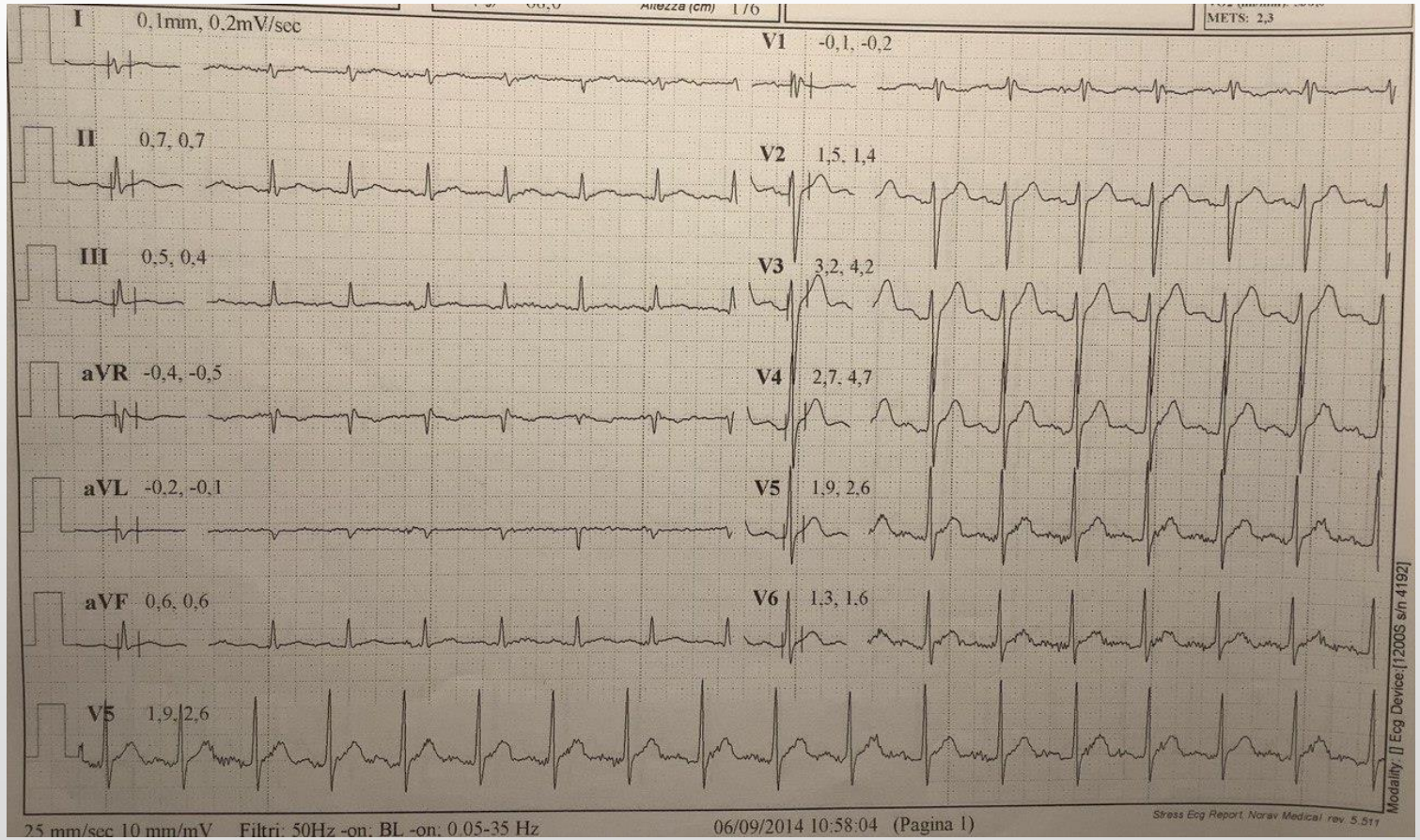
Heartbeat Classification

MATTEO RAZZAI

Project objective

- The goal of this project is to recognize the type of a patient's heartbeats. The beats are described by some features extracted reading two leads of the patient's ECG.
- The type of an heartbeat can be: Normal beat(N), Supraventricular ectopic beat (S), Ventricular ectopic beat(V), Fusion beat (fusion between N beat and V beat) and Unknown beat (Q)
- The detection of the heartbeat's type can be very useful for recognize case of arrhythmia.

