Introduction

Alzheimer’s disease (AD) is a progressive neurological disorder that affects memory, thinking, and behavior. Early detection is crucial for improving patient care and management, but it remains a significant challenge due to the complexity of contributing factors and the long progression period of the disease. This project aims to develop a predictive model to assess the likelihood of an individual having Alzheimer's disease based on a range of features, including demographic details, medical history, lifestyle factors, clinical measurements, and cognitive assessments.

Below is a summary of the modeling approach for this project.

Goals

- To obtain a good baseline model
- Select models that produce high accuracy rates using all the features
- Adapt an approach which improves the accuracy for early detection
    - By excluding the highly correlated features in modeling
    - Restricting to the dataset with no memory complaint

Models Explored

We will be exploring the following models in this project.

1. Logistic Regression
2. Random forest classifier
3. XGBoost classifier
4. Gradient boosting classifier
5. Adaboost classifier
6. KNN classifier
7. Support Vector machine
8. Decision Tree
9. LDA
10. QDA
11. Naĭve Bayes

Metrics Checked

To evaluate the performance of the models, we will be considering the following metrics.

- Accuracy Score
- F1 Score
- Precision Score
- Recall Score
- Confusion Matrices

Methodology

For each method explored, we considered analysis on two broad scenarios:

1. The entire dataset
2. A restriction on the dataset based on only patients who had no memory complaints.

Additionally, we grouped explored the models further under these two broad scenarios by grouping the features based on similarities. The groupings are:

- All features (all_features)
- Demographic features (demo)
- Lifestyle features (lifestyle)
- Medical history features (medic)
- Clinical features (clinical)
- Cognitive features (cognitive)
- Features related to symptoms (symptoms)
- Non-cognitive features (no_cognitive)

Results

In this section, we show the results for model exploration. The scores for the various error metrics are tabulated, while we also include confusion matrices for all features, cognitive features and non-cognitive features.

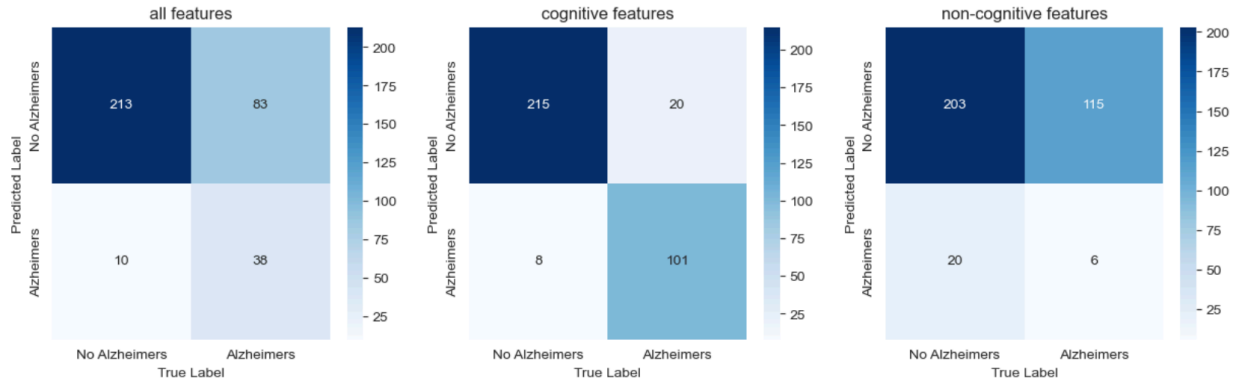
KNN Classifier

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.729651	0.633721	0.654070	0.604651	0.627907	0.918605	0.566860	0.607558
Precision Score	0.314050	0.066116	0.082645	0.107438	0.099174	0.834711	0.272727	0.049587
Recall Score	0.791667	0.380952	0.555556	0.317073	0.387097	0.926606	0.351064	0.230769
F1 Score	0.449704	0.112676	0.143885	0.160494	0.157895	0.878261	0.306977	0.081633

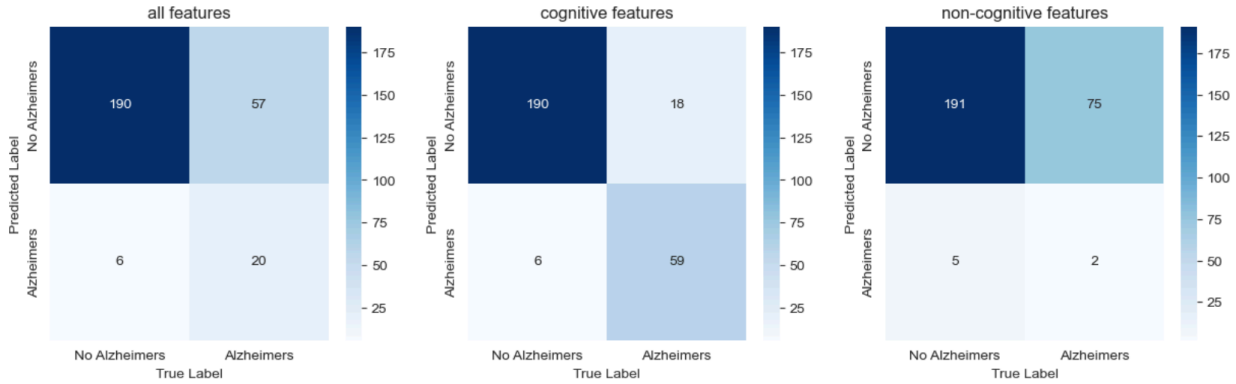
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.769231	0.710623	0.717949	0.688645	0.684982	0.912088	0.699634	0.706960
Precision Score	0.259740	0.038961	0.064935	0.116883	0.012987	0.766234	0.012987	0.025974
Recall Score	0.769231	0.375000	0.500000	0.346154	0.090909	0.907692	0.142857	0.285714
F1 Score	0.388350	0.070588	0.114943	0.174757	0.022727	0.830986	0.023810	0.047619

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

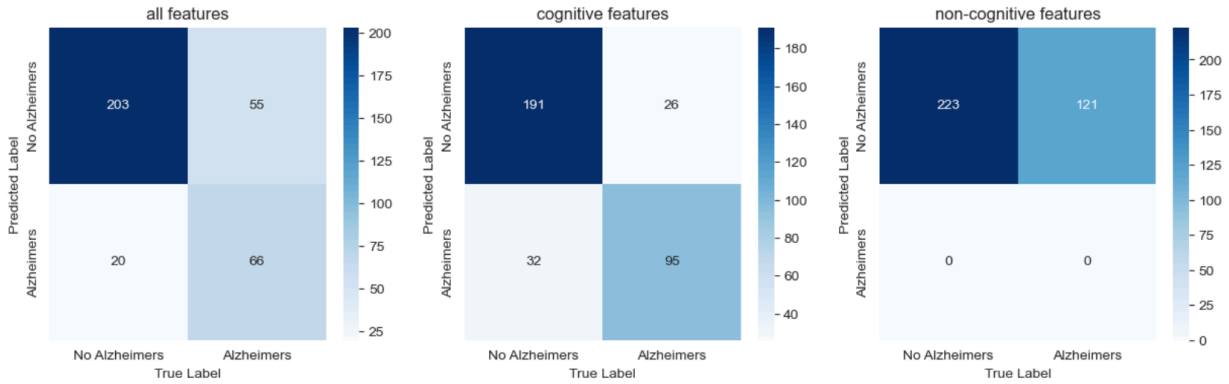
SVM (Polynomial Kernel)

