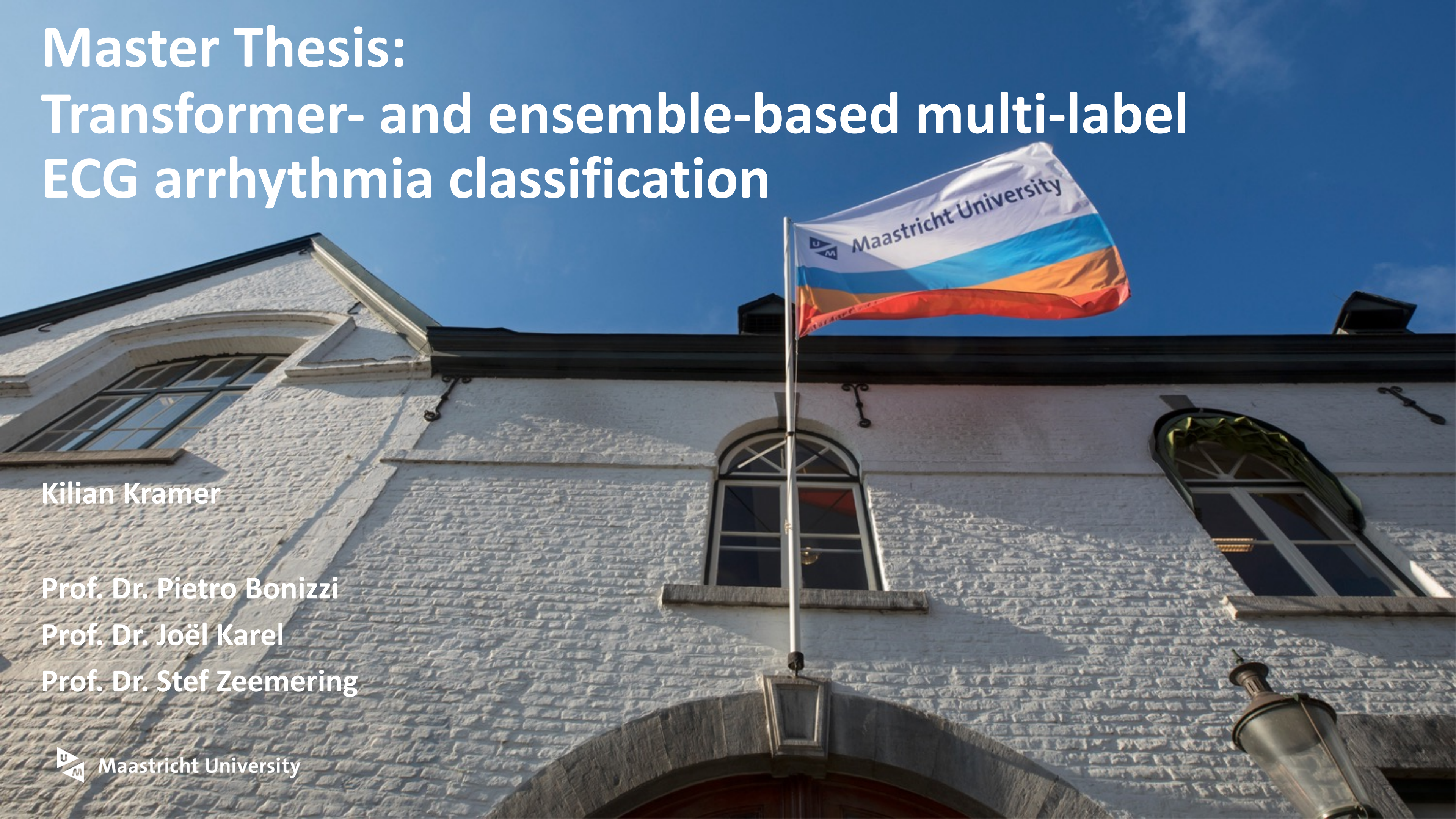


Master Thesis: Transformer- and ensemble-based multi-label ECG arrhythmia classification

Kilian Kramer

Prof. Dr. Pietro Bonizzi
Prof. Dr. Joël Karel
Prof. Dr. Stef Zeemering

Maastricht University

Index

- Problem Introduction
- Related Work
- Research Questions
- Methodology
- Experiments and Evaluation
- Conclusion