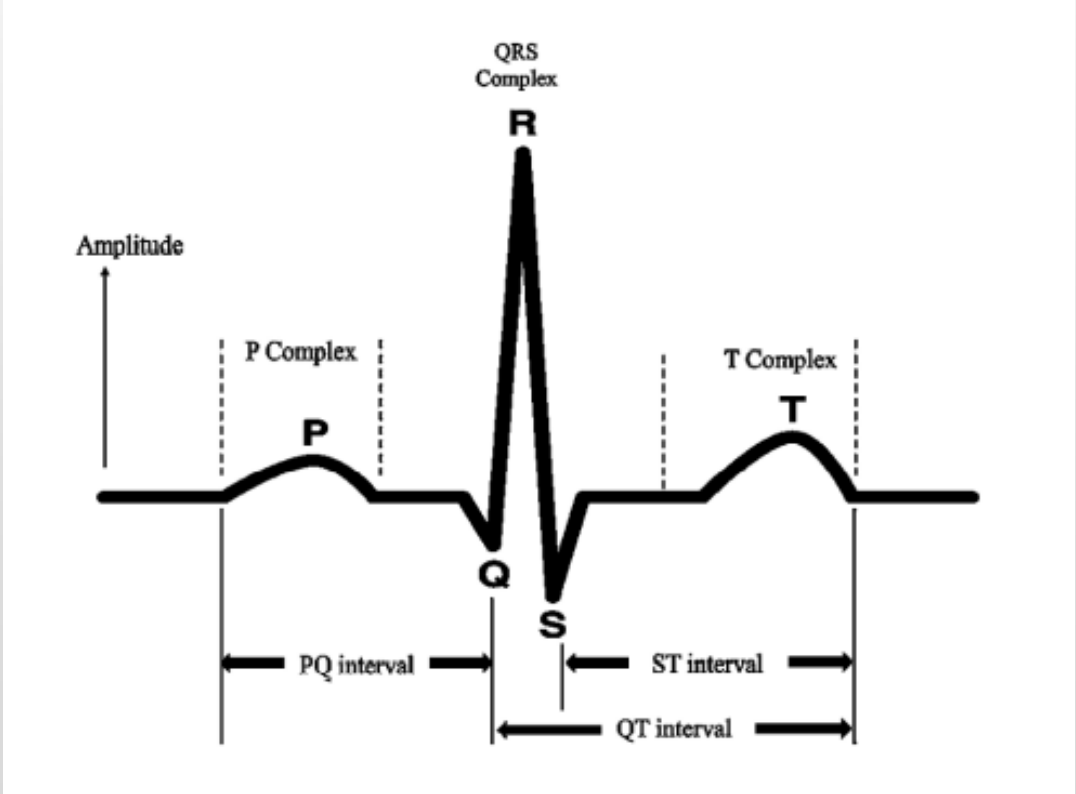


Dataset

- This dataset come from Holter tapes applied on more than 200 patients
- From these holter therapies come out 963654 samples, each of them regarding a single heartbeat
- There are 32 features, summarized in the following table, and other two regarding the heartbeat's type and the patient.

Feature Group	Lead II and V5
RR Intervals	Pre-RR, Post-RR
Heart beats interval	PQ, QT, ST interval QRS duration
Heart beats amplitude	P, T, R, S, Q peak
Morphology QRS	5 samples between onset and offset point of the QRS complex

Source:[ECG Arrhythmia Classification Dataset | Kaggle]



Datasets

Dataset	Number of records	Number of patients	Time of recording for patient	Number of Normal beats	Number of SVEB beats	Number of VEB beats	Number of Fusion beats	Number of Unknown beats
MIT-BIH ArrhythmiaDatabase	100689	44	30 min	90083	7009	2779	803	15
INCART2-leadArrhythmiaDatabase	175729	75	30 min	153546	1958	20000	219	6
MIT-BIH Supraventricular Arrhythmia Database	184428	78	30 min	162195	12194	9937	23	79
Sudden Cardiac Death Holter Database	426591	12	12 hours	403528	1609	14723	211	6520

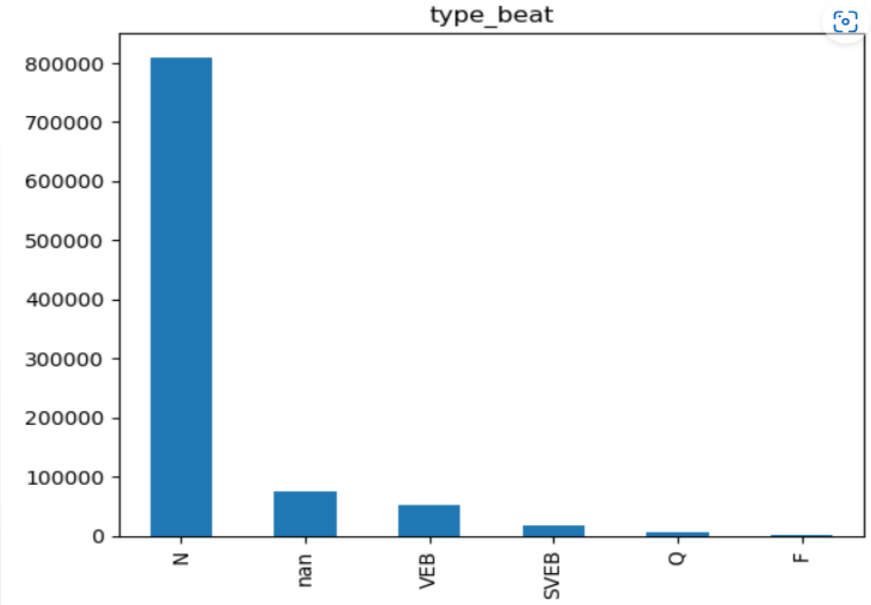
Preprocessing

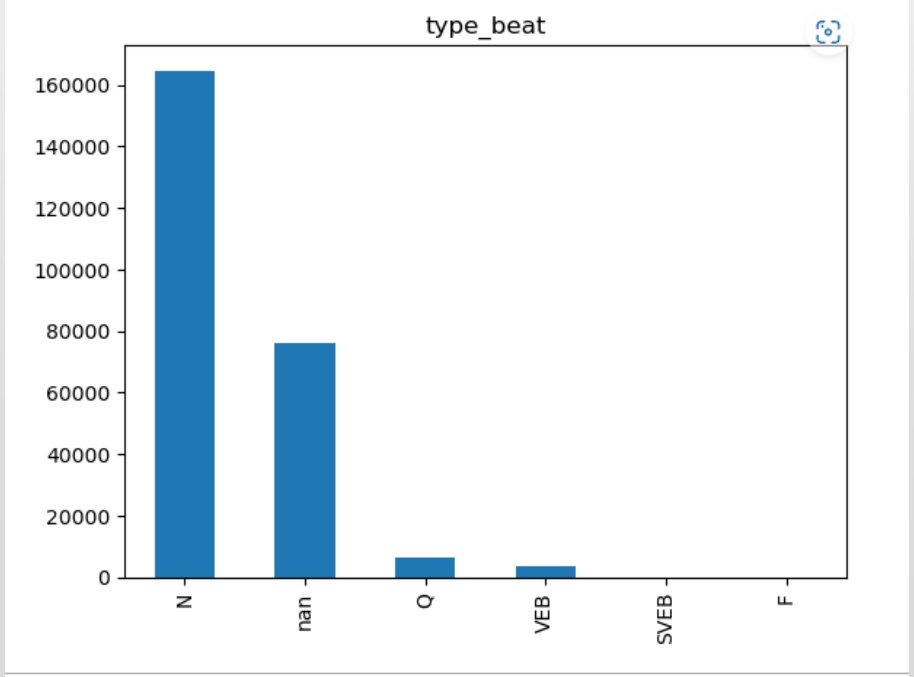
Handling with missing values

```
df_cardio.isna().sum()
```

