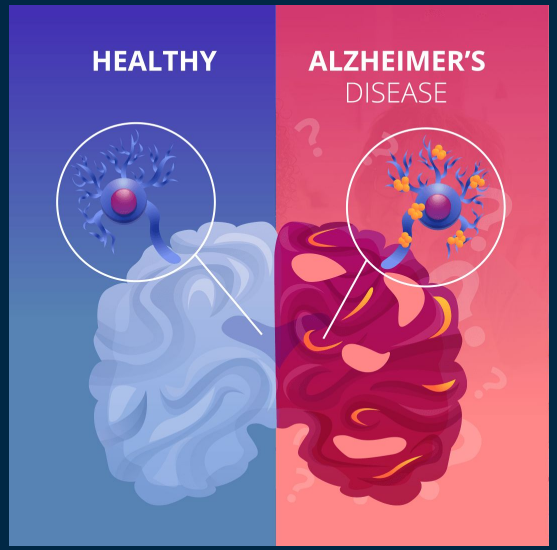
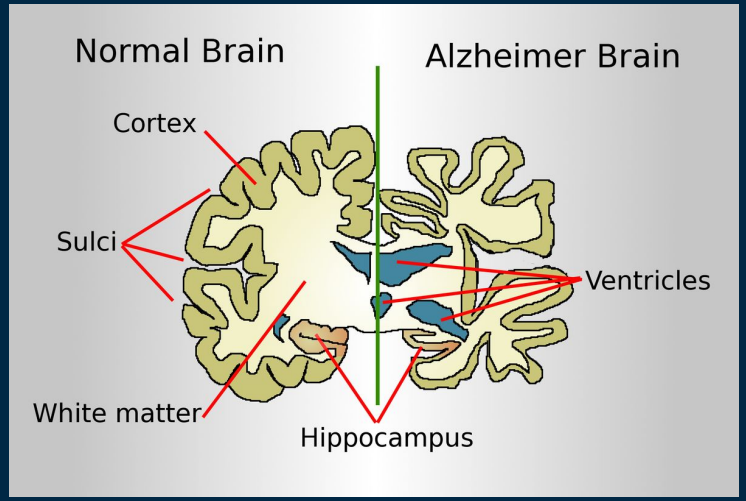
Diagnosing Alzheimer's disease from handwriting

Gabriele Billi Ciani, Gabriele Giudici
University of Pisa
Data Mining and Machine Learning Course
A.Y. 2023-2024

Neurodegenerative Disorders

Neurodegenerative disorders: Progressive cognitive and motor function decline.





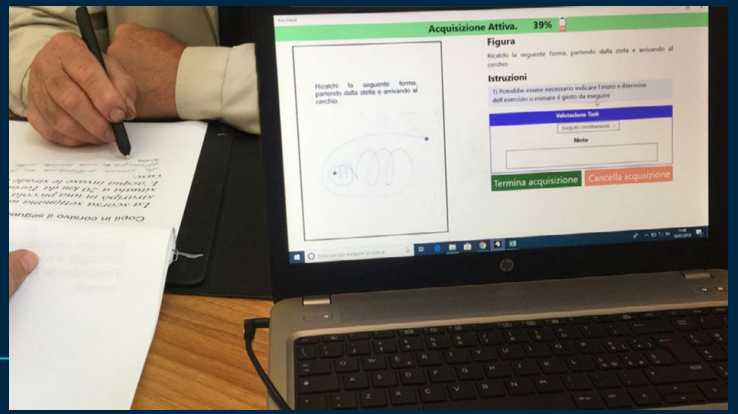
Alzheimer's disease (AD): Leads to memory loss, cognitive impairments, and, in later stages, can also affect motor functions, causing coordination and balance issues.

The Role of Handwriting Analysis

Handwriting: Requires coordination of multiple brain regions.

Offers a non-invasive method for early detection of cognitive impairments.

Changes in handwriting can indicate early stages of Alzheimer's.



Data acquisition is performed using a graphic tablet

The DARWIN Data Set

Created by Cilia et al. (2022) for Alzheimer's research. https://doi.org/10.1016/j.engappai.2022.104822

Aimed at supporting early diagnosis through handwriting.

