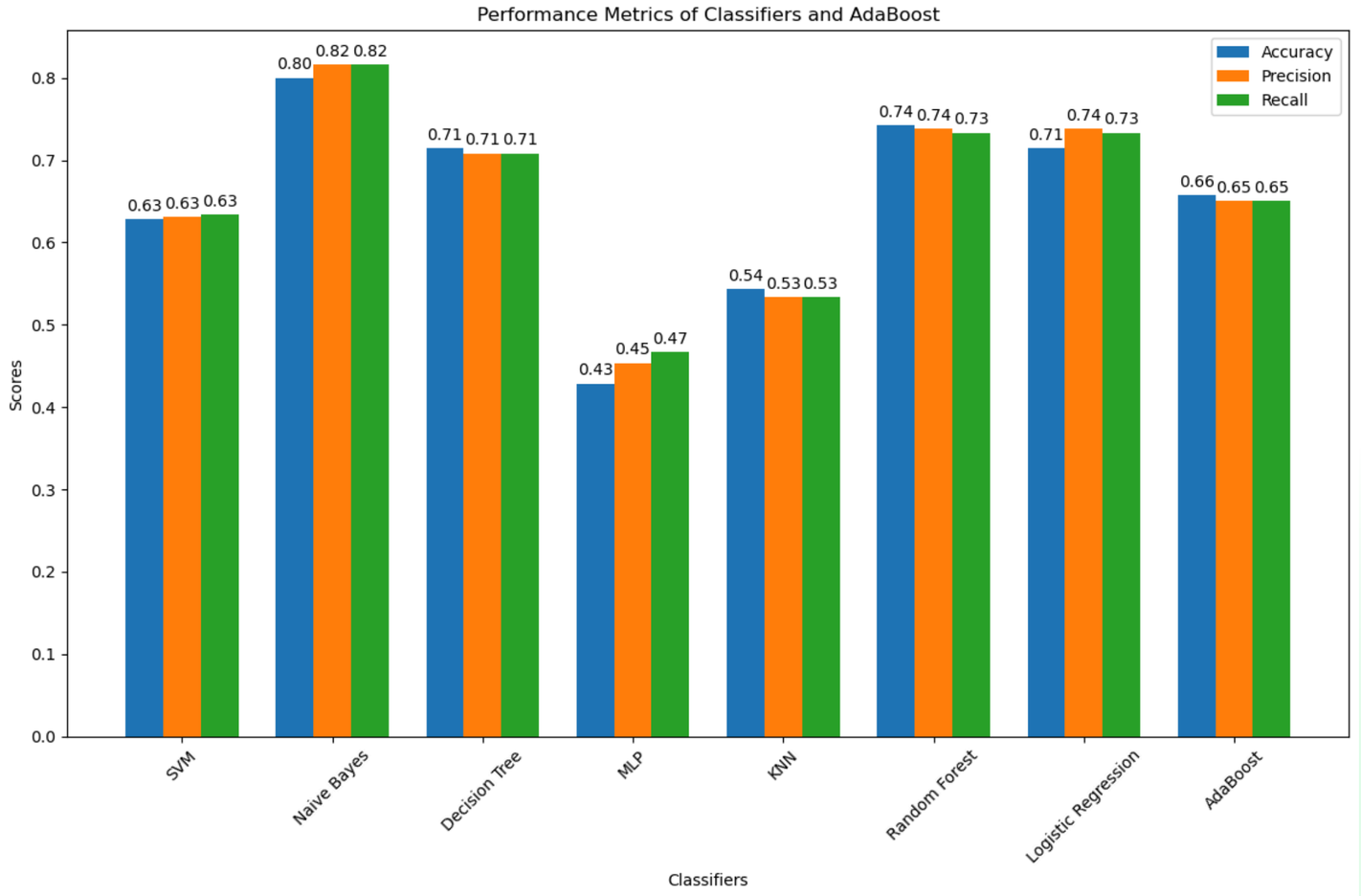
- Population Size: 50
- Generations: 20
- Crossover Rate: 0.8
- Mutation Rate: 0.02