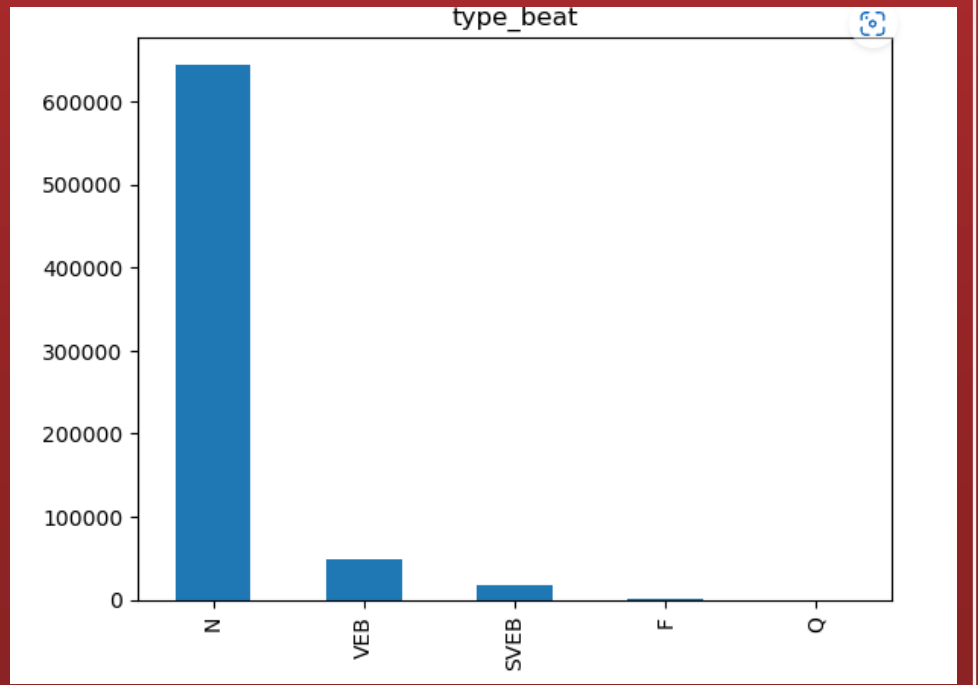
```
record              76217
type                76217
0_pre-RR            76217
0_post-RR           76217
0_pPeak             76217
0_tPeak             76217
0_rPeak             76217
0_sPeak             76217
0_qPeak             76217
0_qrs_interval      76217
0_pq_interval       76217
0_qt_interval       76217
0_st_interval       76217
0_qrs_morph0        76217
0_qrs_morph1        76217
0_qrs_morph2        76217
0_qrs_morph3        76217
0_qrs_morph4        76217
1_pre-RR           174615
1_post-RR          174615
1_pPeak            174615
1_tPeak            174615
1_rPeak            174615
1_sPeak            174615
1_qPeak            174615
1_qrs_interval     174615
1_pq_interval      174615
1_qt_interval      174615
1_st_interval      174615
1_qrs_morph0       174615
1_qrs_morph1       174615
1_qrs_morph2       174615
1_qrs_morph3       174615
1_qrs_morph4       174615
dtype: int64
```

Before





After



- Attribute ‘record’

```
df_cardio=df_cardio.rename(columns={'record':'patient'})
```

- Patient’s recording code

```
df_cardio['patient']=df_cardio['patient'].astype('category').cat.codes
```