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.781977	0.648256	0.648256	0.651163	0.648256	0.831395	0.648256	0.648256
Precision Score	0.545455	0.000000	0.000000	0.024793	0.000000	0.785124	0.000000	0.000000
Recall Score	0.767442	0.000000	0.000000	0.600000	0.000000	0.748031	0.000000	0.000000
F1 Score	0.637681	0.000000	0.000000	0.047619	0.000000	0.766129	0.000000	0.000000

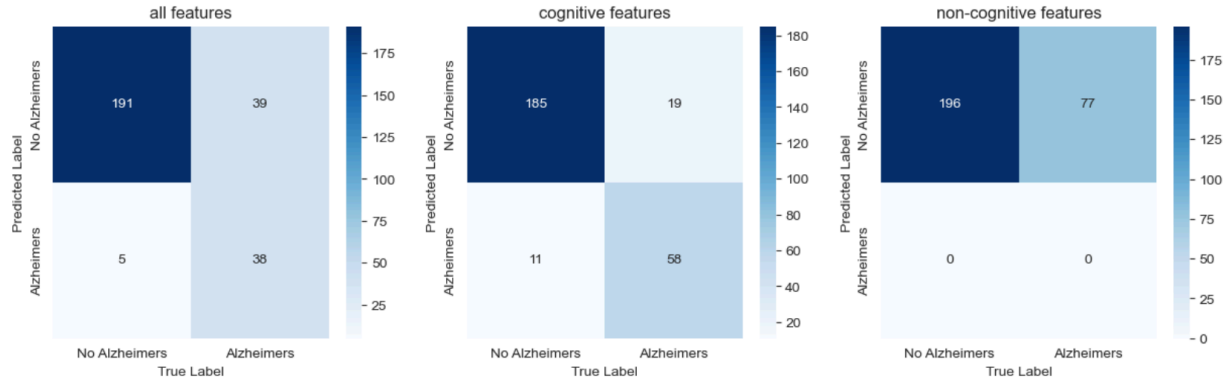
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.838828	0.717949	0.717949	0.717949	0.717949	0.890110	0.717949	0.717949
Precision Score	0.493506	0.000000	0.000000	0.000000	0.000000	0.753247	0.000000	0.000000
Recall Score	0.883721	0.000000	0.000000	0.000000	0.000000	0.840580	0.000000	0.000000
F1 Score	0.633333	0.000000	0.000000	0.000000	0.000000	0.794521	0.000000	0.000000

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

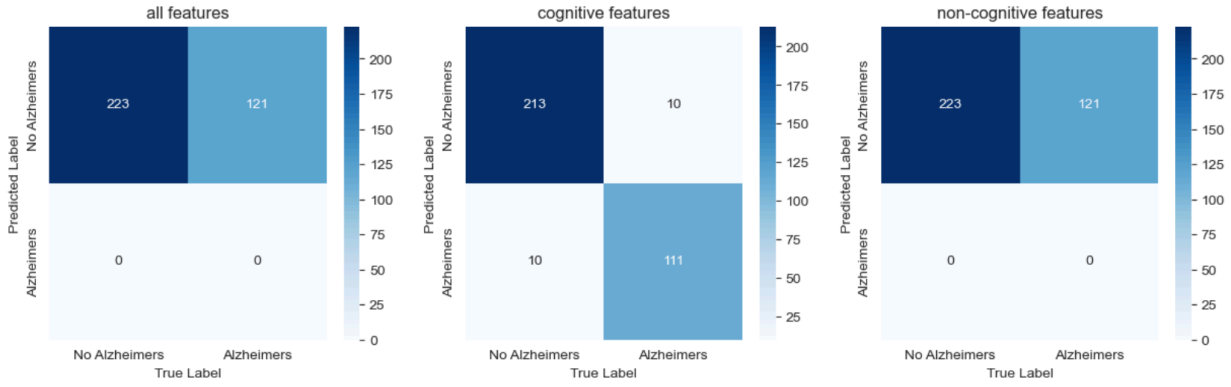
SVM (RBF Kernel)

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.648256	0.654070	0.651163	0.654070	0.648256	0.941860	0.648256	0.648256
Precision Score	0.000000	0.041322	0.016529	0.016529	0.000000	0.917355	0.000000	0.000000
Recall Score	0.000000	0.625000	0.666667	1.000000	0.000000	0.917355	0.000000	0.000000
F1 Score	0.000000	0.077519	0.032258	0.032520	0.000000	0.917355	0.000000	0.000000

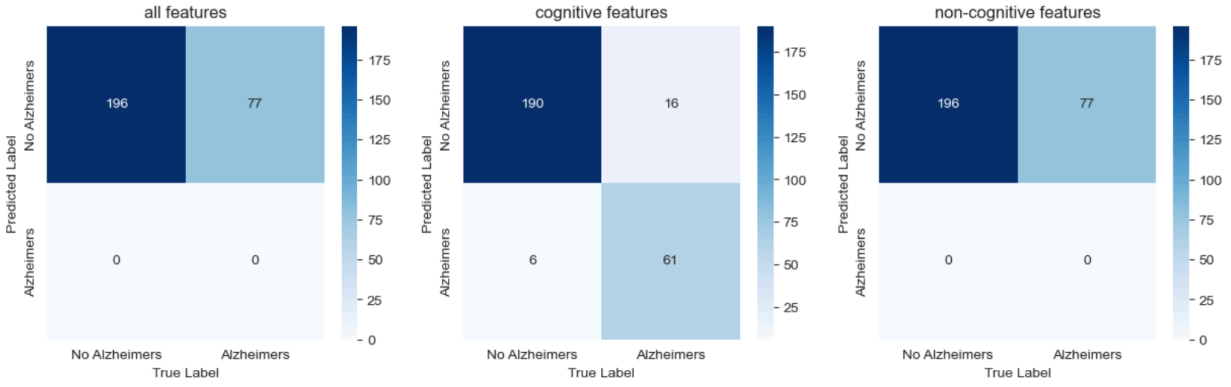
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.717949	0.717949	0.717949	0.717949	0.717949	0.919414	0.717949	0.717949
Precision Score	0.000000	0.000000	0.000000	0.000000	0.000000	0.792208	0.000000	0.000000
Recall Score	0.000000	0.000000	0.000000	0.000000	0.000000	0.910448	0.000000	0.000000
F1 Score	0.000000	0.000000	0.000000	0.000000	0.000000	0.847222	0.000000	0.000000