Classifier	Mean Fitness Score	Selected Features
MLP	0.9143	23
Random Forest	0.9143	29
Decision Tree	0.8952	21
Naive Bayes	0.8571	29
SVM	0.8	24
Logistic Regression	0.8	25
KNN	0.6857	26

GA - Ensemble Learning and Performance Metrics

Ensemble Learning Techniques:

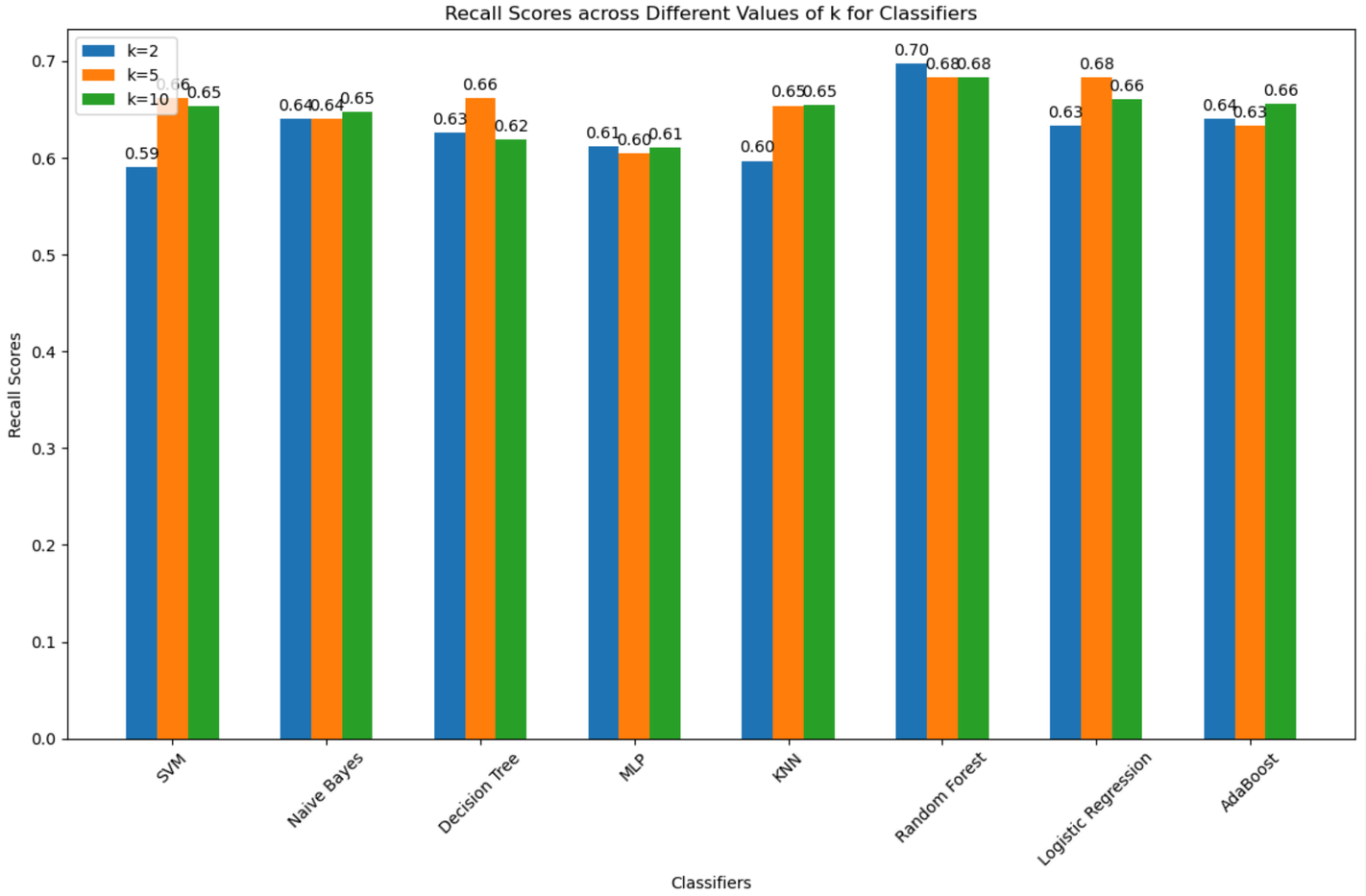
- Basic Learners: SVM, Naive Bayes, Decision Tree, MLP, KNN, RFC, LR
- Meta-Learner: AdaBoost