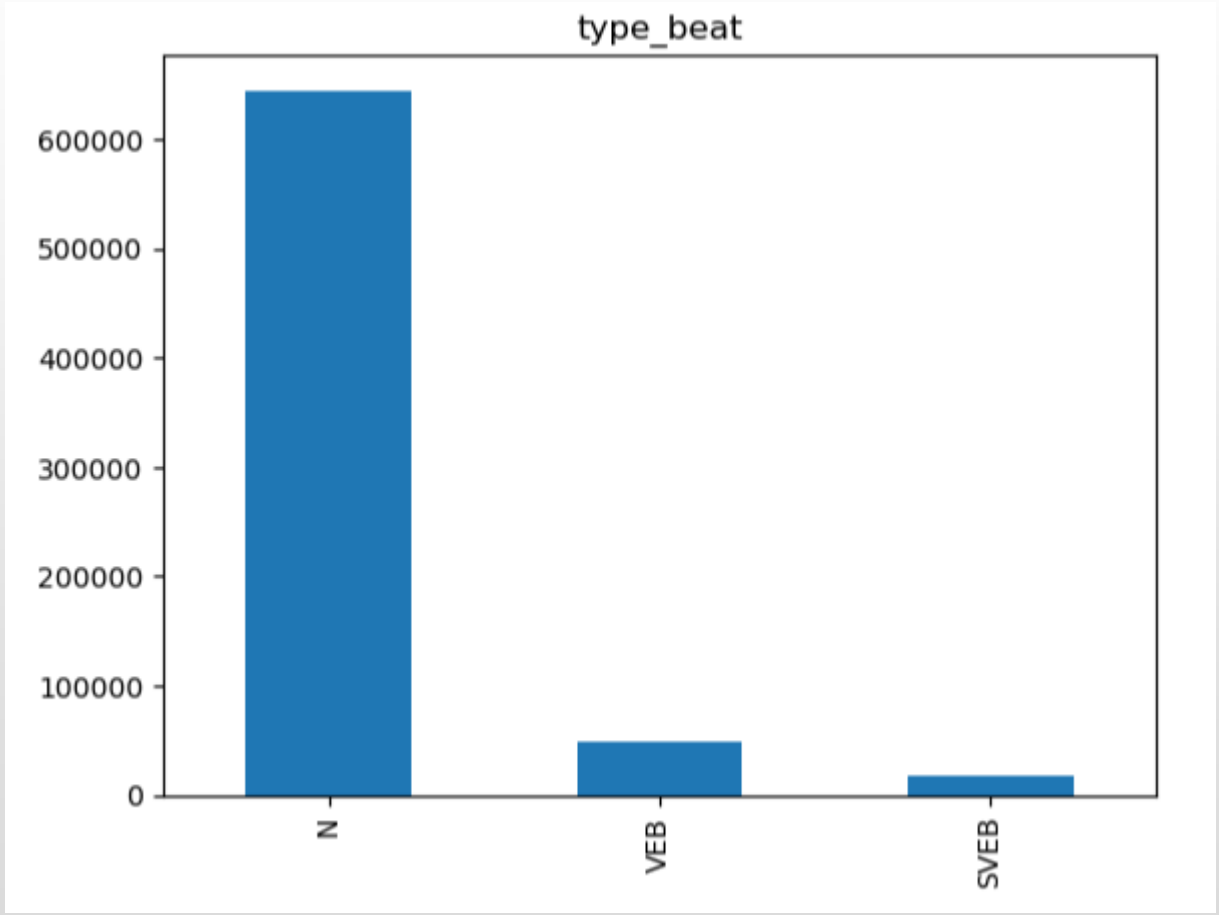
- Encoding ‘type’ feature

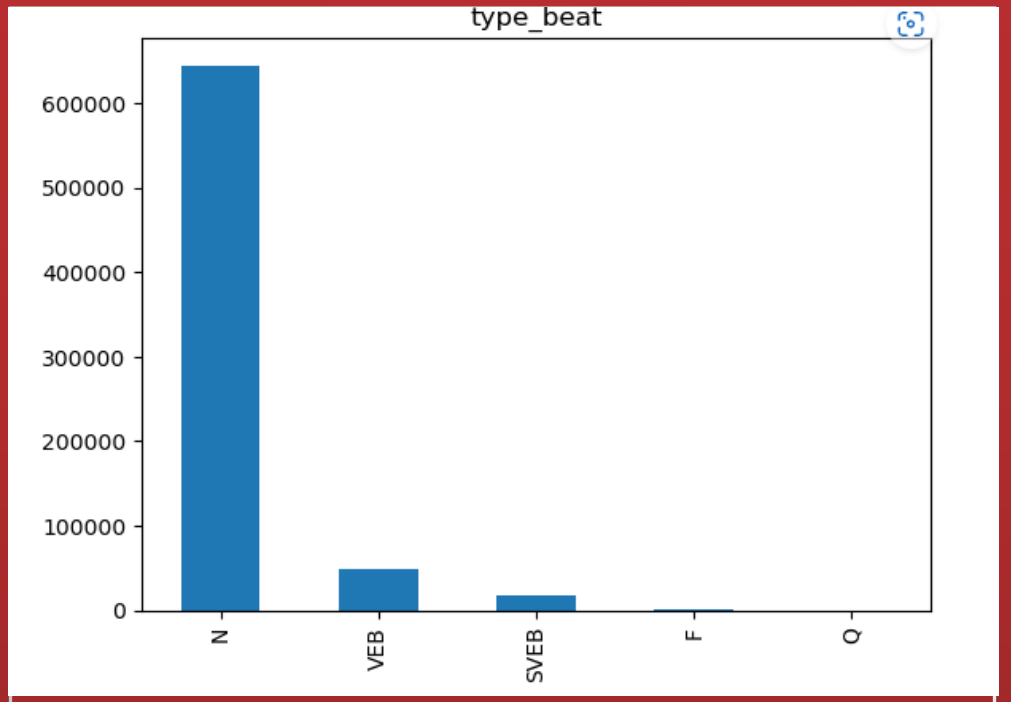
```
from sklearn.preprocessing import OrdinalEncoder,OneHotEncoder
enc=OrdinalEncoder()
X=df_cardio.drop('type',axis=1)
y=df_cardio['type']
encoded_class=enc.fit_transform(y.values.reshape(-1,1))

df_cardio['type']=encoded_class
df_cardio
```

Data Integration and Data Cleaning

- Fusion of the 4 datasets
- Encoding categorical features

```
df_cardio=df_cardio[df_cardio['type']!= 'Q']
df_cardio=df_cardio[df_cardio['type']!= 'F']
```





Imbalanced Class

An atrial premature beats usually means the presence of a Supraventricular anomalies, and a premature ventricular contraction beats usually means the presence of a Ventricular anomalies.

The inconsistency of the heart rates between ECG recordings would reduce the classification performance of the RR interval features.

Z-score normalization

$$v' = \frac{v - \mu_A}{\sigma_A}$$

RR interval features were normalized, using the mean and the standard deviation of the values of that specific features for a specific patient.

Data Normalization

- **Handling RR-interval features**

1	0_pre-RR	0.127404
17	1_pre-RR	0.124532
2	0_post-RR	0.046870
18	1_post-RR	0.045539

Normalization

	Feature	MI
1	0_pre-RR	0.267951
17	1_pre-RR	0.267835
18	1_post-RR	0.207661
2	0_post-RR	0.207591