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

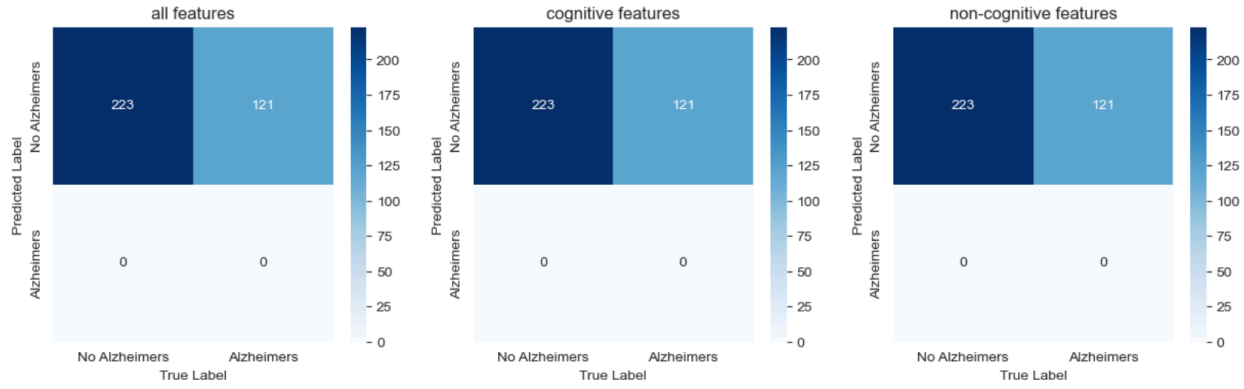
SVM (Sigmoid Kernel)

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.648256	0.648256	0.648256	0.648256	0.648256	0.648256	0.648256	0.648256
Precision Score	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Recall Score	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
F1 Score	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

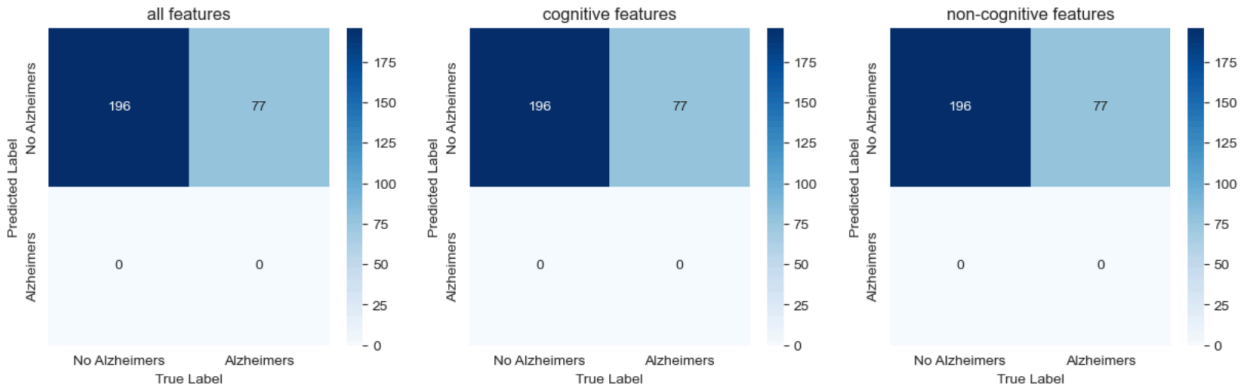
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.717949	0.717949	0.717949	0.717949	0.717949	0.717949	0.717949	0.717949
Precision Score	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Recall Score	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
F1 Score	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

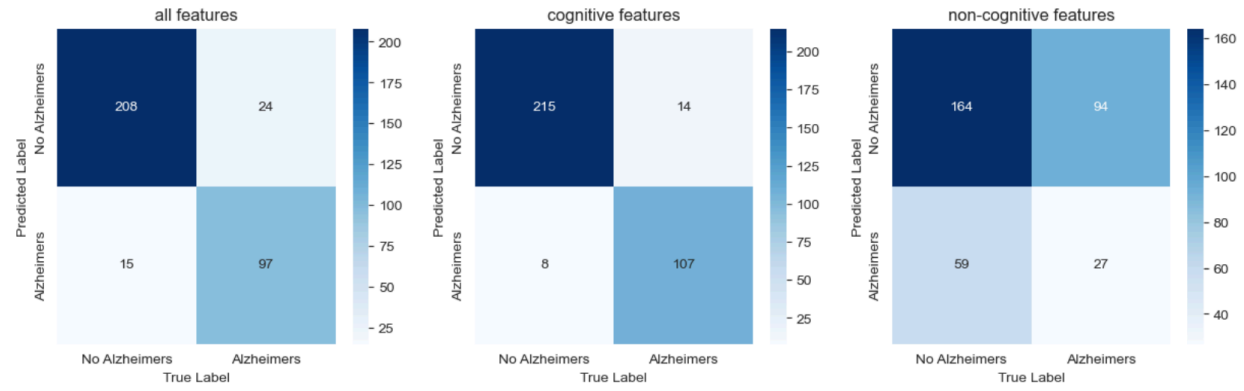
Decision Tree Classifier

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.886628	0.578488	0.601744	0.651163	0.581395	0.936047	0.648256	0.555233
Precision Score	0.801653	0.289256	0.157025	0.024793	0.223140	0.884298	0.000000	0.223140
Recall Score	0.866071	0.372340	0.351852	0.600000	0.350649	0.930435	0.000000	0.313953
F1 Score	0.832618	0.325581	0.217143	0.047619	0.272727	0.906780	0.000000	0.260870

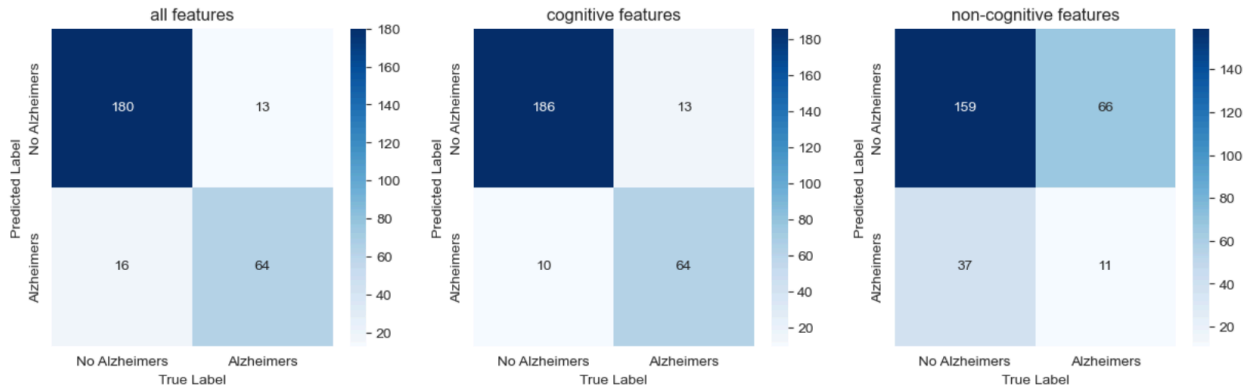
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.893773	0.681319	0.644689	0.710623	0.630037	0.915751	0.717949	0.622711
Precision Score	0.831169	0.142857	0.181818	0.038961	0.155844	0.831169	0.000000	0.142857
Recall Score	0.800000	0.343750	0.291667	0.375000	0.250000	0.864865	0.000000	0.229167
F1 Score	0.815287	0.201835	0.224000	0.070588	0.192000	0.847682	0.000000	0.176000

