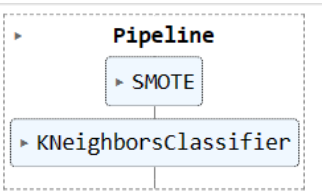
Report classifier

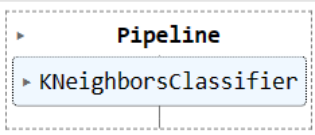
KNN on dataset MIT-BIHArrhythmiaDatabase.csv

Pipeline	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
pipe_normal	0.897752	0.069725	0.394468	0.439789
pipe_smote	0.790652	0.094975	0.364399	0.384815
pipe_PCA	0.906342	0.246932	0.464344	0.481820
pipe_f30	0.794358	0.083933	0.370107	0.381427
pipe_f19	0.794009	0.137183	0.347129	0.394117

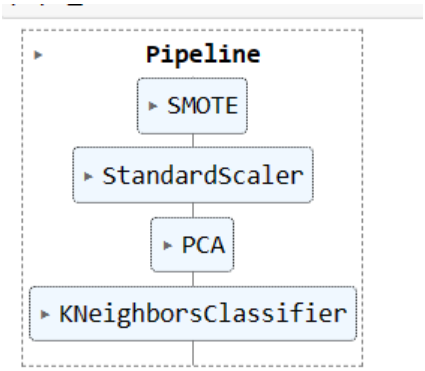
- **Pipe_smote →**



- **Pipe_normal →**



- **Pipe_PCA →**



- **Pipe_f30 →**

```
pipe_smote = Pipeline([
        ('f_classif', SelectKBest(mutual_info_classif, k=30)), #
        ('sampling', SMOTE(random_state = 123)),
        ('classification', KNeighborsClassifier(n_neighbors=10))
    ])
pipe_smote
```

- **Pipe_f19 →**

```
pipe_smote_19 = Pipeline([
        ('f_classif', SelectKBest(mutual_info_classif, k=19)), #
        ('sampling', SMOTE(random_state = 123)),
        ('classification', KNeighborsClassifier(n_neighbors=10))
    ])
pipe_smote_19
```

Wilcoxon Test on **macro avg** F1-score

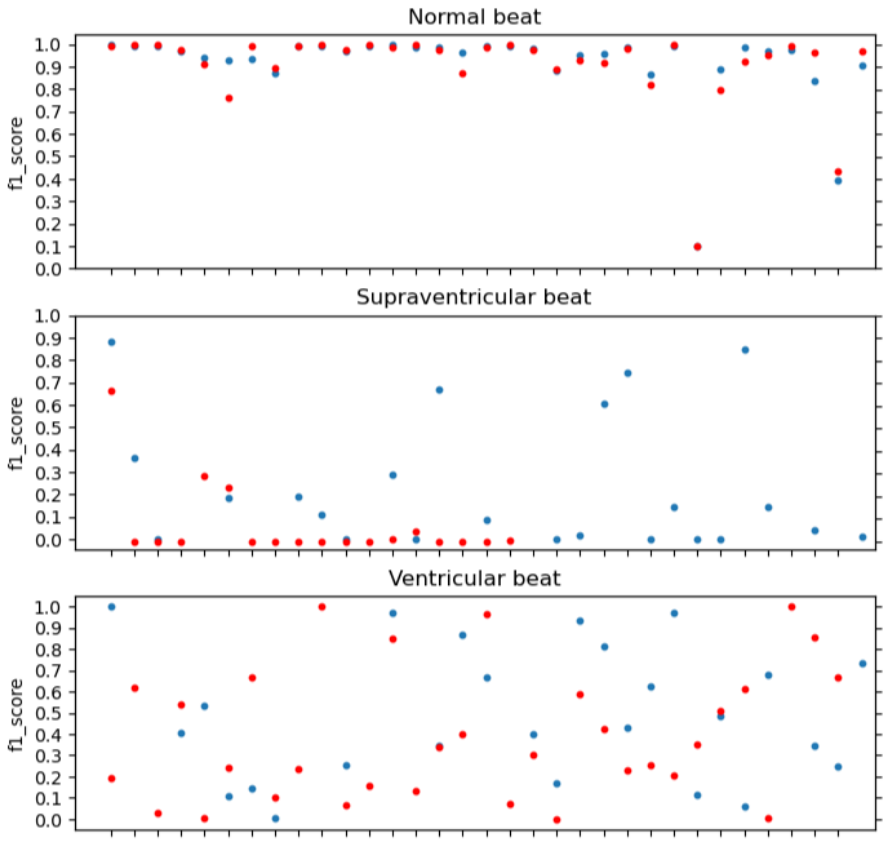
pipe 1	**pipe 2**	**p-value**
Pipe normal	Pipe smote	0.0037
Pipe smote	Pipe PCA	0.000035
Pipe normal	Pipe PCA	0.0201
Pipe_f30	Pipe_PCA	0.000025
Pipe_f19	Pipe_f30	0.733
Pipe_normal	Pipe_f19	0.07

From the classification report the PCA pipeline seems the best, and applying Wilcoxon test we see that all the other classification pipeline are not significantly different, but there is a significant difference between the classification with pipe_PCA and all the other, and also between the normal pipeline and the pipeline with only smote. We can say that the normal pipeline is better than the pipe_smote, but the pipe_PCA is better than the normal one.