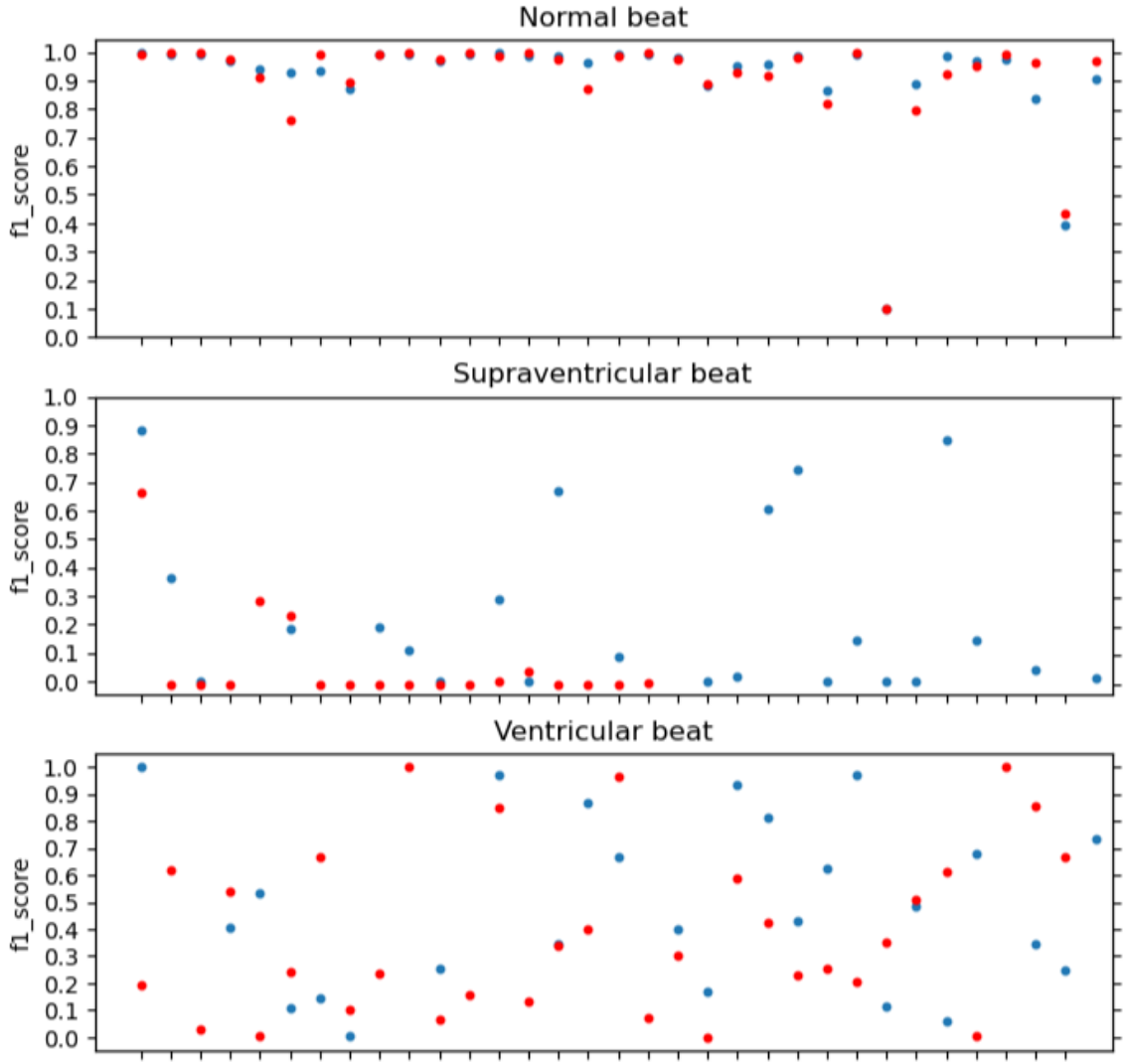
Classification on dataset MIT-BIHArrhythmiaDatabase.csv

K-Nearest-Neighbors classifier

pipe_normal
pipe_PCA

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 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