GA - Cross-Validation

Cross-Validation Approach:

- Importance of Cross-Validation
- Results for k=2, 5, 10
- Table of Cross-Validation Results

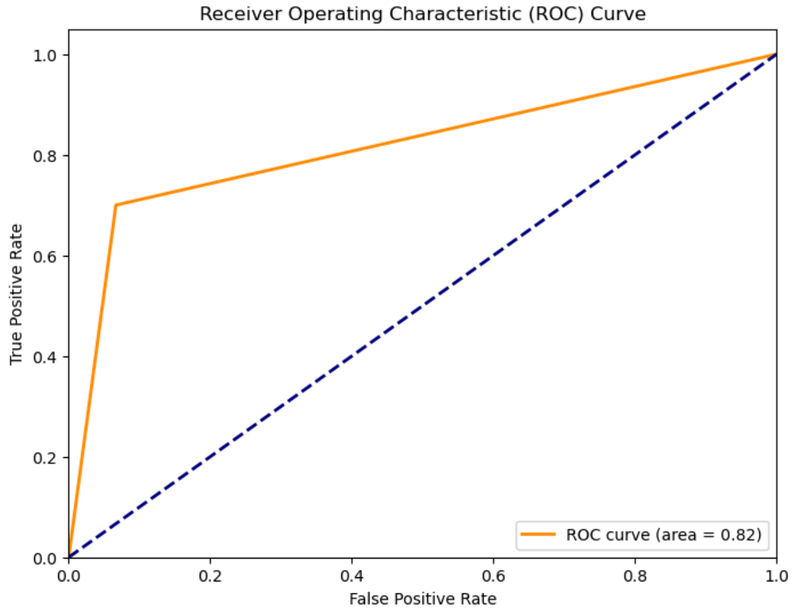


GA - Best clasifier: Random Forest

Validating model performance using cross-validation, it highlights Random Forest as the best classifier based on **performance metrics** and **cross-validation stability**.