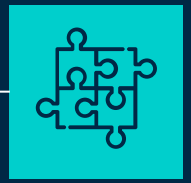
- 174 participants: 89 with AD, 85 healthy people.
- 25 handwriting tasks: graphic, copy, memory, and dictation.
- 450 attributes extracted, all of them are numeric (coordinates, pressure, timestamps, ...).

https://archive.ics.uci.edu/dataset/732/darwin

Methodology

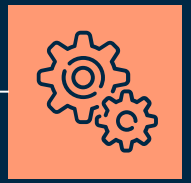
- **Grid-search** to optimize the hyperparameters for all models.
- **5-fold cross-validation** to validate the model's performance.
- As an alternative to cross-validation, **holdout validation method** (used by Cilia et al.): dataset is shuffled and split into training (80%) and test (20%) sets for 20 iterations.

EXPERIMENTAL RESULTS



01

Classification on the Full Data Set



02

Task-Specific Classification

03

Ensemble Modelling via Majority Voting

04

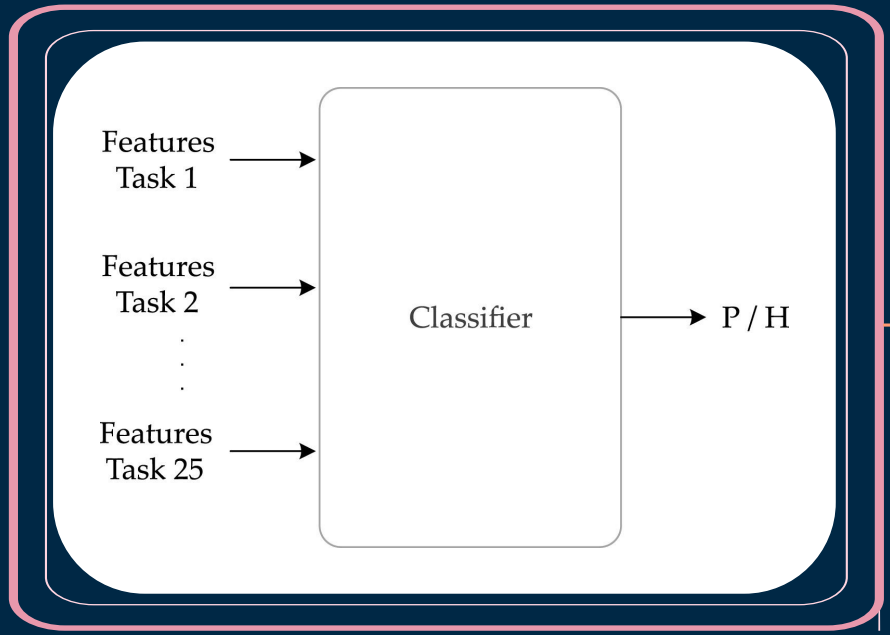
Task Reduction

05

Feature Reduction

Classification on the Full Data Set

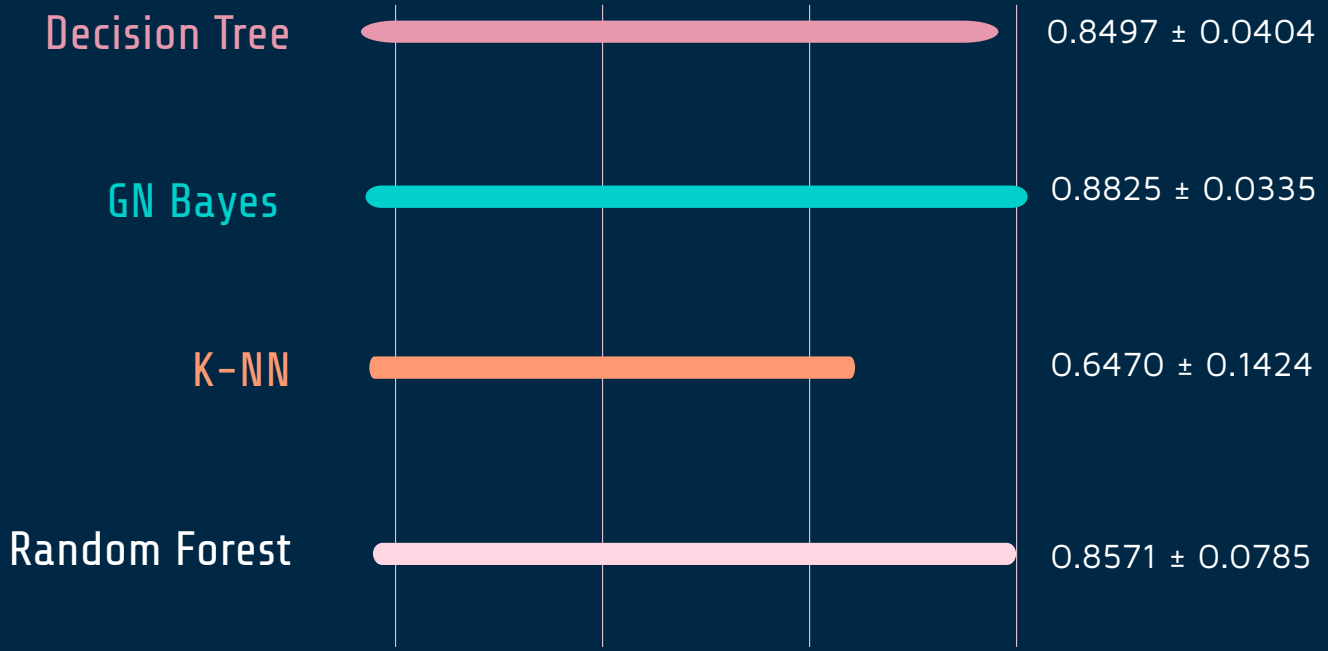
- We used the complete DARWIN data set with all 450 features.
- Aimed at establishing baseline performance for Alzheimer's detection.
- We evaluated our four Classifiers:
  -Decision Tree,
  -Random Forest,
  -K-NN,
  -Gaussian Naive Bayes.



P: Patient, H: Healthy

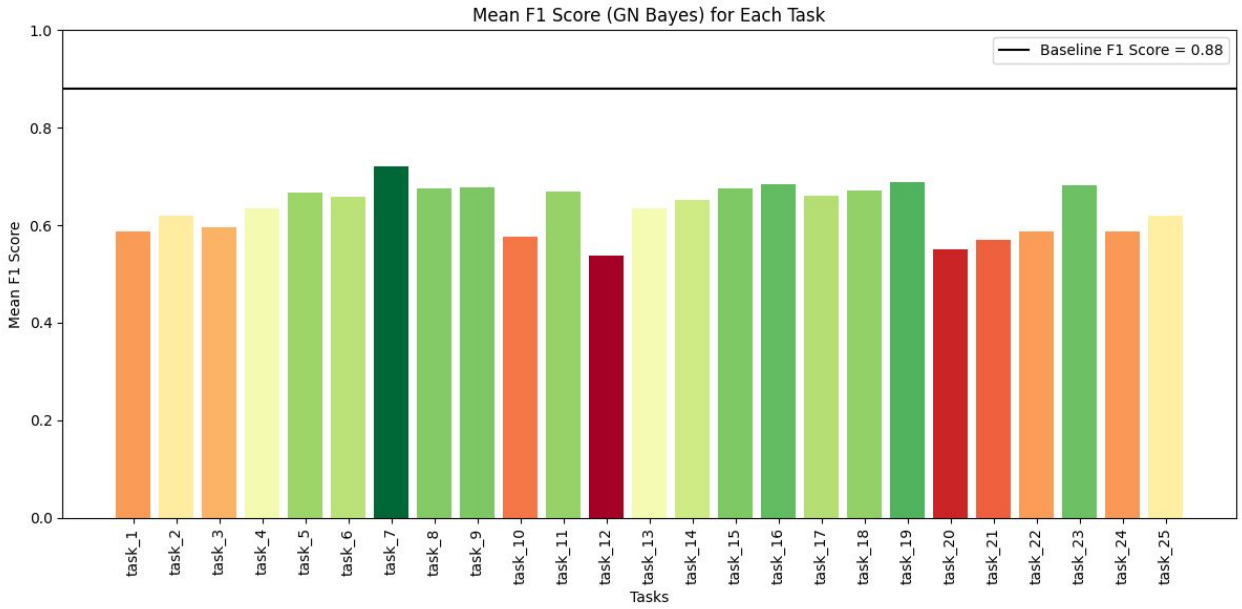
Performance Comparison (F1 Score)

- Results in line with those of reference paper.



