**Experiments with Increasing DataBases**

**Baseline Model:**

Our Baseline Model is a Random Forest Classifier. We perform 10 fold Cross-validation based on the Competition scoring in order to tune the hyperparameters of our model. The hyperparameters that are being tuned are:

|  |  |  |
| --- | --- | --- |
| Parameter | Range | Step |
| Number of trees in the forest | 50-200 | 40 |
| Number of features at every split | ‘auto’, ‘sqrt’ |  |
| Maximum depth of a tree | 10-70 | 15 |
| Minimum Sample Split | 2-15 | 4 |
| Minimum Sample Leaf | 1-5 | 5 |
| Bootstrap | ‘True’, ‘False’ |  |
| Weighted | ‘Balanced’, ‘Balanced\_subsample’ |  |

**DataBases:**

What I am going to do is the following:

I will use a fixed test Set: I will use the split we devised. I will not use the feature selection model since the features selected are not being yet selected on the whole database. The number of features extracted are therefore 1360. Furthermore, in the training examples, I will discard examples not succeeding in the preprocessing validation check (from **Ribeiro Paper**).

I will use an increasing database while always maintaining the same Test Set. I will iterate over the different DataBases in the Training split, in the following order:

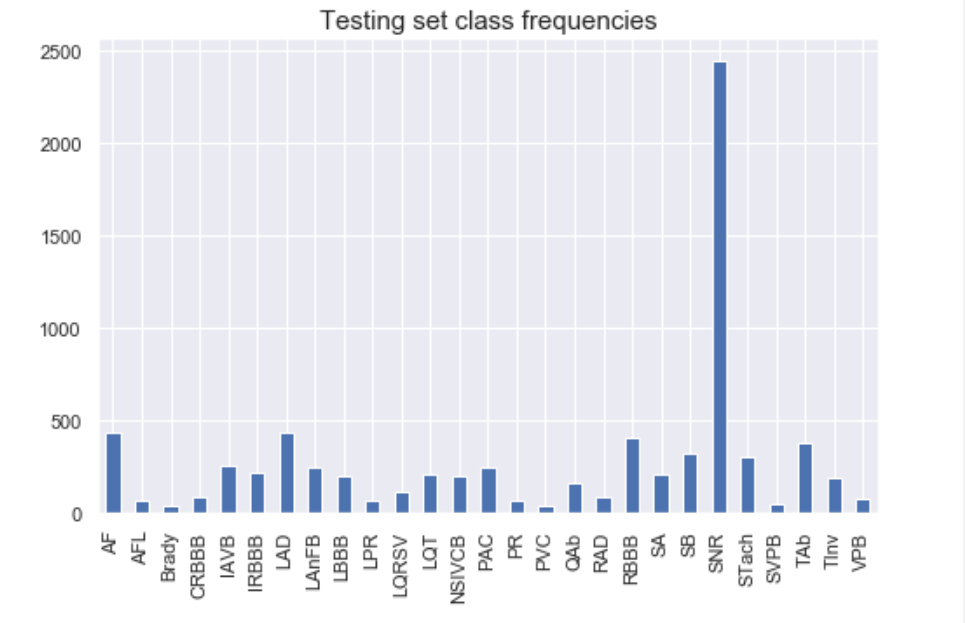
* Georgia DataBase, adding the following examples: **PR, LPR, SVPB** from PTB\_XL and **PVC** and **Brady** from CPSC extra while being careful about the fact that I add the examples that are not already in the Test Set.
* I will then add the training examples not already used from the DataBases:
  + PTBXL
  + PTB (**Note: pay attention to the changing sampling frequency when extracting features**)
  + CPSC
  + CPSC-Extra

The training of my model will be done via 10 fold - Grid Search Cross Validation. I will use single label classification and assign multi labels based on a threshold on probabilities (for now set to 0.3).

**Last,** in the training examples, I will discard examples not succeeding in the preprocessing validation check (from **Ribeiro Paper**) and compare the performances I get when including or not these examples.

**TestSet:**

Composition of the Test Set.



There are no multi-label cases in the Test Set.