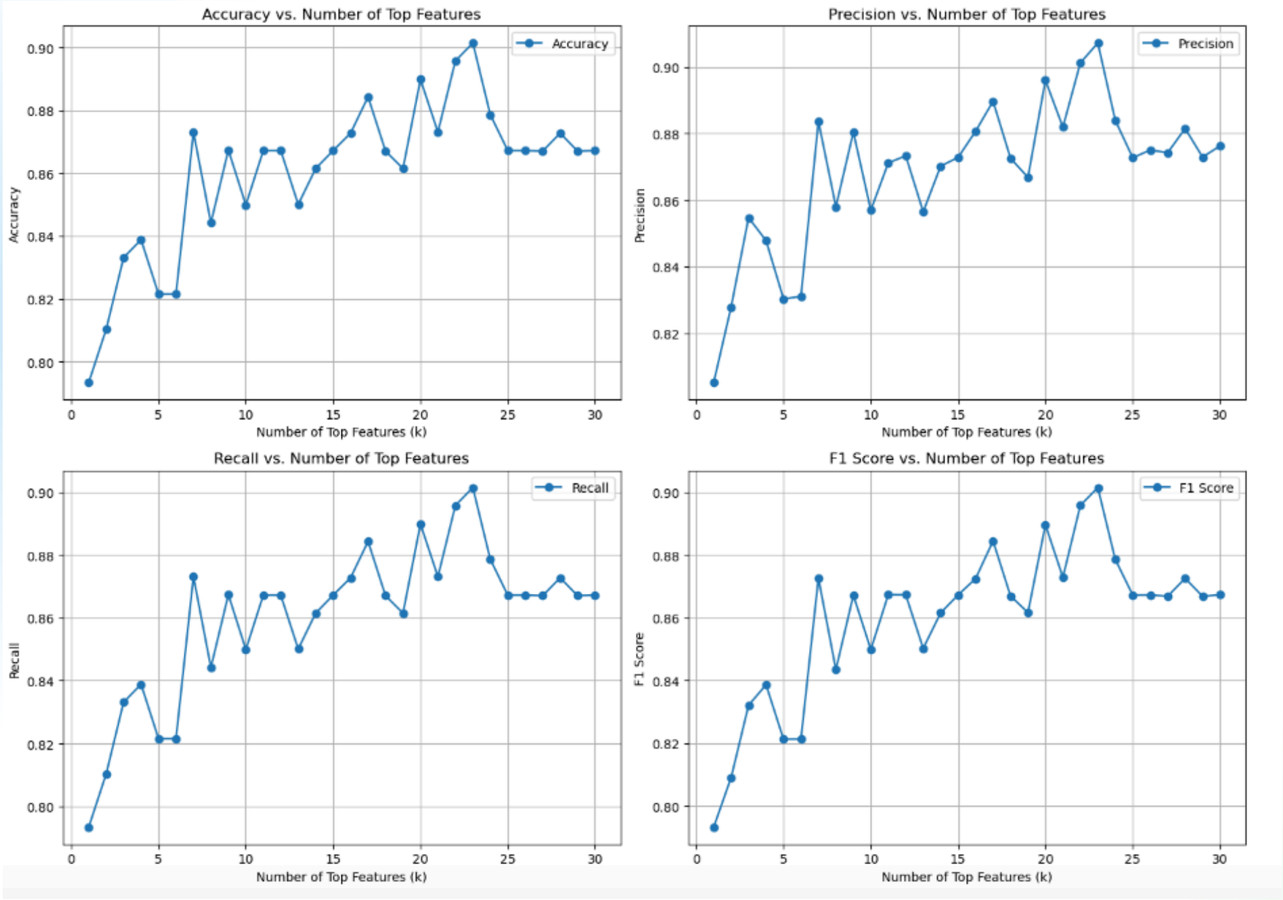
- Final Model Accuracy: 82%
- Confusion Matrix

	Predicted Patient (P)	Predicted Healthy (H)
True Patient (P)	14	6
True Healthy (H)	1	14



Multitask - Select the Best Features

Aggregated feature importance from Random Forest classifiers across all tasks.