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

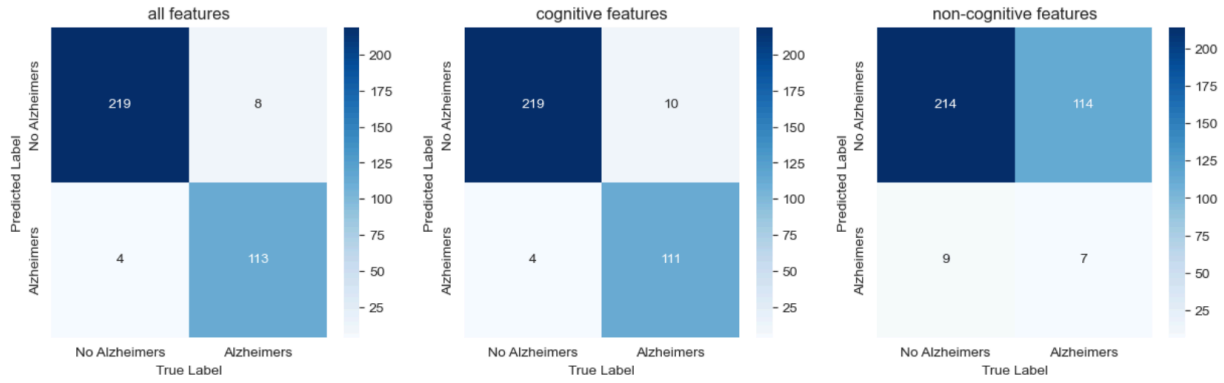
AdaBoost Classifier

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.965116	0.645349	0.654070	0.648256	0.639535	0.959302	0.639535	0.642442
Precision Score	0.933884	0.057851	0.016529	0.024793	0.066116	0.917355	0.008264	0.057851
Recall Score	0.965812	0.466667	1.000000	0.500000	0.421053	0.965217	0.200000	0.437500
F1 Score	0.949580	0.102941	0.032520	0.047244	0.114286	0.940678	0.015873	0.102190

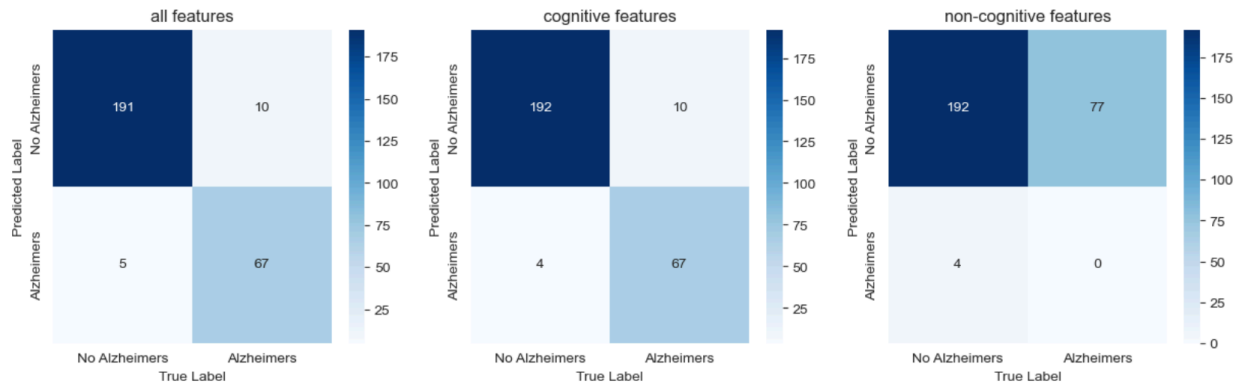
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.945055	0.710623	0.714286	0.725275	0.695971	0.948718	0.717949	0.703297
Precision Score	0.870130	0.025974	0.038961	0.038961	0.012987	0.870130	0.000000	0.000000
Recall Score	0.930556	0.333333	0.428571	0.750000	0.125000	0.943662	0.000000	0.000000
F1 Score	0.899329	0.048193	0.071429	0.074074	0.023529	0.905405	0.000000	0.000000

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

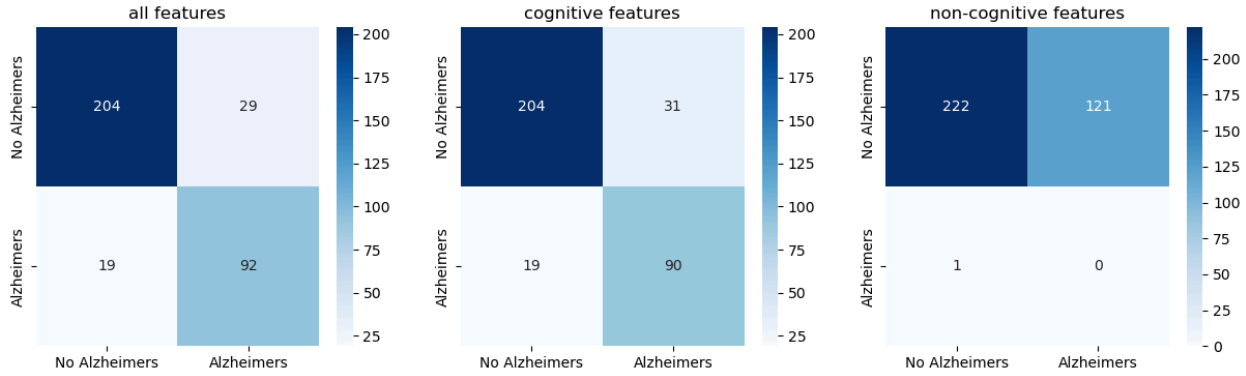
Logistic Regression

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.860465	0.648256	0.648256	0.648256	0.648256	0.854651	0.648256	0.645349
Precision Score	0.760331	0.000000	0.000000	0.000000	0.000000	0.743802	0.000000	0.000000
Recall Score	0.828829	0.000000	0.000000	0.000000	0.000000	0.825688	0.000000	0.000000
F1 Score	0.793103	0.000000	0.000000	0.000000	0.000000	0.782609	0.000000	0.000000