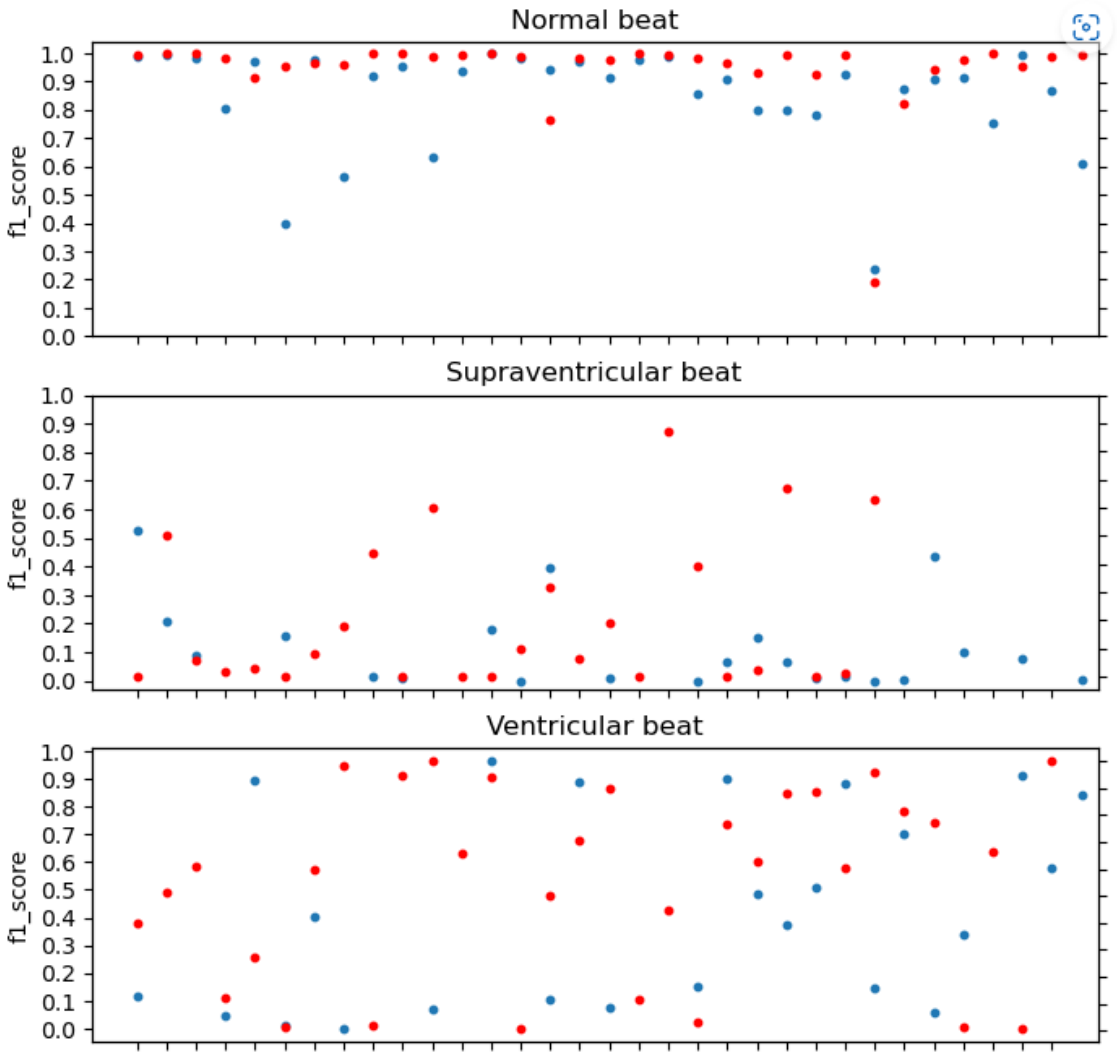


AdaBoost classifier

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

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

pipe_normal
pipe_PCA

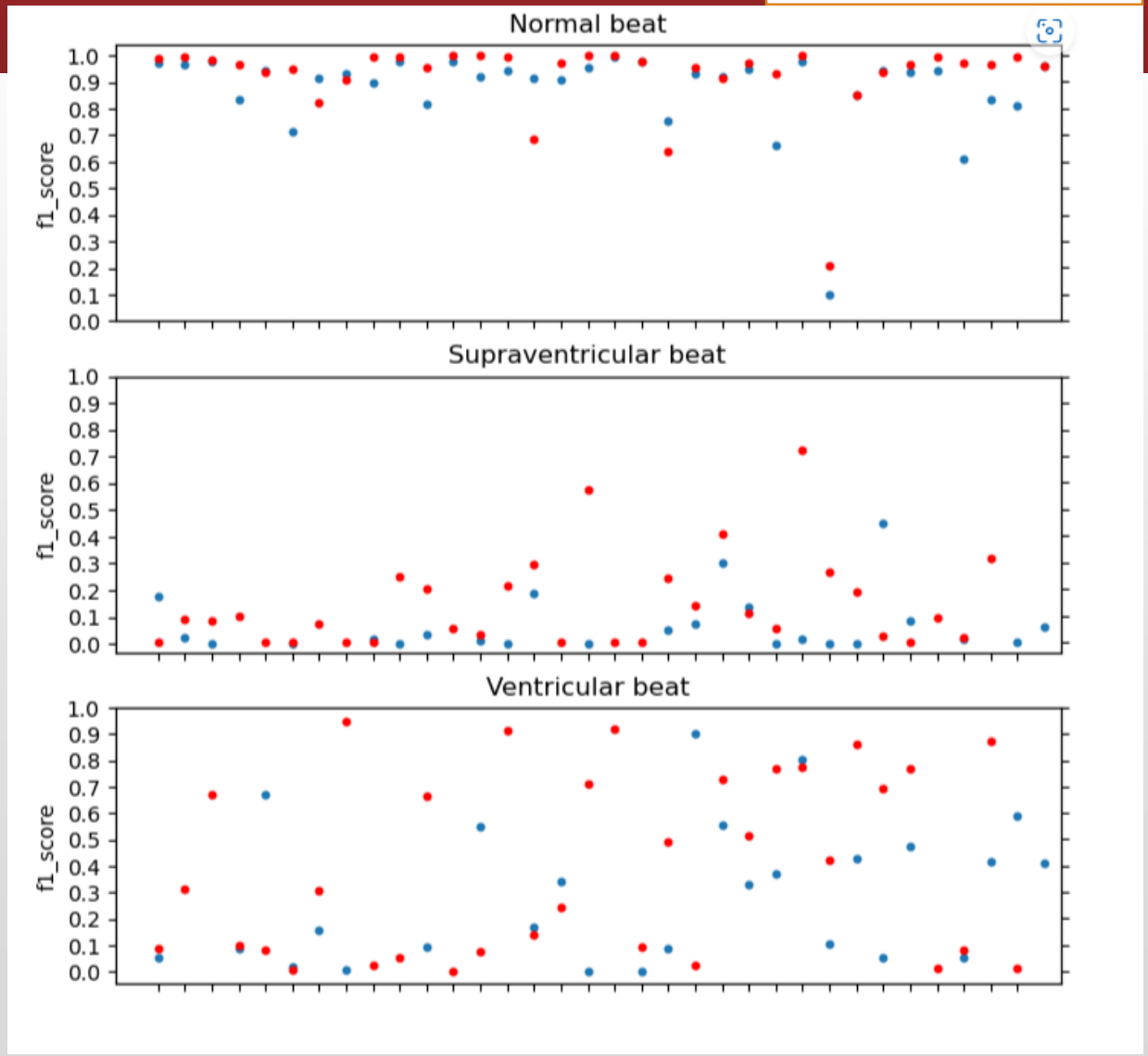


DecisionTree classifier

pipe_normal
pipe_PCA

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

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



RandomForest classifier

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