Task-Specific Classification

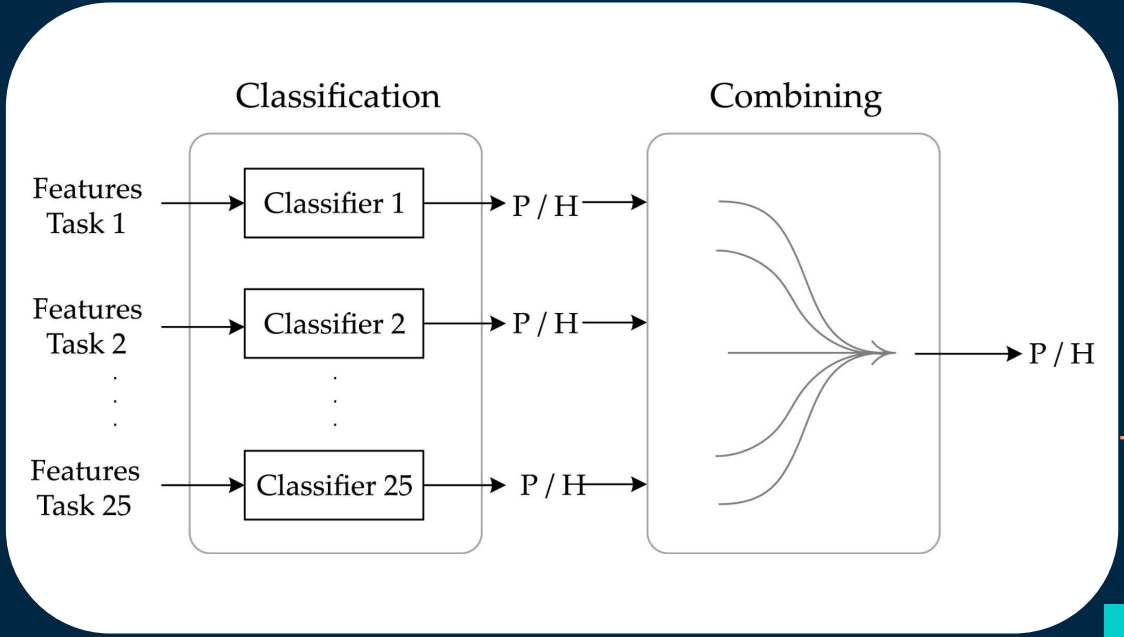
- Explores the diagnostic potential of individual handwriting tasks.
- Training and evaluating our four classifiers on 25 distinct handwriting tasks.
- Results in line with those of reference paper.