But the f1-score macro average doesn't give us the complete information about the classifier, because is more important the f1-score on the two minority classes which represent the heart anomalies.

In the following image is possible compare the f1-score on each patient (on the x-axis) for each class. This make understand better if a pipeline really works better than another for each class.

In this case we compare the pipe_normal with the pipe_PCA. It is visible especially for the SVEB class that the pipe_PCA is better.



AdaBoost on dataset MIT-BIHArrhythmiaDatabase.csv

Pipeline	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
pipe_normal	0.936992	0.199248	0.555593	0.530120
pipe_smote	0.867423	0.163976	0.474478	0.449714
pipe_PCA	0.821393	0.126328	0.430322	0.412356
pipe_f30	0.870194	0.152633	0.498226	0.461365
pipe_f19	0.869095	0.163587	0.424841	0.443797

Same pipe than KNN case, but obviously with AdaBoost instead of KNN

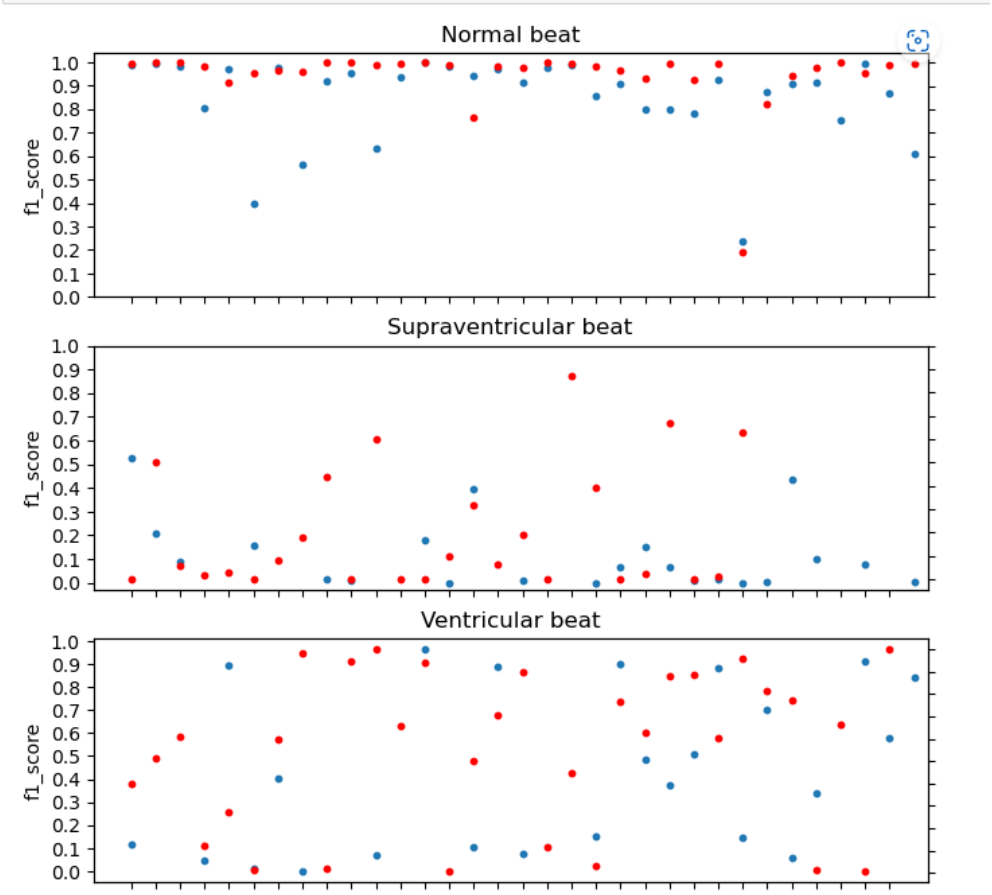
Wilcoxon Test on **macro avg** F1-score

pipe 1	**pipe 2**	**p-value**
Pipe normal	Pipe smote	0.0073
Pipe smote	Pipe PCA	0.2435
Pipe normal	Pipe PCA	0.00026

Pipe_f30	Pipe_smote	0.477
Pipe_f19	Pipe_smote	0.894
Pipe_normal	Pipe_f30	0.0044

From the classification report the normal pipeline seems the best, and applying Wilcoxon test we see that all the other classification pipeline are not significantly different, but there is a significant difference between the classification with pipe_normal and all the other. So, we can conclude that for AdaBoost the best pipeline is the simplest one.

In this case we compare the pipe_normal with the pipe_PCA. It is visible especially for the SVEB class that the pipe normal is better, but also in the Ventricular beat class the number of red dots above the blue ones are more.