**Training DataBases:**

We are going to train our model on DataBases in a progressive order. The order in which I am going to add the databases is:

* Georgia DataBase with missing pathologies from PTB\_XL and CPSCB
* PTB\_XL
* CPSCA
* CPSCB + PTB

**Results on Test Set:**

**Georgia DataBase**:

* Fbeta score:

{'AF': 0.6855404773046326, 'AFL': 0.3391959798994975, 'Brady': 0.38205980066445183, 'IAVB': 0.39159109645507006, 'IRBBB': 0.039414414414414414, 'LAnFB': 0.16422435573521982, 'LAD': 0.07835455435847209, 'LBBB': 0.8126195028680688, 'LPR': 0.278372591006424, 'LQRSV': 0.24208566108007448, 'LQT': 0.39404553415061294, 'NSIVCB': 0.012376237623762377, 'PR': 0.37722419928825623, 'PAC': 0.7512953367875648, 'PVC': 0.3282275711159737, 'QAb': 0.03765060240963856, 'RAD': 0.0847457627118644, 'RBBB': 0.5510116229014206, 'SA': 0.48633440514469456, 'SB': 0.6157112526539278, 'SNR': 0.640810152113623, 'STach': 0.7962740384615384, 'TAb': 0.1868377021751255, 'TInv': 0.032981530343007916}

* Competition metrics: 0.30

**Georgia + PTB\_XL:**

* Fbeta score:

{'AF': 0.8142006802721088, 'AFL': 0.30405405405405406, 'Brady': 0.32653061224489793, 'IAVB': 0.37916666666666665, 'IRBBB': 0.1076923076923077, 'LAnFB': 0.44131028207461326, 'LAD': 0.53515625, 'LBBB': 0.8349328214971209, 'LPR': 0.1746031746031746, 'LQRSV': 0.18962075848303392, 'LQT': 0.3477822580645161, 'NSIVCB': 0.09003601440576231, 'PR': 0.3325942350332594, 'PAC': 0.7917888563049853, 'PVC': 0.28938906752411575, 'QAb': 0.0, 'RAD': 0.3838582677165354, 'RBBB': 0.3729281767955801, 'SA': 0.5096660808435852, 'SB': 0.6424433912585571, 'SNR': 0.8002177463255308, 'STach': 0.8158368326334733, 'TAb': 0.14796547472256474, 'TInv': 0.03942181340341656}

* Competition metrics: 0.41

**Georgia + PTB\_XL + CPSCA:**

* Fbeta score:

{'AF': 0.8478802992518704, 'AFL': 0.26223776223776224, 'Brady': 0.2631578947368421, 'IAVB': 0.6198960653303638, 'IRBBB': 0.03954802259887006, 'LAnFB': 0.4360867558837102, 'LAD': 0.5638801261829653, 'LBBB': 0.8570075757575758, 'LPR': 0.132013201320132, 'LQRSV': 0.1934826883910387, 'LQT': 0.32388663967611336, 'NSIVCB': 0.06105006105006105, 'PR': 0.536309127248501, 'PAC': 0.8235294117647058, 'PVC': 0.16917293233082706, 'QAb': 0.007680491551459293, 'RAD': 0.25065963060686014, 'RBBB': 0.7069881487535759, 'SA': 0.511049723756906, 'SB': 0.6336528221512248, 'SNR': 0.8004385964912281, 'STach': 0.8166363084395871, 'TAb': 0.16257668711656442, 'TInv': 0.0461133069828722}