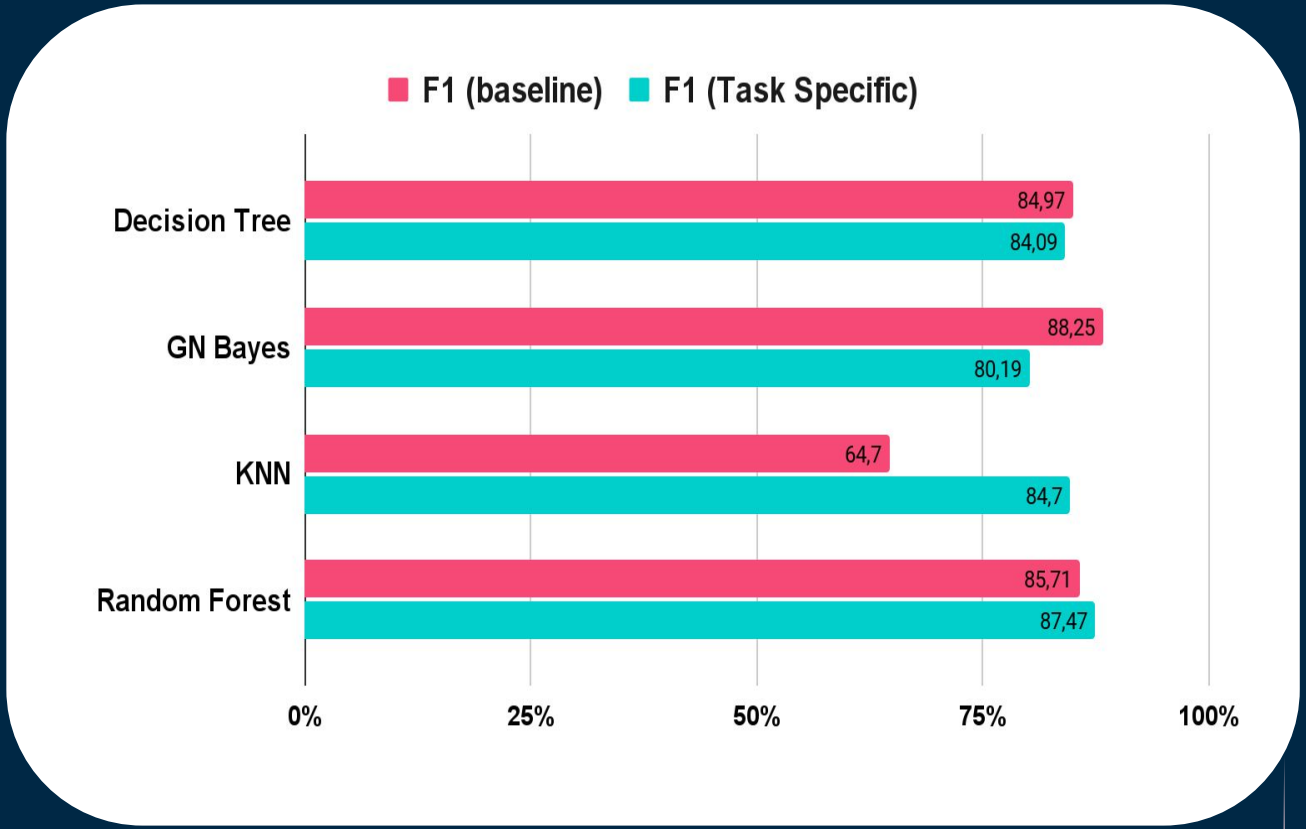
GN Bayes mean F1 metrics for Each Task

Ensemble Modelling via Majority Voting

- Each ensemble combines predictions from the 25 task-specific classifiers using majority voting.
- Image description:
  - P: Patient
  - H: Healthy

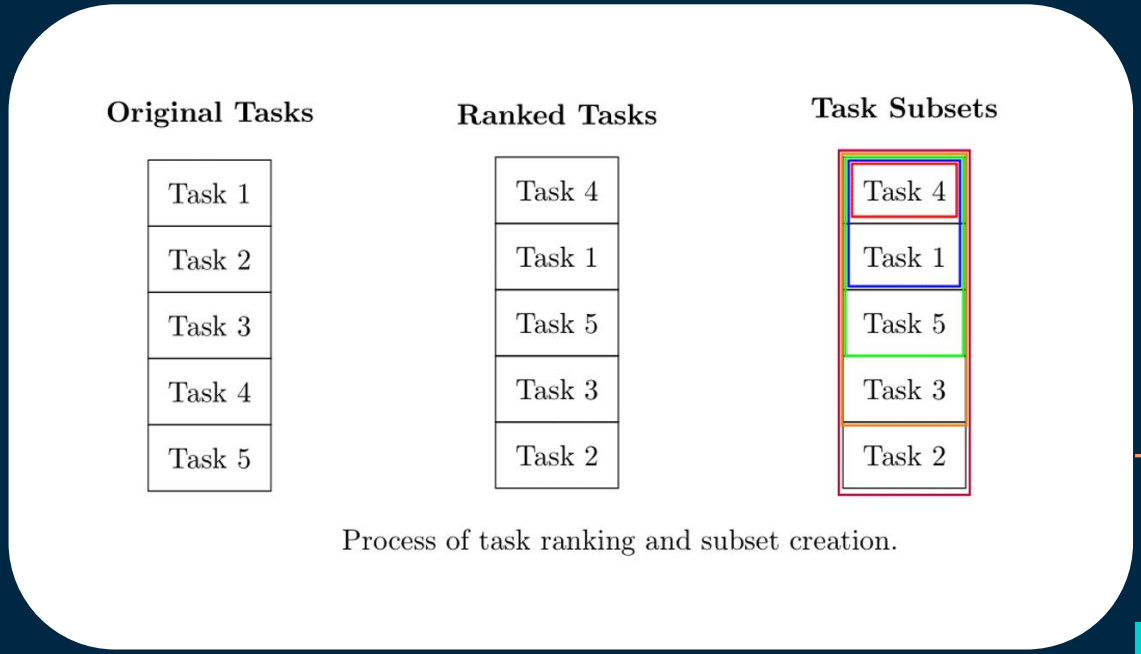


- RF and K-NN outperformed baseline, proving the strength of ensemble models in improving diagnostic accuracy.
- Our results align with Cilia et al., but we did not replicate the exceptional DT performance reported in their study.