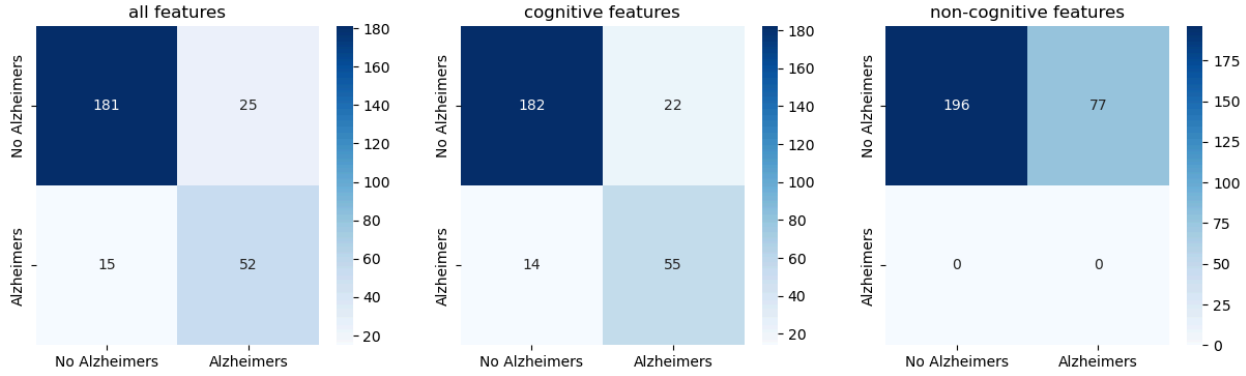
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.853480	0.717949	0.717949	0.717949	0.717949	0.868132	0.717949	0.717949
Precision Score	0.675325	0.000000	0.000000	0.000000	0.000000	0.714286	0.000000	0.000000
Recall Score	0.776119	0.000000	0.000000	0.000000	0.000000	0.797101	0.000000	0.000000
F1 Score	0.722222	0.000000	0.000000	0.000000	0.000000	0.753425	0.000000	0.000000

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

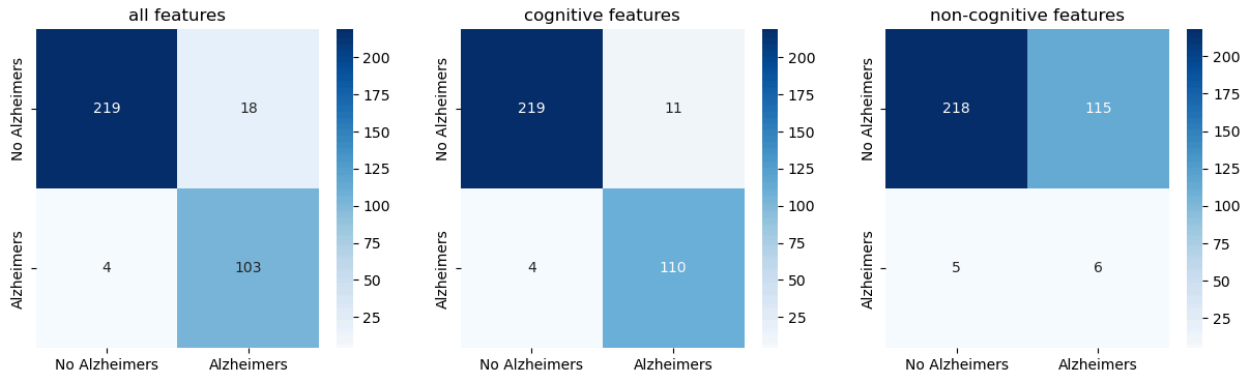
Random Forest Classifier

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.936047	0.508721	0.607558	0.648256	0.610465	0.956395	0.639535	0.651163
Precision Score	0.851240	0.297521	0.074380	0.024793	0.082645	0.909091	0.008264	0.049587
Recall Score	0.962617	0.300000	0.281250	0.500000	0.303030	0.964912	0.200000	0.545455
F1 Score	0.903509	0.298755	0.117647	0.047244	0.129870	0.936170	0.015873	0.090909

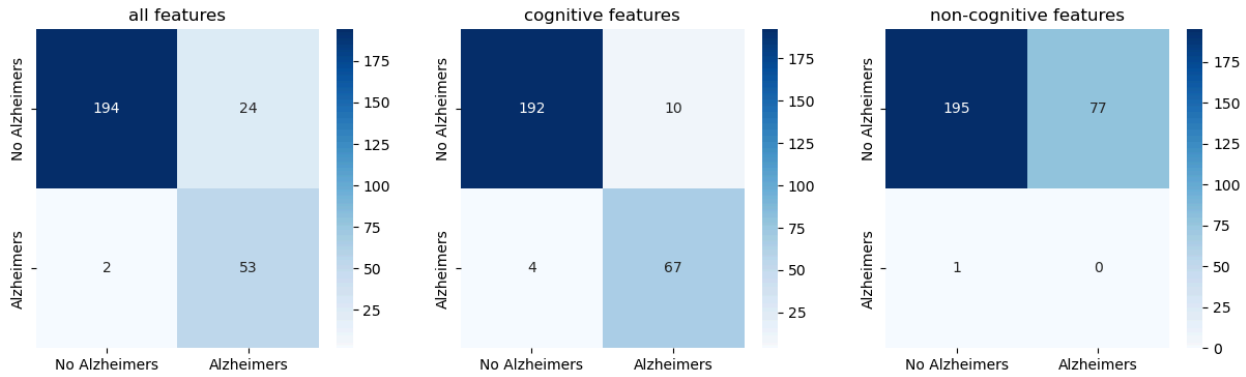
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.904762	0.622711	0.710623	0.695971	0.692308	0.948718	0.714286	0.714286
Precision Score	0.688312	0.168831	0.077922	0.038961	0.064935	0.870130	0.000000	0.000000
Recall Score	0.963636	0.250000	0.428571	0.250000	0.294118	0.943662	0.000000	0.000000
F1 Score	0.803030	0.201550	0.131868	0.067416	0.106383	0.905405	0.000000	0.000000

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

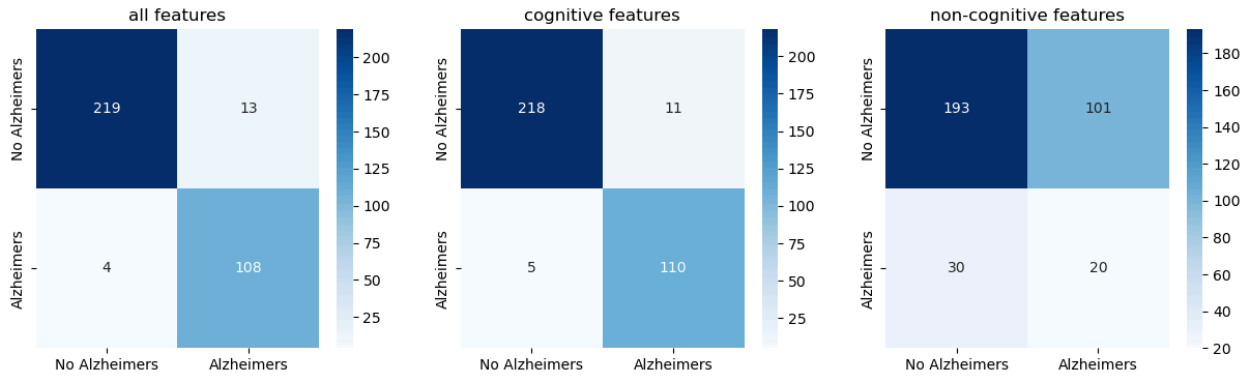
XGBoost Classifier

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.950581	0.552326	0.619186	0.645349	0.595930	0.953488	0.639535	0.619186
Precision Score	0.892562	0.280992	0.264463	0.041322	0.280992	0.909091	0.008264	0.165289
Recall Score	0.964286	0.336634	0.432432	0.454545	0.395349	0.956522	0.200000	0.400000
F1 Score	0.927039	0.306306	0.328205	0.075758	0.328502	0.932203	0.015873	0.233918