DecisionTree on dataset MIT-BIHArrhythmiaDatabase.csv

Pipeline	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
pipe_normal	0.911661	0.141635	0.404338	0.446878
pipe_smote	0.864459	0.134598	0.346011	0.407286
pipe_PCA	0.868581	0.103693	0.304521	0.390795
pipe_f30	0.868581	0.103693	0.304521	0.390795
pipe_f19	0.873182	0.179294	0.357913	0.424004

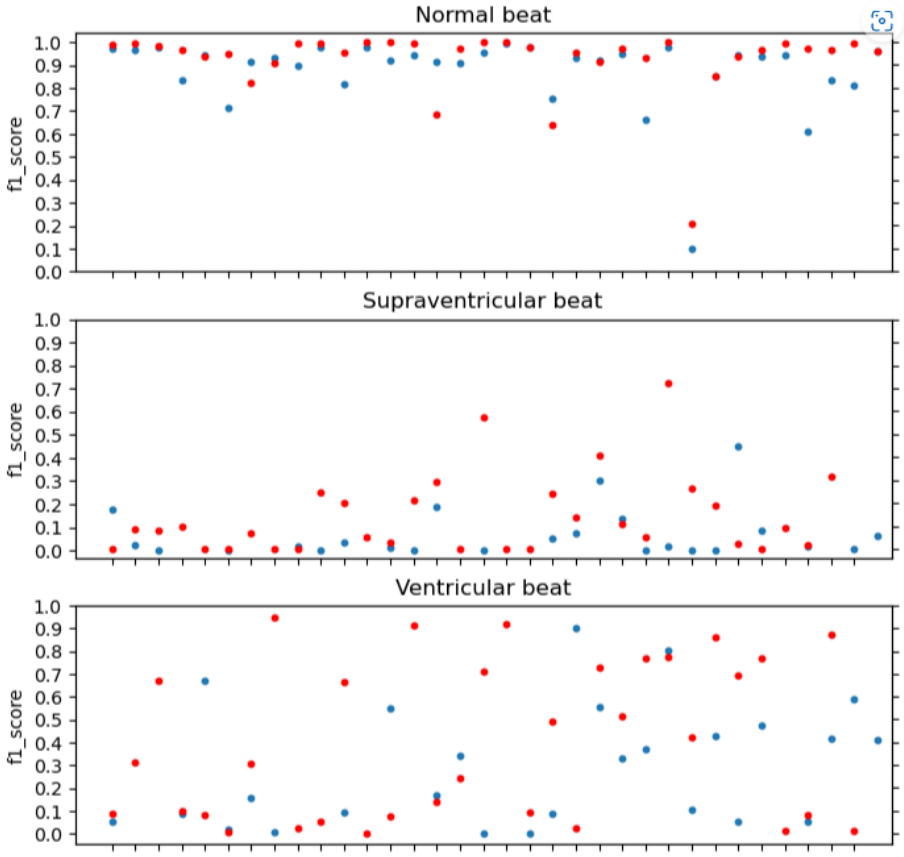
Same pipe than KNN case, but obviously with DecisionTree instead of KNN

Wilcoxon Test on **macro avg** F1-score

pipe 1	**pipe 2**	**p-value**
Pipe normal	Pipe smote	0.0208
Pipe smote	Pipe PCA	0.1576
Pipe normal	Pipe PCA	0.00261
Pipe_f30	Pipe_smote	0.462
Pipe_f19	Pipe_smote	0.414
Pipe_normal	Pipe_f30	0.147

From the classification report the normal pipeline seems the best, and applying Wilcoxon test we see that all the other classification pipeline are not significantly different, but there is a significant difference between the classification with pipe_normal and some of the other. So, we can conclude that for DecisionTree the best pipeline is the simplest one.

But the f1-score macro average doesn't give us the complete information about the classifier, because is more important the f1-score on the two minority classes which represent the heart anomalies.



In the previous image is possible compare the f1-score on each patient (on the x-axis) for each class. This make understand better if a pipeline really works better than another for each class.

In this case we compare the pipe_normal with the pipe_PCA. It is visible especially for the SVEB class that the pipe normal is better.

RandomForest on dataset MIT-BIHArrhythmiaDatabase.csv

Pipeline	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
pipe_normal	0.946720	0.285172	0.600983	0.563359
pipe_smote	0.934377	0.297163	0.591672	0.535853
pipe_PCA	0.934581	0.231283	0.556517	0.535899
pipe_f30	0.933537	0.309624	0.559328	0.559543

Same pipe than KNN case, but obviously with RandomForest instead of KNN

Wilcoxon Test on **macro avg** F1-score

pipe 1	**pipe 2**	**p-value**
Pipe normal	Pipe smote	0.522
Pipe normal	Pipe PCA	0.209
Pipe normal	Pipe f30	0.965

Applying Wilcoxon test we can see that all the classification pipelines are not significantly different from the simplest one. The improvement brought by the other pipelines are not significant, so the simplest classification pipeline should be chosen.