Task Reduction

- We ranked tasks based on their performance metrics for each classifier.
- We created subsets of tasks, starting with the top-ranked ones, and progressively added tasks according to their rank.
- Each subset was evaluated using the majority voting strategy.



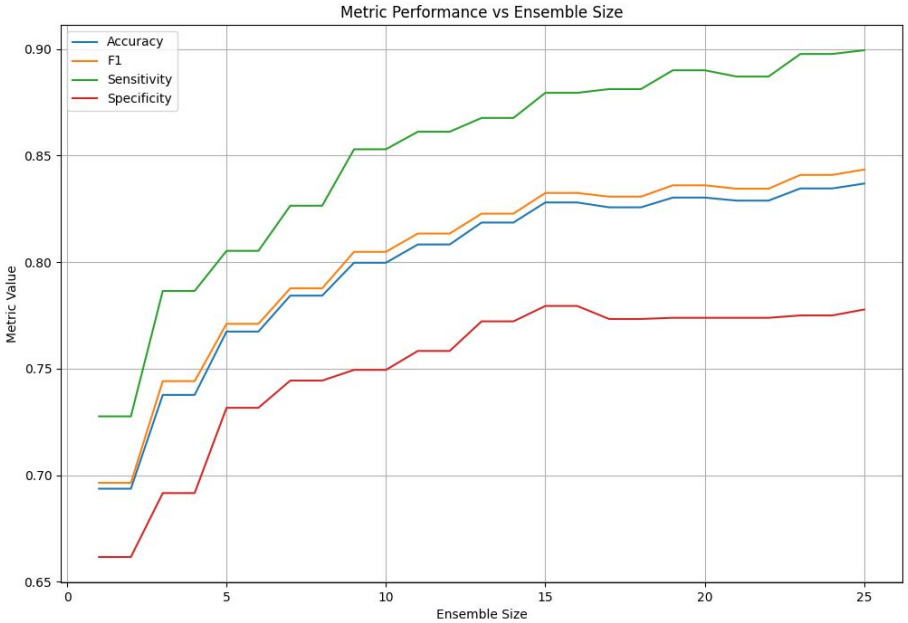
Process of task ranking and subset creation.

Example with 5 Tasks

- To ensure the integrity and validity of our model evaluation, the ranking of tasks was performed **exclusively** on the training set. This precaution was taken to avoid data leakage, which can occur when the same data is used for both ranking tasks and testing the model's performance.

- If the same data were used for both ranking tasks and evaluating the model, it could artificially inflate the performance metrics. This happens because the model would have 'seen' information from the test set during the training process, leading to overly optimistic results that don't reflect true performance on unseen data.

- None of the models achieved performance improvements by using a subset of tasks compared to using all tasks.
- However, task reduction might still be considered through statistical analysis to identify subsets that provide comparable performance to using all tasks.



K-NN Performance Metrics for all subsets of tasks

Please note that the Ensemble Method result for K-NN (F1) was 84.7%

Feature Reduction

- In addition to task reduction, we explored **standard feature reduction methods** to enhance classifier performance.
- Unlike task reduction, feature reduction simplifies the feature set by reducing its dimensionality, without considering task-specific structures.
- We applied four techniques: PCA, RFE, NMIFS, and mRMR.

- A consistent pipeline approach was applied to each feature reduction technique: Data Preparation, Feature Reduction, Classifier Evaluation.

- This is something reference paper did not explore.

Principal Component Analysis (PCA)

- We explored variance thresholds from 35% to 98% across all four classifiers.
- Results for each classifier were mixed:

GNB: No improvement over baseline.

DT: Comparable performance at a 35% variance threshold, but improved only after parameter adjustments.

K-NN: F1 score improved to 0.7478 at a 35% variance threshold (baseline F1: 0.6470), but still below task reduction.