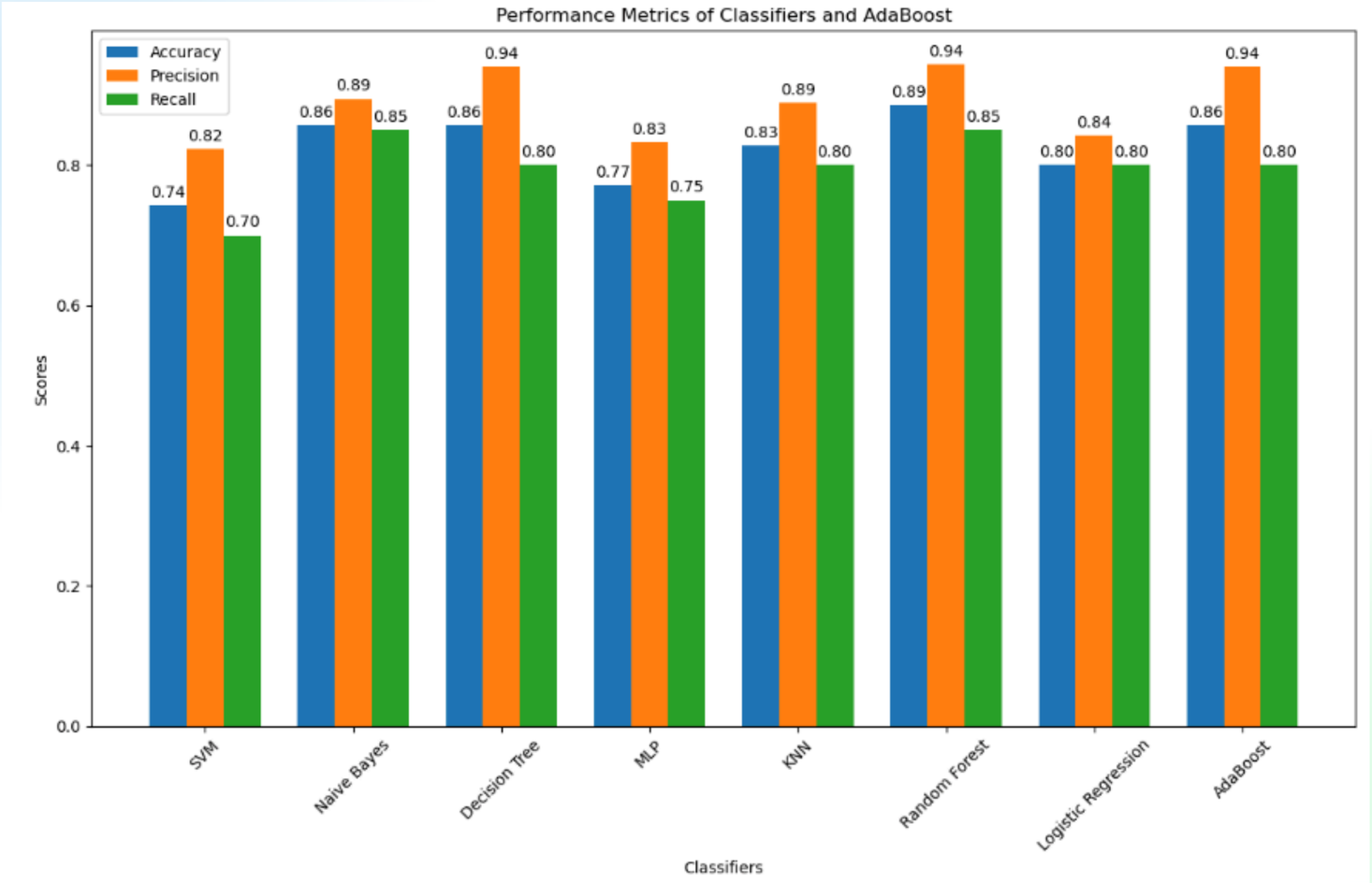


Optimal Number of Features: 22, where performance metrics stabilize, ensuring the model's effectiveness without unnecessary complexity.

Multitask - Ensemble Learning and Performance Metrics

Ensemble Learning Techniques:

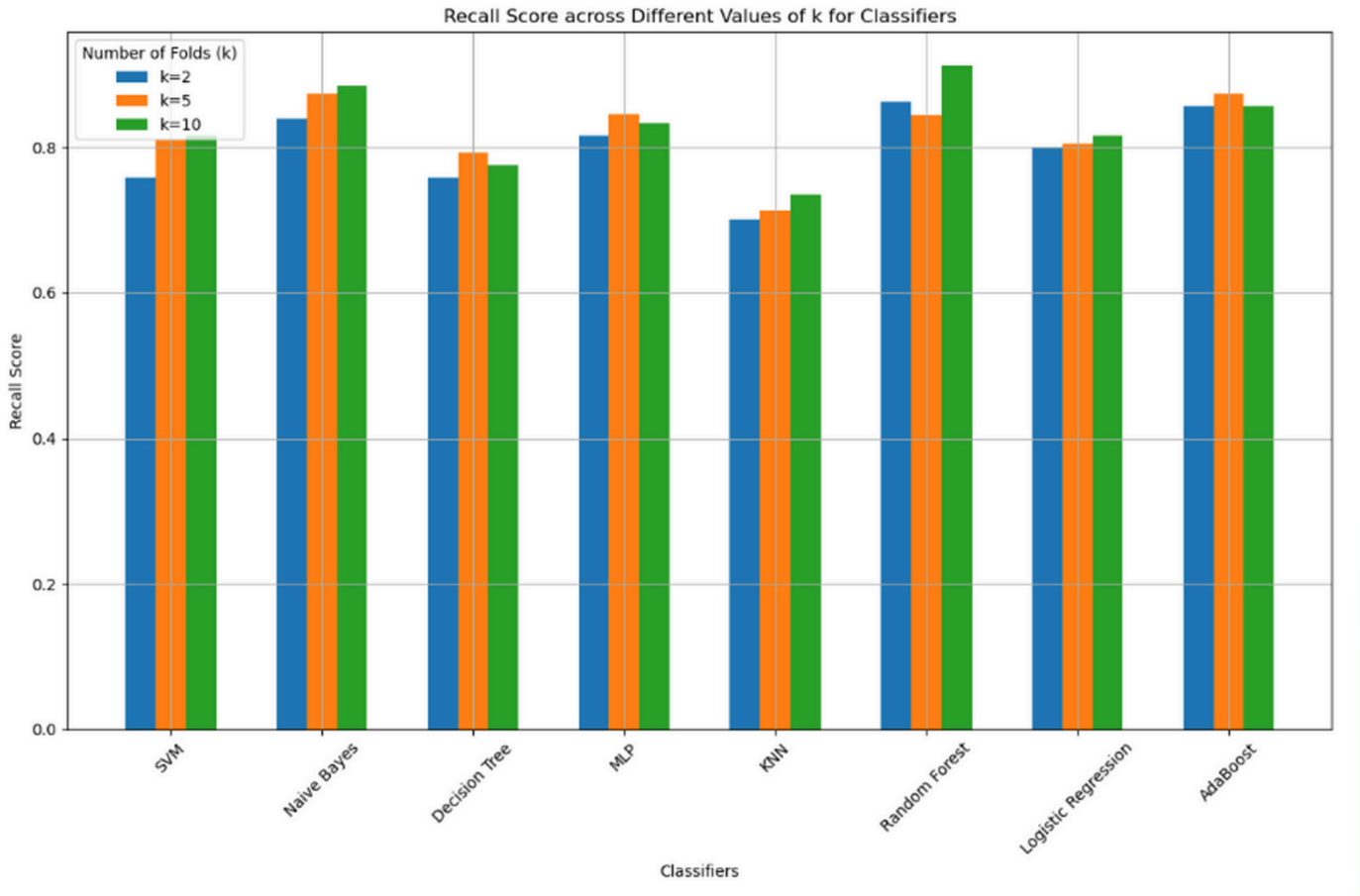
- Basic Learners: SVM, Naive Bayes, Decision Tree, MLP, KNN, RFC, LR
- Meta-Learner: AdaBoost



Multitask - Cross-Validation

Cross-Validation Approach:

- Importance of Cross-Validation
- Results for k=2, 5, 10
- Table of Cross-Validation Results