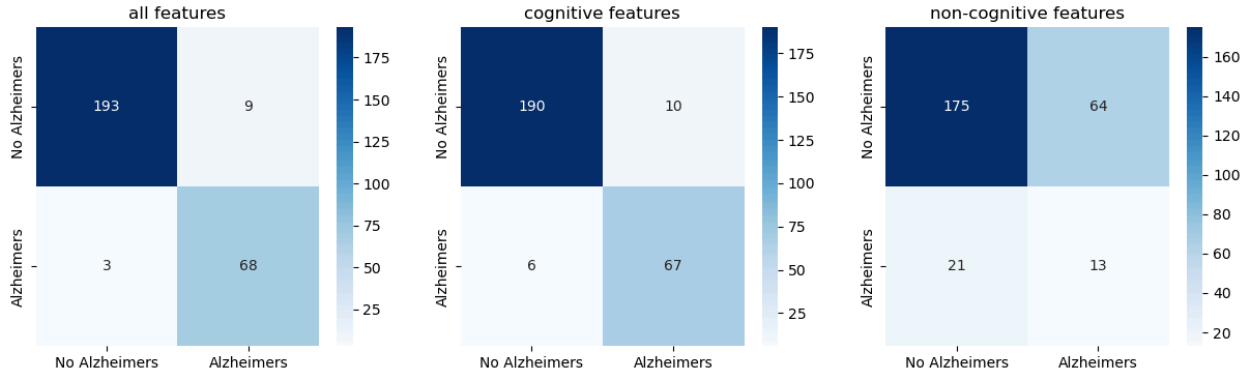
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.956044	0.681319	0.659341	0.695971	0.626374	0.941392	0.717949	0.688645
Precision Score	0.883117	0.155844	0.181818	0.038961	0.129870	0.870130	0.000000	0.168831
Recall Score	0.957746	0.352941	0.318182	0.250000	0.222222	0.917808	0.000000	0.382353
F1 Score	0.918919	0.216216	0.231405	0.067416	0.163934	0.893333	0.000000	0.234234

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

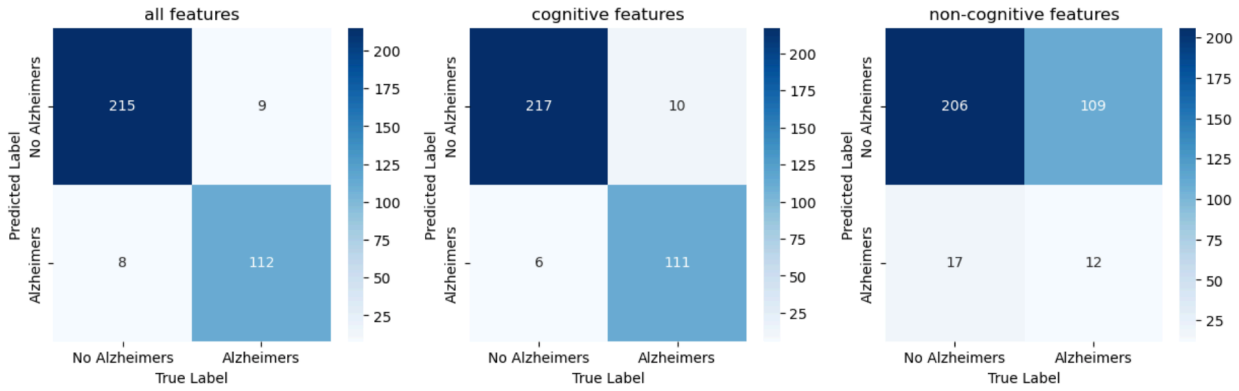
Gradient Boosting Classifier

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.950581	0.648256	0.639535	0.648256	0.625000	0.953488	0.639535	0.633721
Precision Score	0.925620	0.066116	0.066116	0.024793	0.123967	0.917355	0.008264	0.099174
Recall Score	0.933333	0.500000	0.421053	0.500000	0.394737	0.948718	0.200000	0.413793
F1 Score	0.929461	0.116788	0.114286	0.047244	0.188679	0.932773	0.015873	0.160000

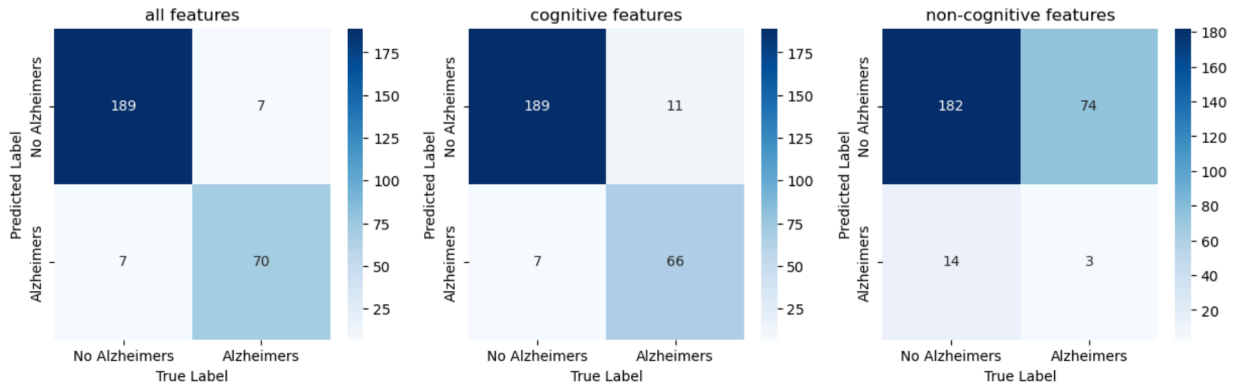
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.948718	0.688645	0.695971	0.714286	0.695971	0.934066	0.717949	0.677656
Precision Score	0.909091	0.012987	0.064935	0.000000	0.012987	0.857143	0.000000	0.038961
Recall Score	0.909091	0.100000	0.312500	0.000000	0.125000	0.904110	0.000000	0.176471
F1 Score	0.909091	0.022989	0.107527	0.000000	0.023529	0.880000	0.000000	0.063830