Compare classifiers results on MIT-BIHArrhythmiaDatabase.csv

Classifier	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
RandomForest	0.946720	0.285172	0.600983	0.563359
AdaBoost	0.936992	0.199248	0.555593	0.530120
KNN	0.906342	0.246932	0.464344	0.481820
DecisionTree	0.911661	0.141635	0.404338	0.446878

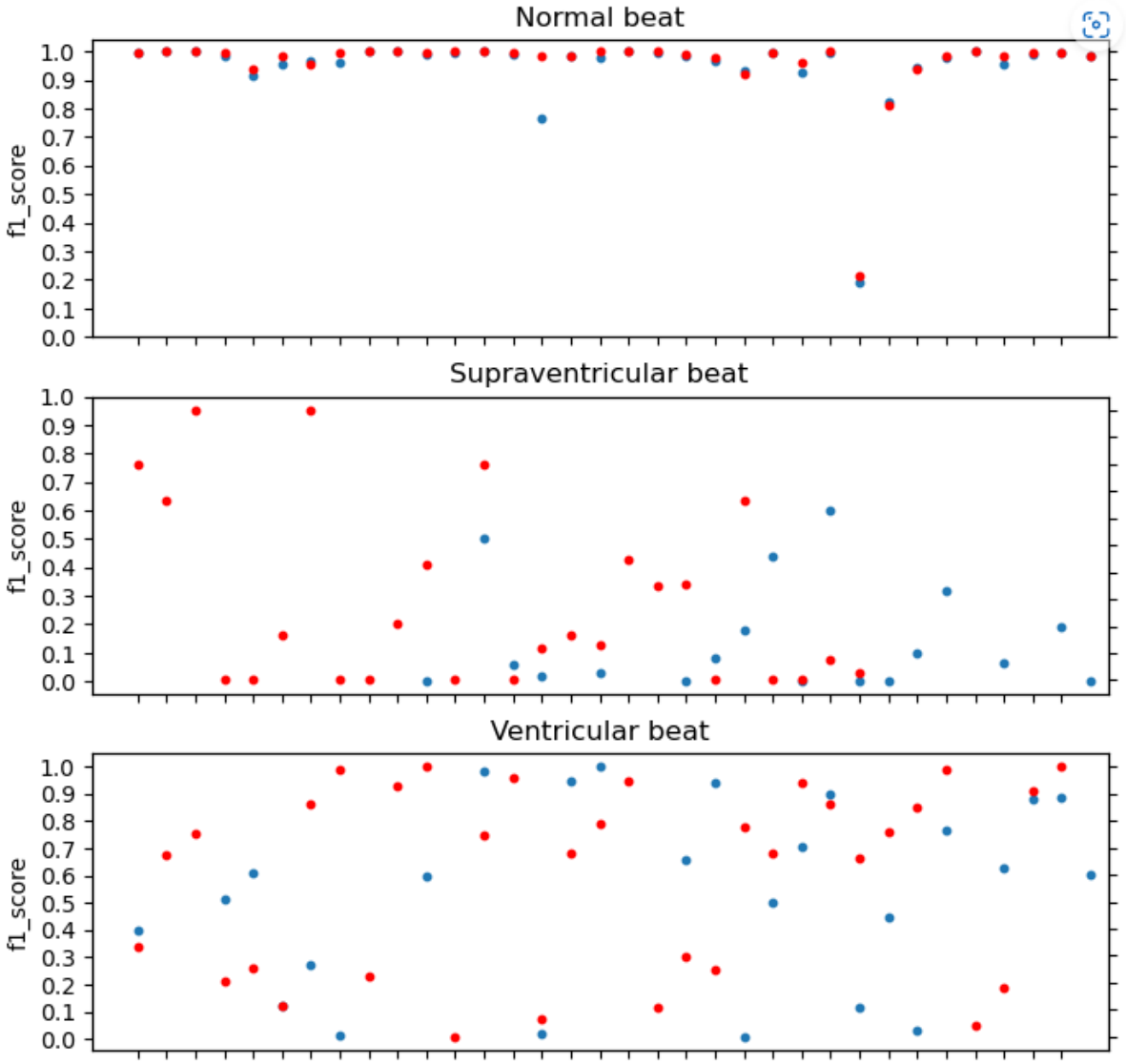
Wilcoxon Test on **macro avg** F1-score

Classifier 1	**Classifier 2**	**p-value**
KNN	DecisionTree	0.2076
KNN	AdaBoost	0.04
KNN	RandomForest	0.007
AdaBoost	RandomForest	0,12
DecisionTree	AdaBoost	0.0005
DecisionTree	RandomForest	0,0000027

Computational time for the best pipeline of each classifier

Classifier	Time
KNN	151 sec
RandomForest	2661 sec
AdaBoost	883 sec
DecisionTree	379 sec

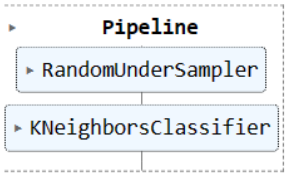
The two best classifiers from the Wilcoxon test and the classification report seems to be AdaBoost and RandomForest, seeing the computational report AdaBoost has a very lower time of execution than RandomForest, but comparing the f1-score of each of the two classifiers the RandomForest models is better on the Supraventricular beat class.



Comparison between different datasets and complete dataset (the four datasets merged together)

KNN on dataset INCART2-leadArrhythmiaDatabase.csv

On this dataset I tried also the following pipeline



Pipeline	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
pipe_normal	0.985236	0.440093	0.785343	0.711076
pipe_smote	0.978471	0.254832	0.770841	0.622649
pipe undersamp	0.925827	0.175294	0.587798	0.545243
pipe_PCA	0.983127	0.353399	0.850980	0.687838
pipe_f30	0.978584	0.255145	0.772441	0.618888
pipe_f19	0.980106	0.252890	0.776002	0.649595

The undersampling rebalancing approach works worse than the smote approach.