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

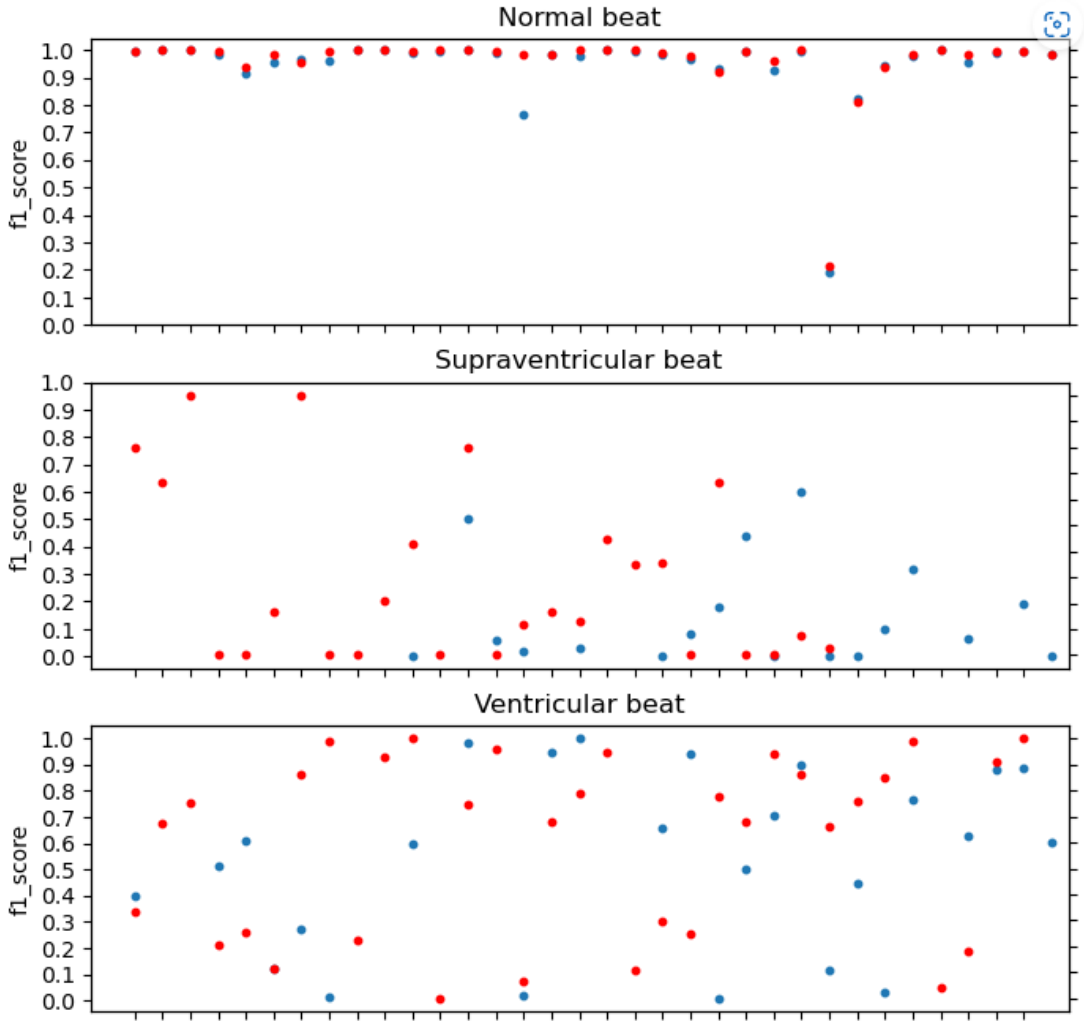
Compare classifiers

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

RandomForest
AdaBoost

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

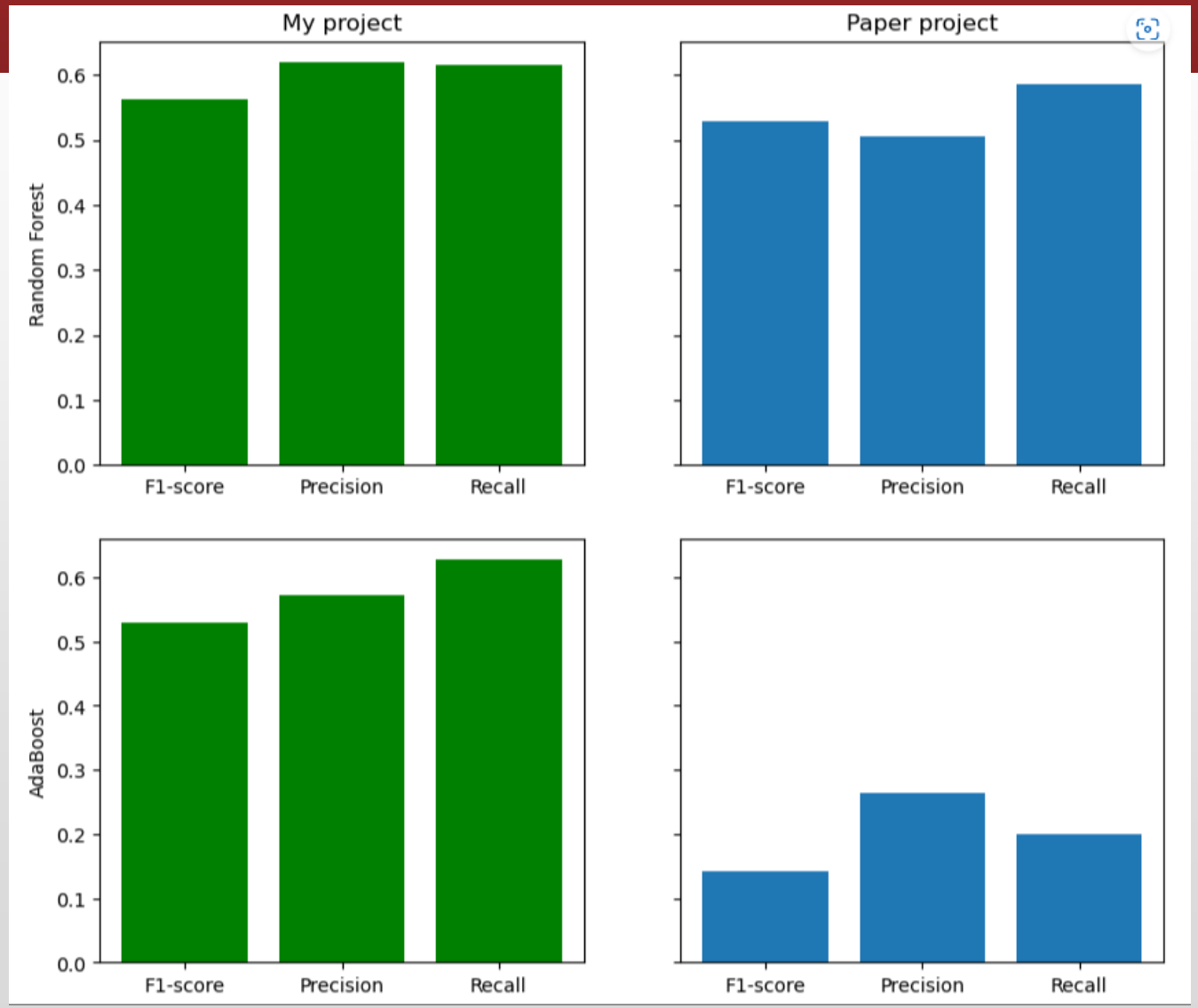
Classifier 1	Classifier 2	p-value
KNN	DecisionTree	0.2076
KNN	AdaBoost	0.04
KNN	RandomForest	0.007
AdaBoost	RandomForest	0,12
DecisionTree	AdaBoost	0.0005
DecisionTree	RandomForest	0,0000027