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

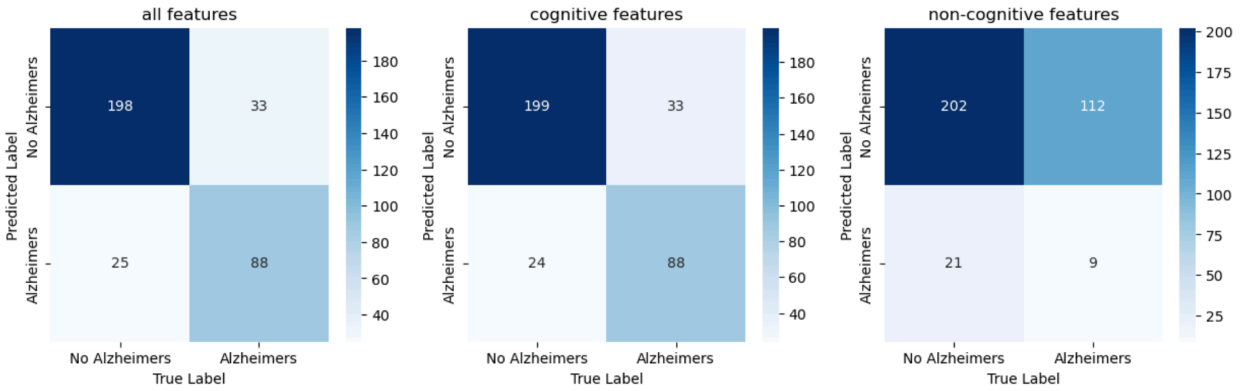
Gaussian Naïve Bayes

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.831395	0.648256	0.648256	0.633721	0.648256	0.834302	0.648256	0.613372
Precision Score	0.727273	0.000000	0.000000	0.024793	0.000000	0.727273	0.000000	0.074380
Recall Score	0.778761	0.000000	0.000000	0.272727	0.000000	0.785714	0.000000	0.300000
F1 Score	0.752137	0.000000	0.000000	0.045455	0.000000	0.755365	0.000000	0.119205

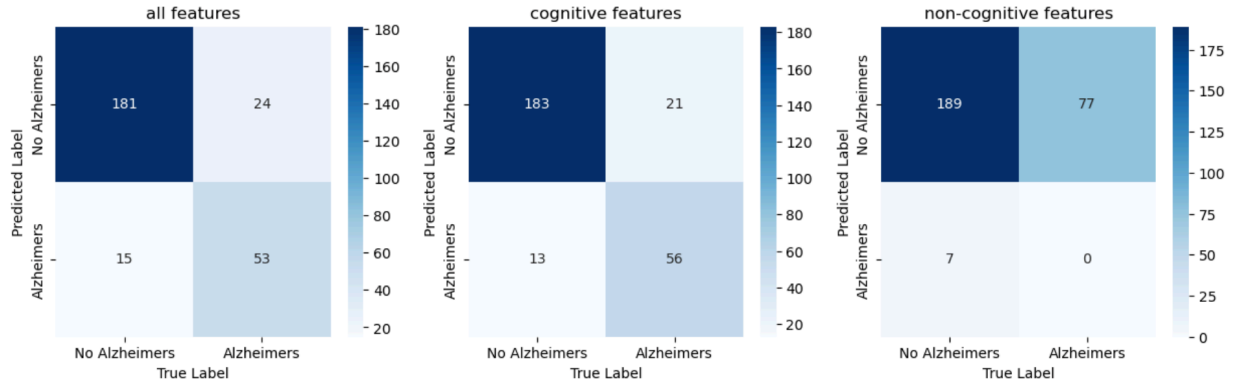
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.857143	0.717949	0.717949	0.717949	0.717949	0.875458	0.717949	0.692308
Precision Score	0.688312	0.000000	0.000000	0.000000	0.000000	0.727273	0.000000	0.000000
Recall Score	0.779412	0.000000	0.000000	0.000000	0.000000	0.811594	0.000000	0.000000
F1 Score	0.731034	0.000000	0.000000	0.000000	0.000000	0.767123	0.000000	0.000000

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

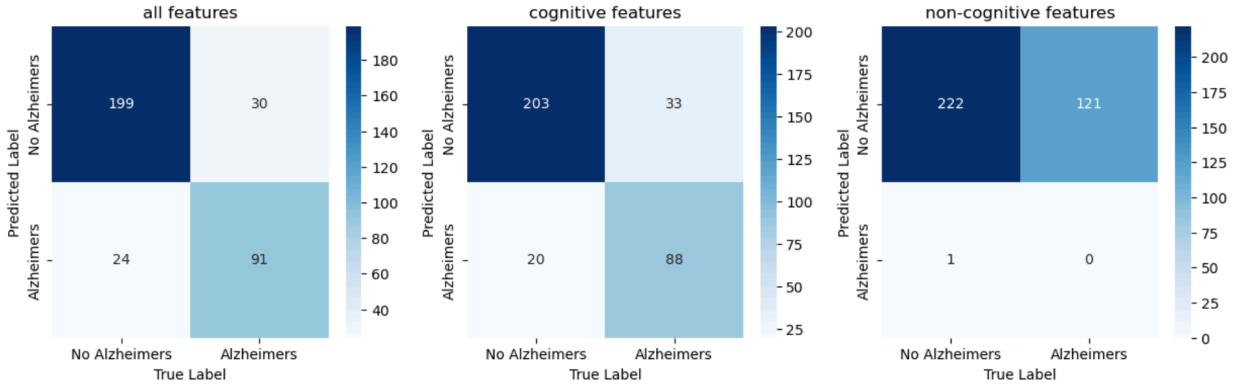
Linear Discriminant Analysis

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.843023	0.648256	0.648256	0.648256	0.648256	0.845930	0.648256	0.645349
Precision Score	0.752066	0.000000	0.000000	0.000000	0.000000	0.727273	0.000000	0.000000
Recall Score	0.791304	0.000000	0.000000	0.000000	0.000000	0.814815	0.000000	0.000000
F1 Score	0.771186	0.000000	0.000000	0.000000	0.000000	0.768559	0.000000	0.000000

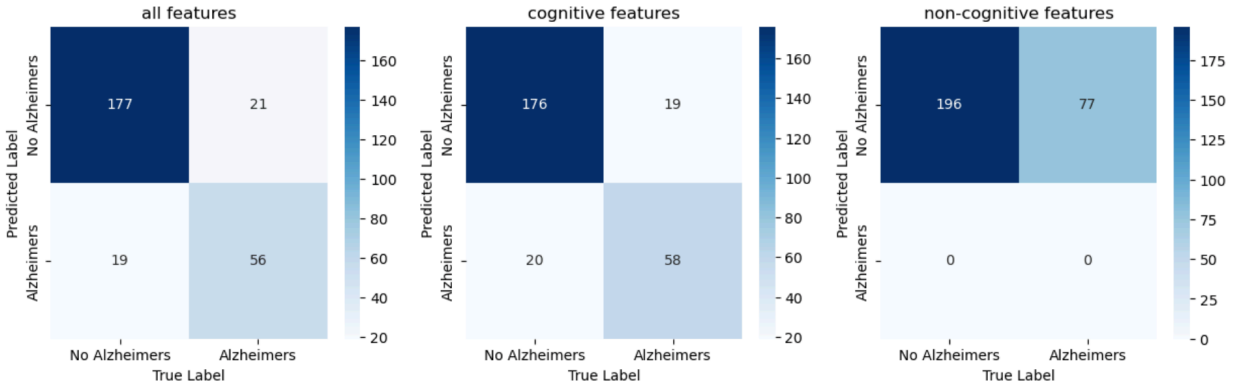
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.853480	0.717949	0.717949	0.717949	0.717949	0.857143	0.717949	0.717949
Precision Score	0.727273	0.000000	0.000000	0.000000	0.000000	0.753247	0.000000	0.000000
Recall Score	0.746667	0.000000	0.000000	0.000000	0.000000	0.743590	0.000000	0.000000
F1 Score	0.736842	0.000000	0.000000	0.000000	0.000000	0.748387	0.000000	0.000000