RF: PCA did not surpass baseline performance.

Normalized Mutual Information Feature Selection (NMIFS)

- NMIFS was applied exclusively to the RF model due to high computational demands.
- We evaluated selected components ranging from 2 to 50 features.
- All metrics improved as more features were included, suggesting that exploring higher numbers of features could yield better results.

Recursive Feature Elimination (RFE)

- RFE was applied to the RF model, with the number of selected features ranging from 5 to 450.
- All metrics peaked at 450 features, indicating that RFE does not significantly benefit the RF model.

Minimum-Redundancy Maximum-Relevance (mRMR)

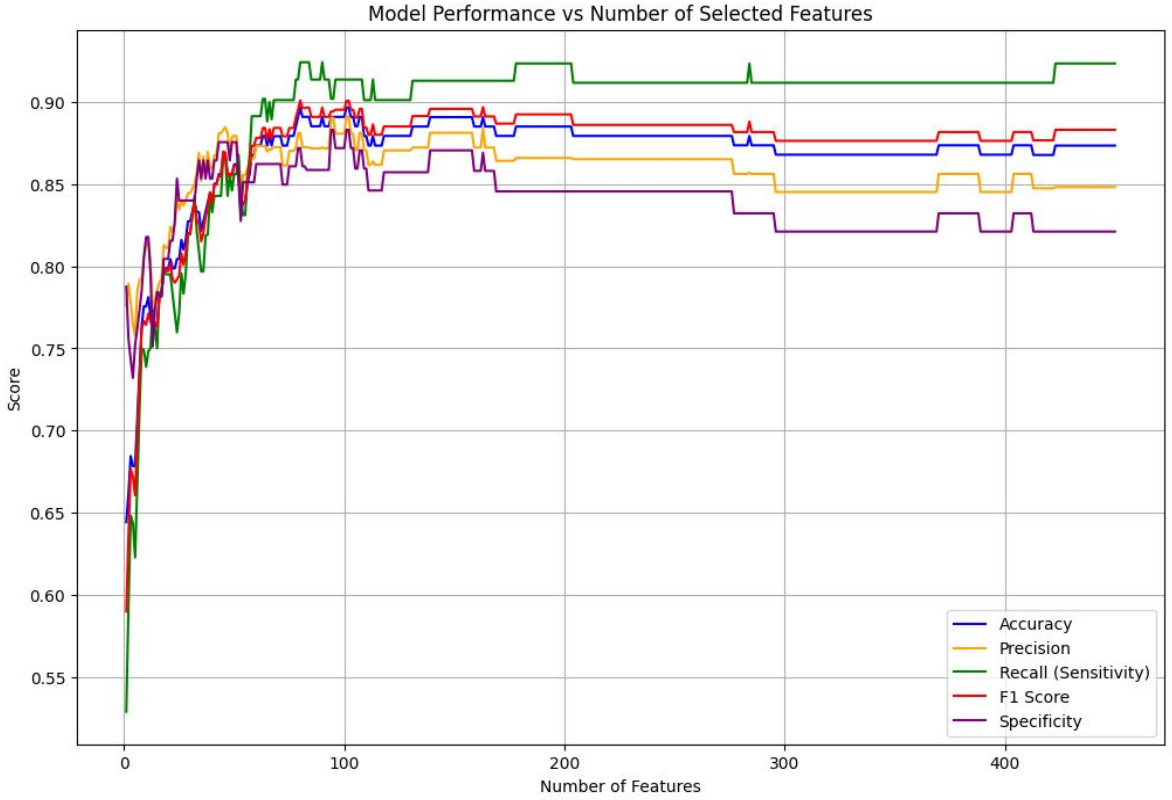
- mRMR was applied to all four classifiers, selecting between 1 and 450 features.

- **RF**: Comparable to baseline only with more than 300 features.