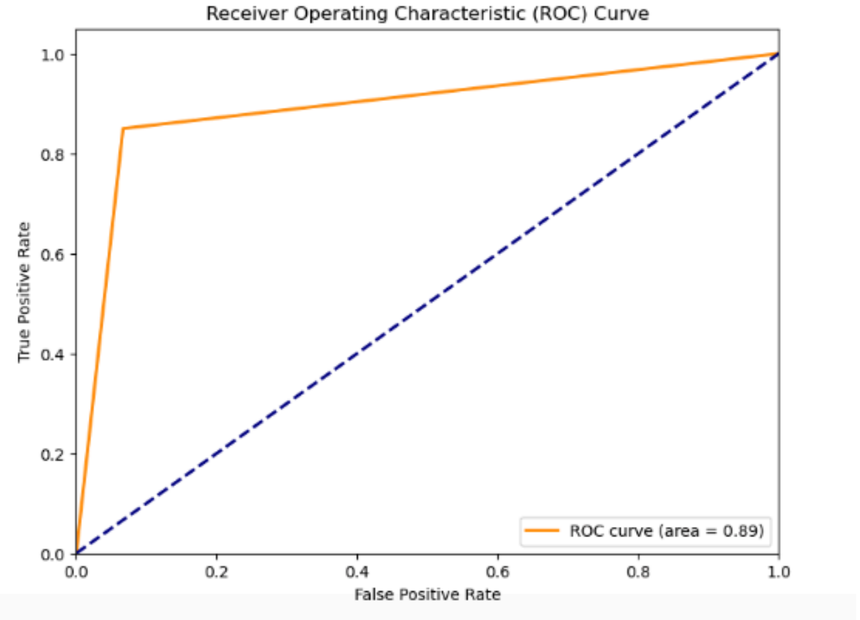


Multitask - Best clasifier: Random Forest

Random Forest emerged as the best model due to its high accuracy, strong recall, and robustness in cross-validation.

- Final Model Accuracy: 88.57%
- Confusion Matrix

	Predicted Patient (P)	Predicted Healthy (H)
True Patient (P)	17	3
True Healthy (H)	1	14



Conclusions

Result

All the methods were run on a MacBook Air M2 with 8 GB of memory.

Feature Selection Method	Number of Features	Best Classifier	Accuracy	Precision	Recall	k=2 Recall	k=5 Recall	k=10 Recall	Computational Time
ML with Original Dataset	450	Random Forest	0.89	0.88	0.89	0.85	0.84	0.9	90 seconds (1 minute 30 seconds)
PCA (Kaiser Criteria)	139	Random Forest	0.89	0.90	0.89	0.81	0.81	0.83	17 seconds
Stepwise	9	Random Forest	0.75	0.70	0.84	0.86	0.84	0.91	2538 seconds (about 42 minutes)
GA-Based	23	Random Forest	0.74	0.74	0.73	0.698	0.669	0.683	1410 seconds (about 23 minutes)
Multi-Task Learning	22	Random Forest	0.8857	0.9412	0.8500	0.862	0.844	0.913	122 seconds (about 2 minutes)

Conclusion

- PCA might have been a first step to **reduce** the very large feature space (450 features!).
- The stepwise selection method proved effective in **finding** a smaller set of relevant features.
- The filter and genetic algorithm (GA) **didn’t perform** well, properly because of outliers. It **takes** a lot of computational resouces.
- When the dataset has a **pattern**, the multi-task learning framework would be a good choice to find the optimal number of features
- The Random Forest classifier consistently **outperformed** other classifiers in terms of accuracy and recall and **stability** across the different k-fold cross-validations.
- Using AdaBoost as a meta-learner to combine the predictions of other classifiers **yielded** good results.

Areas of Improvement

- Use techniques like Grid Search or Random Search to explore a wider range of parameter combinations.
- Investigate Bagging, Gradient Boosting, and Voting Classifiers
- Analyzing the selected features (particularly the top 22) to understand their meaning and relationship
- Consider expanding the dataset by adding more samples or synthesizing new samples using techniques like data augmentation.

Future Work

- Apply in larger Dataset
- Use hyperparameter Tuning
- Deep Learning: CNNs,...

Bibliography

Abdollahi, J., & Nouri-Moghaddam, B. (2022). A hybrid method for heart disease diagnosis utilizing feature selection-based ensemble classifier model generation. Iranian Journal of Computer Science, 5(3), 229–246. https://doi.org/10.1007/s42044-022-00104-x

Cillia, M., Gregorio, C., Fontanella, L., Marcelli, A., & Parziale, A. (2022). An enhanced machine-learning framework for predicting Alzheimer's disease via handwriting analysis. Engineering Applications of Artificial Intelligence, 112, 104902. https://doi.org/10.1016/j.engappai.2022.104902