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

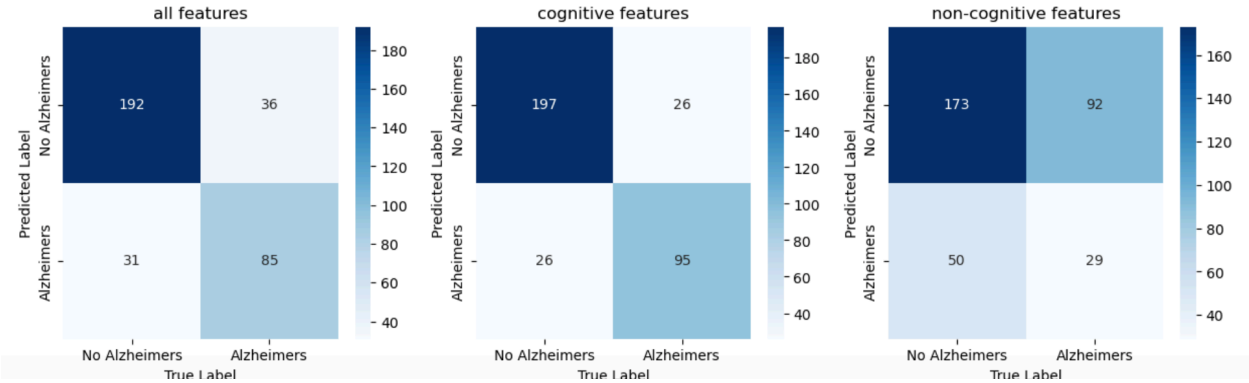
Quadratic Discriminant Analysis

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.843023	0.648256	0.648256	0.648256	0.648256	0.845930	0.648256	0.645349
Precision Score	0.752066	0.000000	0.000000	0.000000	0.000000	0.727273	0.000000	0.000000
Recall Score	0.791304	0.000000	0.000000	0.000000	0.000000	0.814815	0.000000	0.000000
F1 Score	0.771186	0.000000	0.000000	0.000000	0.000000	0.768559	0.000000	0.000000

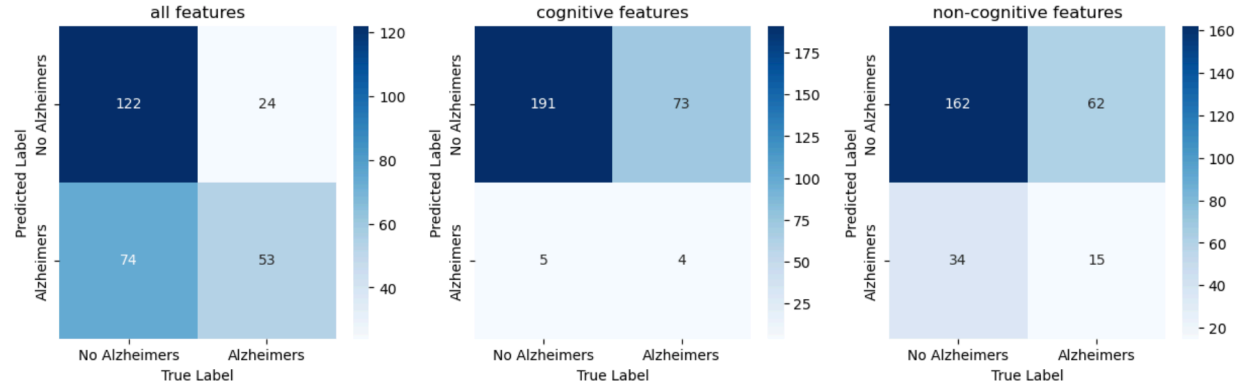
Metrics: Entire dataset

	all_features	demo	lifestyle	medic	clinical	cognitive	symptoms	no_cognitive
Accuracy Score	0.805233	0.648256	0.636628	0.636628	0.651163	0.848837	0.648256	0.587209
Precision Score	0.702479	0.000000	0.008264	0.049587	0.049587	0.785124	0.000000	0.239669
Recall Score	0.732759	0.000000	0.166667	0.375000	0.545455	0.785124	0.000000	0.367089
F1 Score	0.717300	0.000000	0.015748	0.087591	0.090909	0.785124	0.000000	0.290000

Metrics: Dataset restricted to patients with no memory complaints



Confusion matrices: Entire dataset



Confusion matrices: Dataset restricted to patients with no memory complaints

DISCUSSION OF RESULTS

The metrics for when we use the entire dataset seem to be slightly better than when the dataset is restricted to only records of patients with no memory complaints. There was only one case (the recall for QDA) where the difference was significant; except for that, the difference between the metrics for the two datasets were somewhat minute. This was interesting to look at as intuitively memory complaints is one main characteristic of Alzheimer's disease, so it is interesting to see that a restrictive dataset in that regard did not outperform the entire dataset. The table below gives the comparison between using all the dataset and the dataset for only patients with no memory complaints. This comparison was made for when we use all the features.

MODEL	A1	A2	P1	P2	R1	R2	F1	F2
Logistic regression	0.8605	0.85348	0.7603	0.67533	0.829	0.776119	0.7931	0.722222
Random Forest	0.936	0.90476	0.8512	0.68831	0.9623	0.963636	0.9304	0.803030
Gradient boosting	0.9506	0.94872	0.9256	0.90909	0.9333	0.909091	0.9295	0.909091
XGBoost	0.9506	0.95604	0.8926	0.88312	0.9643	0.957746	0.927	0.918919
Adaboost	0.9651	0.94506	0.8678	0.87013	0.9130	0.930556	0.8898	0.899329
KNN	0.7297	0.76923	0.3141	0.25974	0.7917	0.769231	0.4498	0.388350
SVM (Poly Kernel)	0.7820	0.83883	0.5455	0.49351	0.7674	0.883721	0.6377	0.633333
Naïve Bayes	0.8314	0.85714	0.7273	0.68831	0.7788	0.779412	0.7521	0.731034
LDA	0.843	0.85348	0.752	0.72727	0.7913	0.746667	0.7712	0.736842
QDA	0.8052	0.64103	0.7025	0.68831	0.7328	0.417323	0.7173	0.519608

*Dataset for patients with no memory complaints have better score
*Significant difference in metric scores for when using entire dataset vs dataset with only patients with no memory complaints
1=entire data set; 2= data set with only patients with no memory complaints; A = accuracy; P=precision R = recall F= f1 score

Another interesting observation was that the sub categories that gave high accuracy were when we used all features and when we used only cognitive features. This cut across all models discussed and for both scenarios: using the entire dataset or the data set restricted on patients with no memory complaints.