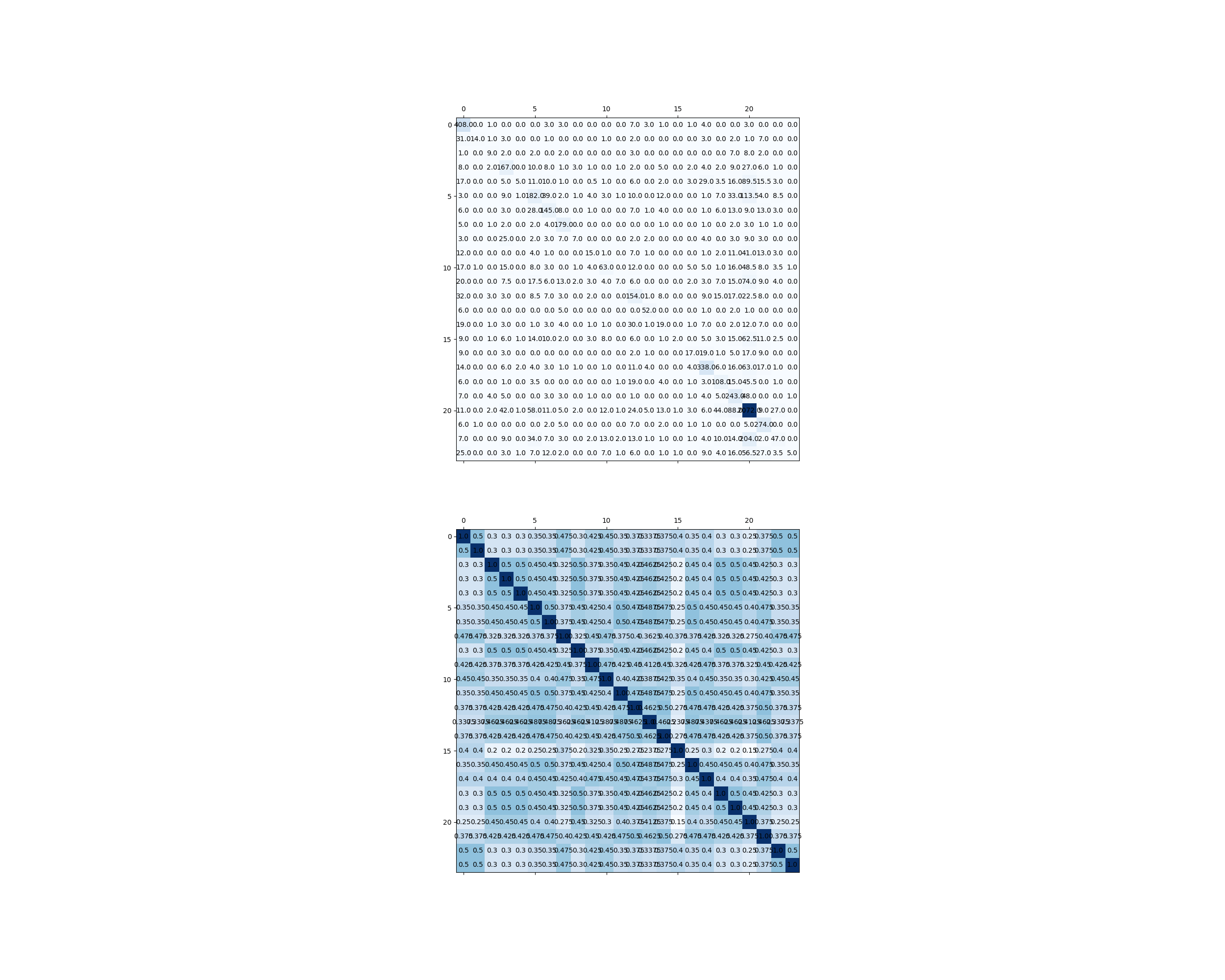
* Competition score: 0.45

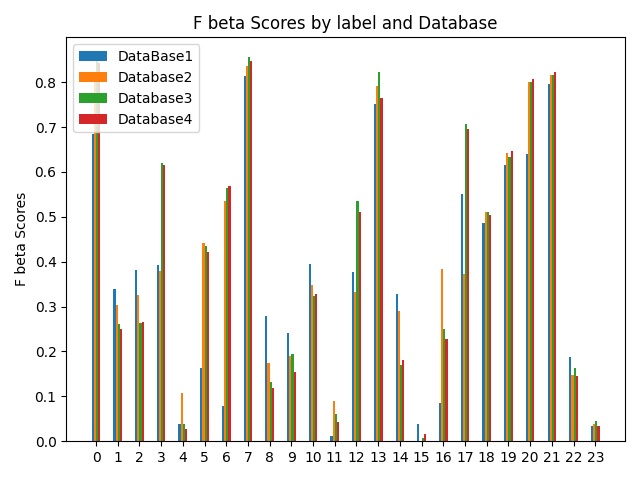
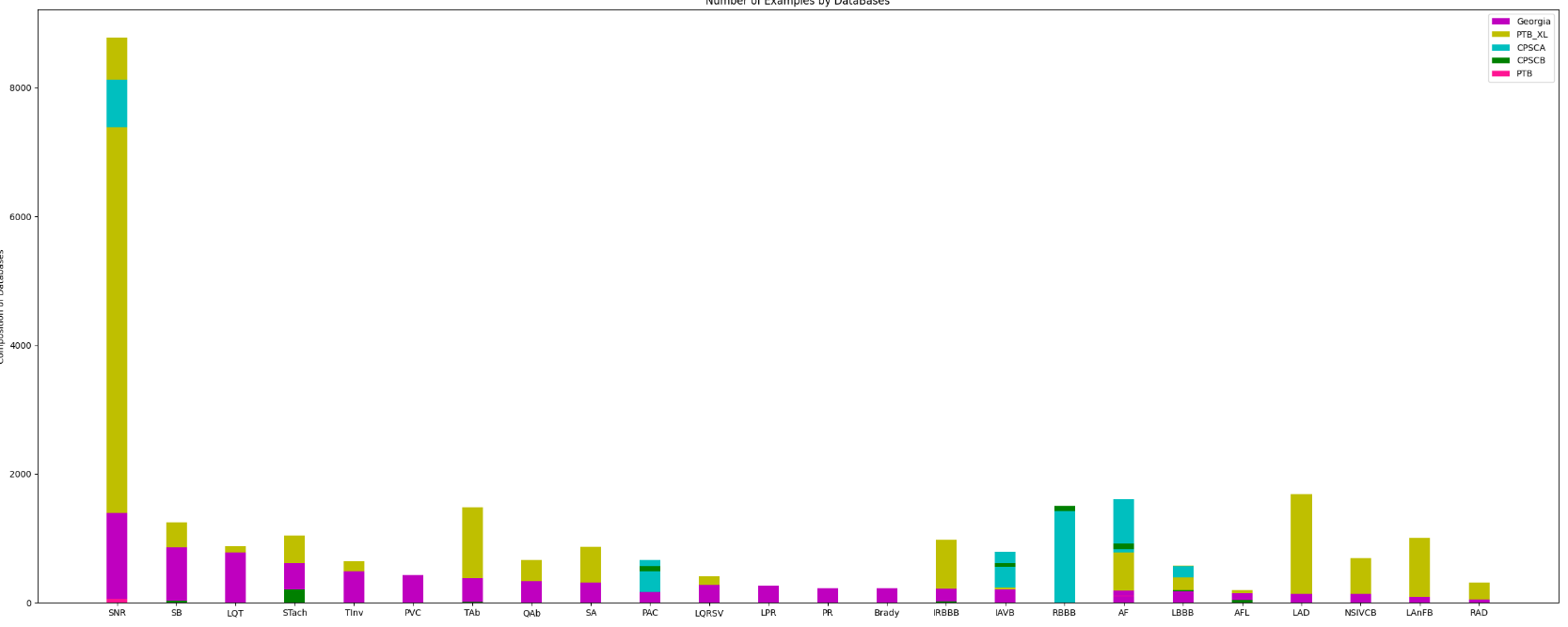
**All DataBases:**

* Fbeta score:

{'AF': 0.8436724565756824, 'AFL': 0.25, 'Brady': 0.26627218934911245, 'IAVB': 0.6153279292557111, 'IRBBB': 0.028312570781426953, 'LAnFB': 0.42266604737575475, 'LAD': 0.5695208169677927, 'LBBB': 0.8467360454115421, 'LPR': 0.11784511784511785, 'LQRSV': 0.1540041067761807, 'LQT': 0.32710280373831774, 'NSIVCB': 0.042997542997543, 'PR': 0.5102717031146454, 'PAC': 0.7647058823529411, 'PVC': 0.18199233716475097, 'QAb': 0.015337423312883436, 'RAD': 0.22727272727272727, 'RBBB': 0.6949013157894737, 'SA': 0.5046728971962616, 'SB': 0.6473095364944059, 'SNR': 0.8074195308237861, 'STach': 0.8233173076923077, 'TAb': 0.1447935921133703, 'TInv': 0.033112582781456956}

* Competition score: 0.48

**Final Confusion Matrix**

**Summary Charts**