Wilcoxon Test on **macro avg** F1-score

pipe 1	**pipe 2**	**p-value**
Pipe normal	Pipe smote	0.00073
Pipe normal	Pipe PCA	0.8
Pipe normal	Pipe_f30	0.00027
Pipe_normal	Pipe_f19	0.03
Pipe_PCA	Pipe_f19	0.003
Pipe_PCA	Pipe_f30	0.0000526
Pipe_PCA	Pipe smote	0,0000014

From the f1-score report the two best classifications pipeline seems to be the pipe normal and the pipe PCA, and the Wilcoxon test confirms that there is a significantly difference between the two quoted pipelines and all the others.

There is no difference between the normal pipeline and the pipe PCA, so the simplest one should be chosen.

KNN on dataset MIT-BIH Supraventricular Arrhythmia Database.csv

Pipeline	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
pipe_normal	0.911370	0.225427	0.372760	0.486619
pipe_PCA	0.911202	0.376924	0.448321	0.548076

Wilcoxon Test on **macro avg** F1-score

pipe 1	**pipe 2**	**p-value**
Pipe normal	Pipe PCA	0.002

After Wilcoxon test we can say that the PCA pipeline is better than the normal one.

KNN on dataset Sudden Cardiac Death Holter Database.csv

Pipeline	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
pipe_normal	0.877491	0.000000	0.215906	0.383212
pipe_PCA	0.770739	0.003063	0.393542	0.376633

Wilcoxon Test on **macro avg** F1-score

pipe 1	**pipe 2**	**p-value**
Pipe normal	Pipe PCA	0.831

After Wilcoxon test we can say that there isn't significative difference between the two pipelines, but the pipe_PCA is better on the two minority classes.

Comparison between the four dataset and the complete dataset

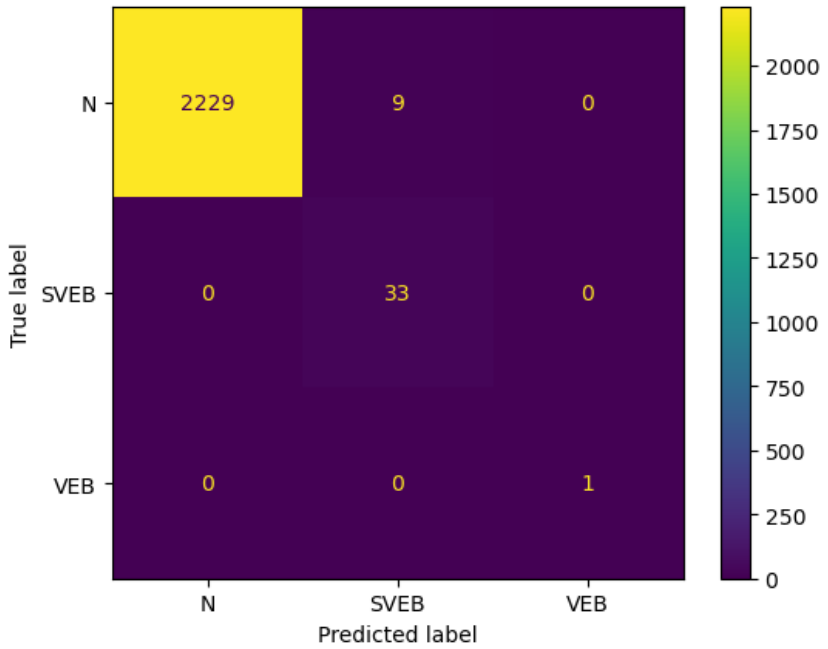
Using the best pipeline of KNN for each dataset

Dataset	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
MIT-BIHArrhythmiaDatabase	0.906342	0.246932	0.464344	0.481820
INCART2-leadArrhythmiaDatabase	0.985236	0.440093	0.785343	0.711076
MIT-BIH Supraventricular Arrhythmia Database	0.911202	0.376924	0.448321	0.548076
Sudden Cardiac Death Holter Database	0.770739	0.003063	0.393542	0.376633
4 datasets merged	0.927160	0.322954	0.629596	0.572556

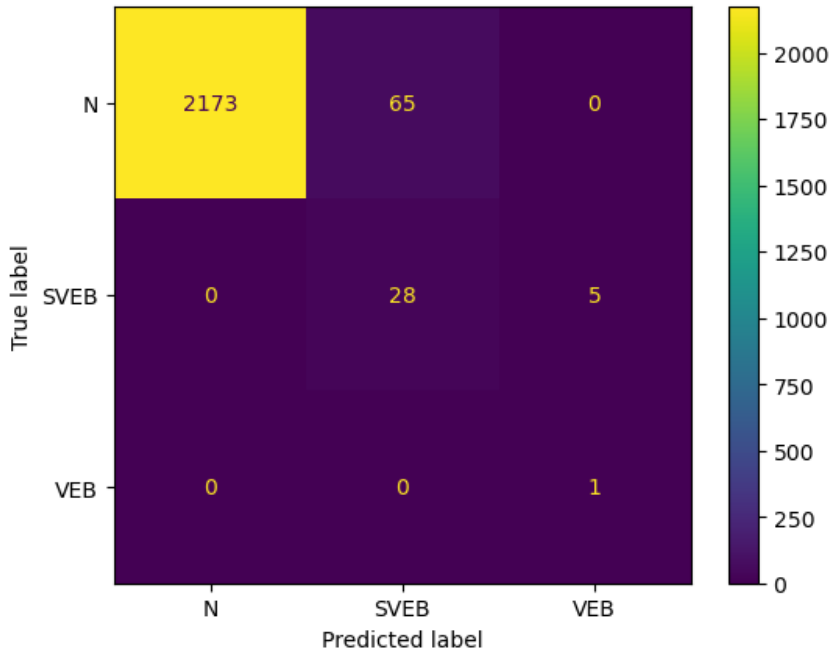
Comparison of performance of pipe PCA on the same patient in case of single dataset and complete one