Compare results with paper

(Paper: DEVELOPMENT OF INTERPRETABLE MACHINE LEARNING MODELS TO DETECT ARRHYTHMIA BASED ON ECG DATA by Shourya Verma)

Project	Classifier	F1-score	Precision	Recall
HC	RandomForest	0.563	0.62	0.616
Paper		0.529	0.506	0.587
HC	AdaBoost	0.53	0.573	0.628
Paper		0.142	0.263	0.199



Comparison between different datasets

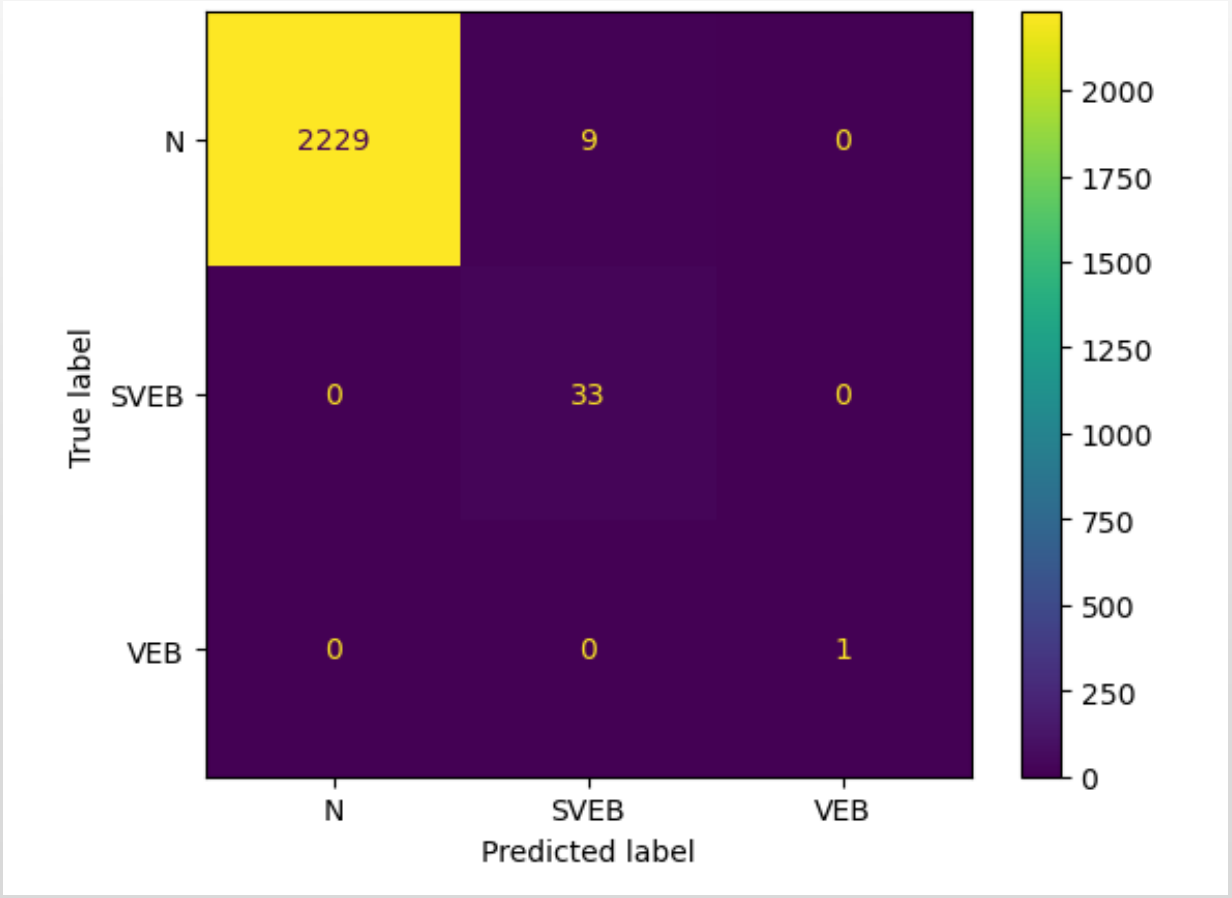
K-Nearest-Neighbors classifier

Dataset	F1-score (N)	F1-score (SVEB)	F1-score (VEB)	F1-score (macro avg)
MIT-BIHArrhythmia Database	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

Confusion matrix comparison

Patient 1 of MIT-BIHArrhythmiaDatabase

KNN on MIT-BIHArrhythmiaDatabase



KNN on merged dataset