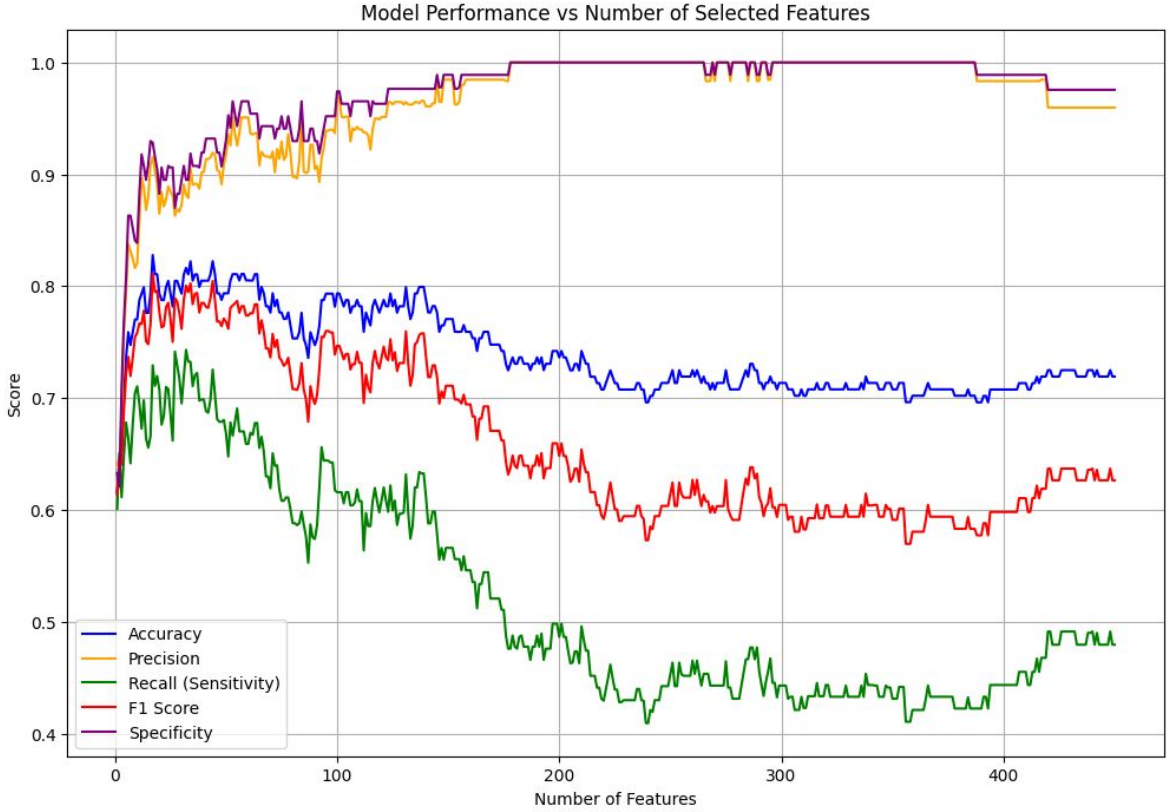
  DT: No feature subset smaller than the maximum improved performance.

  K-NN: Performance peak between 15 and 60 features, significantly surpassing the baseline.

  GNB: Baseline performance achieved with 70 features, with no improvement beyond that.



mRMR on **GN Bayes**


Model Performance vs Number of Selected Features
Score
Number of Features
Accuracy
Precision
Recall (Sensitivity)
F1 Score
Specificity
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
100
200
300
400

mRMR on **K-NN**

Takeaways from Feature Reduction

- For **different Classifiers**, Feature Reduction Techniques have showed **different results**.

- Differently from Task Reduction, which allows to simplify the data acquisition process, feature reduction remains a valuable method for simplifying models and may offer new avenues for **further analysis**.

Proposed Enhancements and Future Experiments

1) Task-Specific Feature Reduction: Apply feature reduction techniques to each handwriting task before classification.

2) Combining Task and Feature Reduction: Use a hybrid approach to reduce both the number of tasks and features within selected tasks.

3) Task Reduction with Redundancy Analysis: Eliminate redundant tasks based on correlation to prioritize the most unique, informative tasks.

Conclusions

- We evaluated various methods for improving classification models in diagnosing Alzheimer's disease using the DARWIN dataset.
- Ensemble methods significantly enhanced diagnostic accuracy
- Task reduction strategies might streamline the process by reducing the number of required tasks without compromising performance.
- Feature reduction holds potential for **further exploration**.
- Each classification model benefits from different approaches.

- Our study underscores the value of combining task-specific and feature-specific strategies to optimize diagnostic performance.
- Future research can integrate these approaches to refine Alzheimer's detection methods.