Patient 1 of **MIT-BIHArrhythmiaDatabase**

Confusion matrix on this patient on single dataset

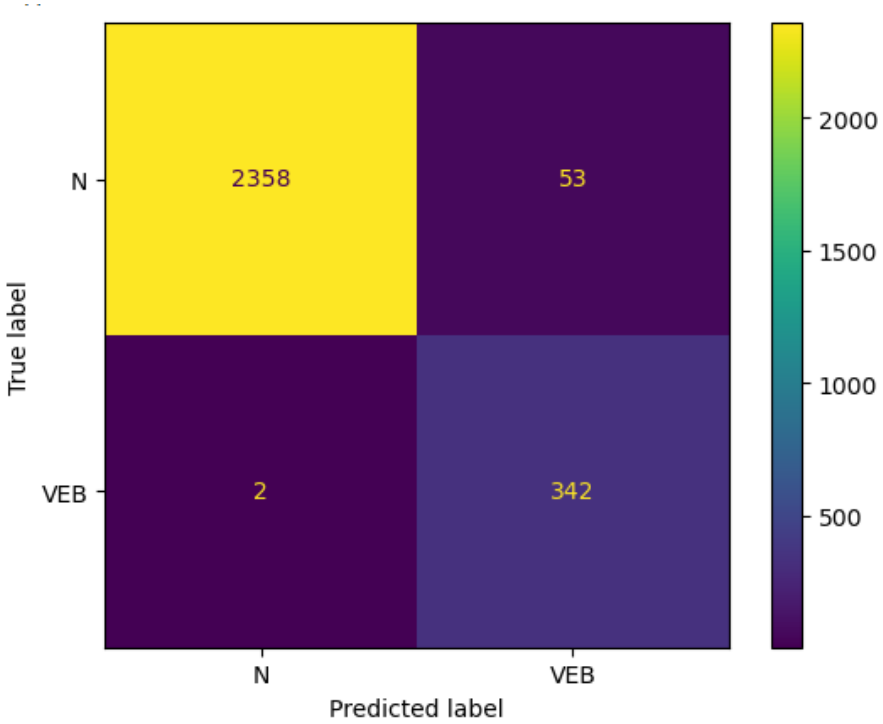


Confusion matrix on this patient on the complete dataset

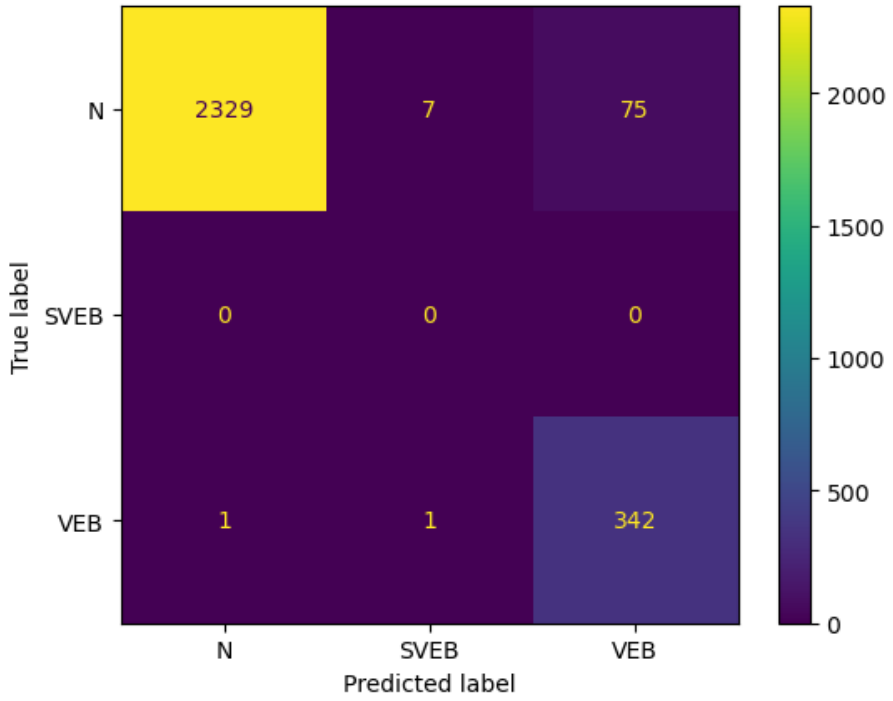