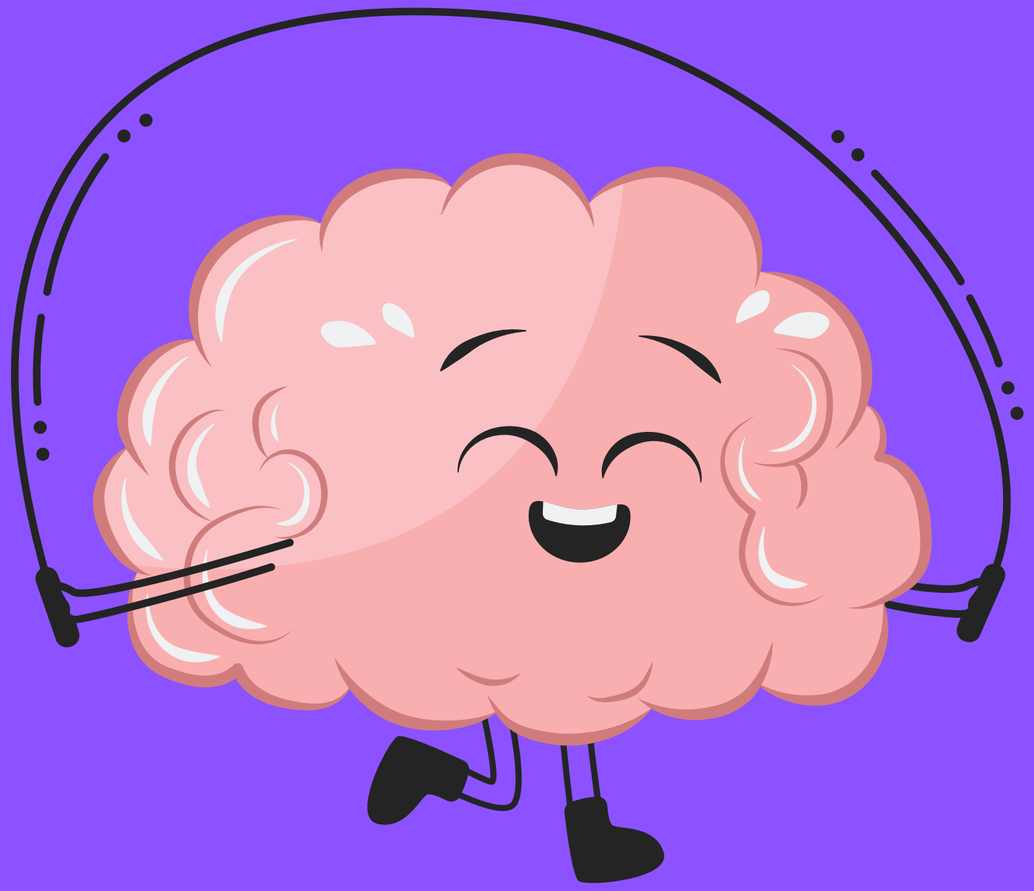
Mujahid, M., Rehman, A., Alam, T., Alamri, F. S., Fati, S. M., & Saba, T. (2023). An efficient ensemble approach for Alzheimer's disease detection using an adaptive synthetic technique and deep learning. Diagnostics, 13(2489). Retrieved from https://www.sciencedirect.com/science/article/pii/S235291482100143X

Öcal, H. (2023). A novel approach to detection of Alzheimer's disease from handwriting: Triple ensemble learning model. Journal of Applied Artificial Intelligence. Retrieved from https://dergipark.org.tr/tr/download/article-file/3518417.

Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked, 26, 100655. https://doi.org/10.1016/j.imu.2021.100655

N. D. Cilia, C. De Stefano, F. Fontanella, A. S. Di Freca, An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis, Procedia Computer Science 141 (2018) 466–471. https://doi.org/10.1016/j.procs.2018.10.141

N. D. Cilia, G. De Gregorio, C. De Stefano, F. Fontanella, A. Marcelli, A. Parziale, Diagnosing Alzheimer's disease from online handwriting: A novel dataset and performance benchmarking, Engineering Applications of Artificial Intelligence, Vol. 111 (20229) 104822. https://doi.org/10.1016/j.engappai.2022.10



THANK YOU!

Q&A