Patient 1 of **INCART2-leadArrhythmiaDatabase**

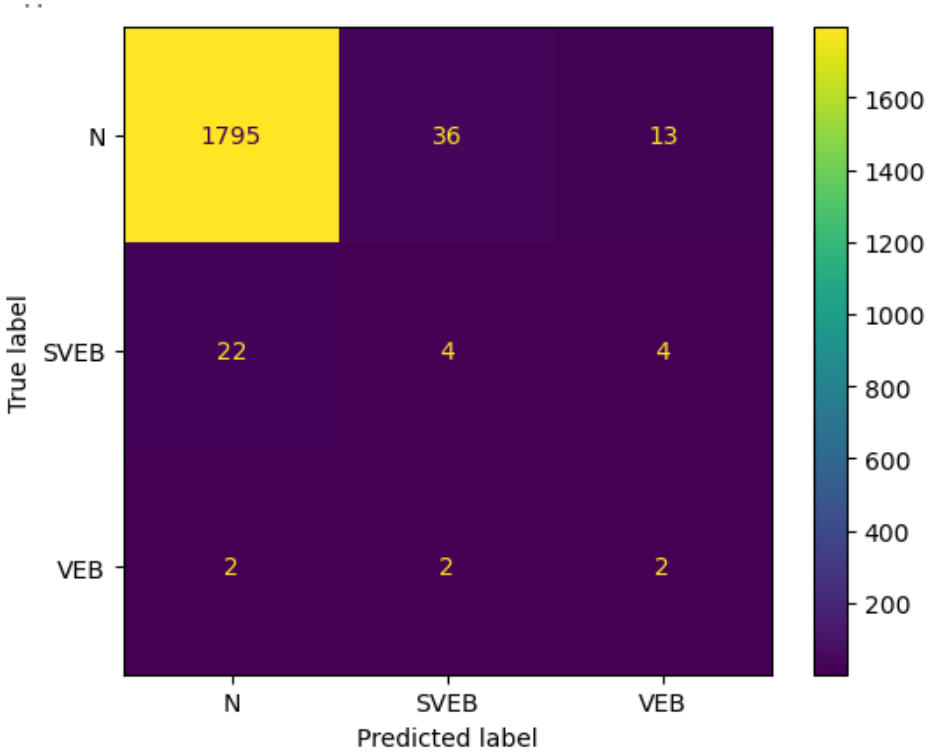
Confusion matrix on this patient on single dataset with pipe PCA



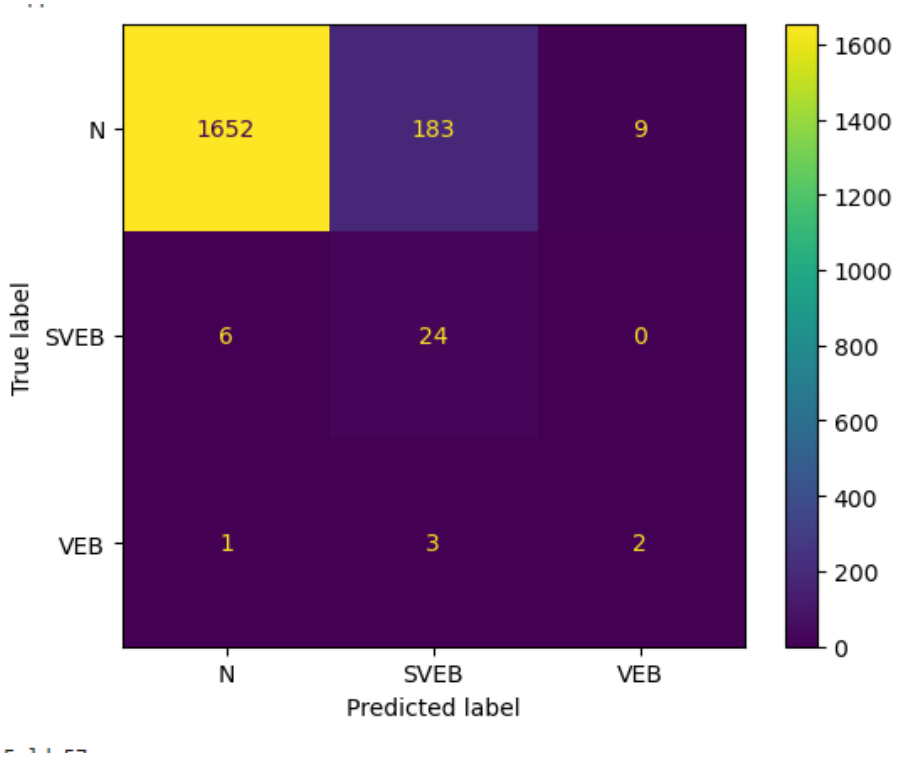
Confusion matrix on this patient on the complete dataset



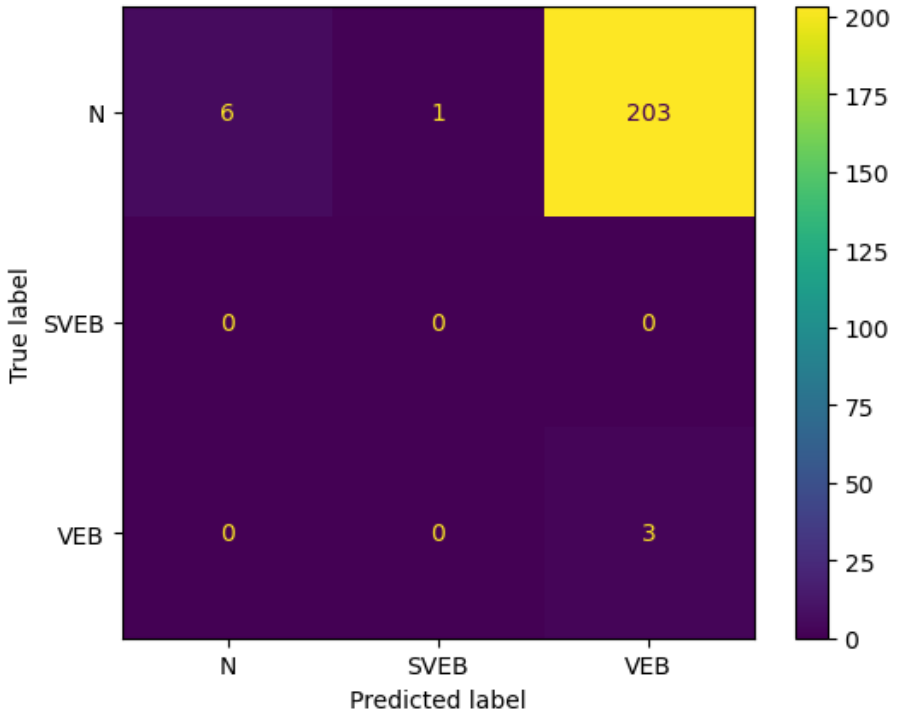
Patient 1 of **MIT-BIH Supraventricular Arrhythmia Database**



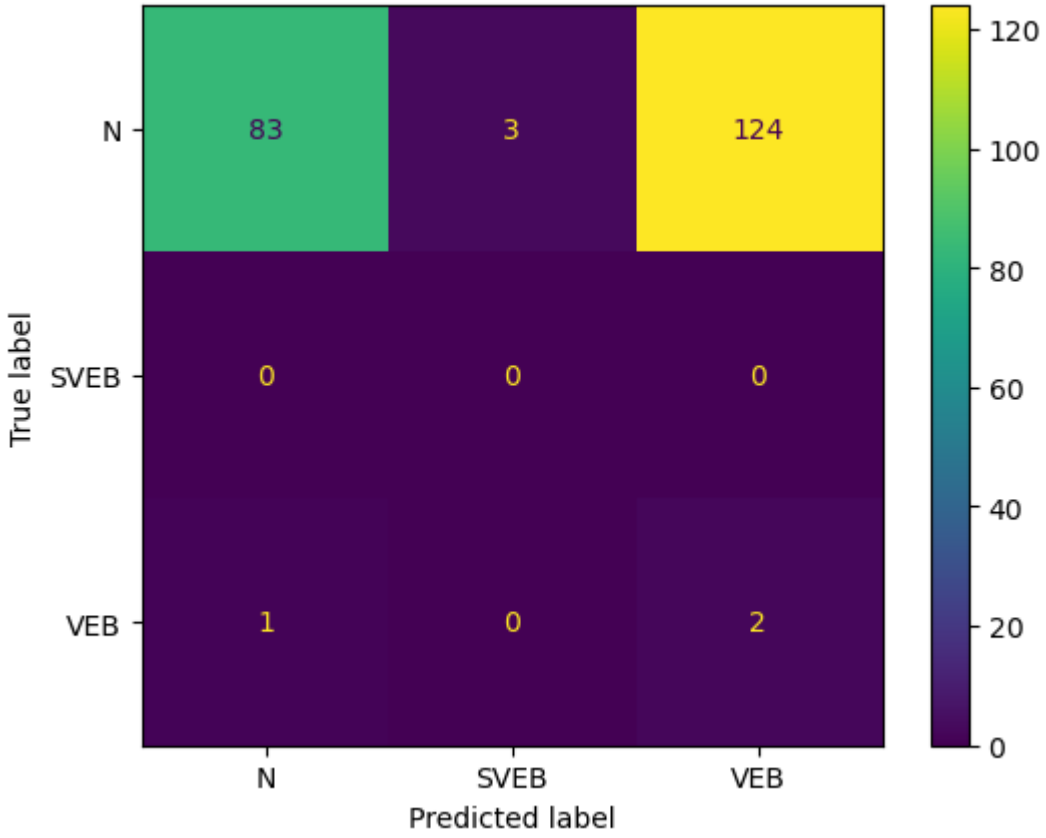
Confusion matrix on this patient on the complete dataset



Patient 0 of **Sudden Cardiac Death Holter Database**



Confusion matrix on this patient on the complete dataset



Comparison results on the entire dataset

Classifier	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
AdaBoost	0.966243	0.331620	0.639934	0.623943
KNN	0.927160	0.322954	0.629596	0.572556
DecisionTree	0.936271	0.249389	0.545603	0.528986

Except for KNN, all the classifiers have too great computational cost, I exclude RandomForest from the esperimentation because it was unreasonable.

AdaBoost seems to be the best approach.