Problem Introduction

- Cardiovascular diseases, such as Atrial Fibrillation, are among the leading causes of death in the population, resulting in an increased demand for cardiac assessment

- Deep learning can be used to develop multi-classification models for arrhythmia detection and improve medical monitoring

- Today, much research and accurate models are available for simpler arrhythmia classification tasks, e.g. AFIB classification (binary) or grouped common arrhythmia types (i.e. AAMI standard)

Problem Introduction

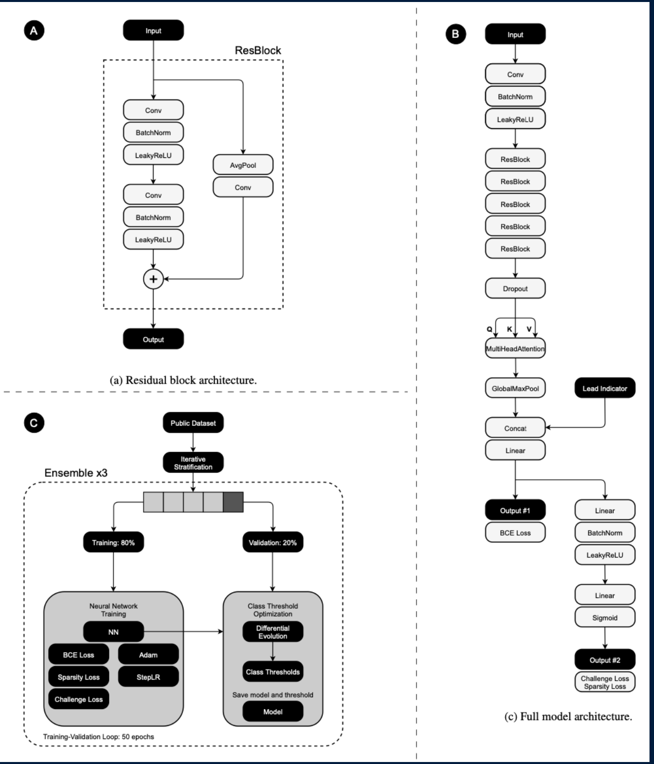
- Problem: Professional treatment requires individual and detailed ECG assessment

- In recent years, Transformer models have gained considerable popularity due to the self-attention mechanism and research papers that apply Transformer models on less comprehensive ECG arrhythmia classification tasks show good results

- However, research and accurate models are limited on comprehensive arrhythmia detection tasks including Transformer models and rare subdiseases, such as Atrial Flutter, Premature Ventricular Contractions, Prolonged QT interval etc.

Related work: Classification of ECG Using Ensemble of Residual CNNs with Attention Mechanism

- Prop. by Nejedly et. al
- Winning paper of the Physionet 2021 challenge
- Achieves 58% accuracy on challenge metrics on all subtasks (2-, 3-, 4-, 6- and 12-leads) for multi-label classification of 26 classes
- Uses several residual CNN blocks in combination with attention (a single encoder-block)
- Specific designed loss function that incorparates class weights of Challenge evaluation metrics
- Authors show in a follow up study “Classification of ECG using ensemble of residual CNNs with or without attention mechanism“ that the Multi-Head attention block does not improve model performance (model achieves 59% challenge score without)



Related work: ECGBERT: Understanding Hidden Language of ECGs with Self-Supervised Representation Learning

- Prop. by Choi et. al
- Based on BERT methodology: authors create own wave segment vocabular to tokenize ECGs, apply similar training approach (MLM) on model, which can be fine-tuned on multiple downstream tasks
- First, ECG signals are preprocessed (filtering etc.) and splitted by fiducial points
- Segment vocabular is previously obtained from clustering ECG segments into 70 distinct clusters using K-mean and Dynamic Time Warping to train four classifiers for P, QRS, T and background wave clusters -> classifiers are trained based on extracted fiducial points from ECG segments resulting in 12 P, 19 QRS, 14 T and 25 background (e.g. PR or ST intervals) wave clusters
- Each segment is classified by the corresponding classifier and assigned to a specific wave cluster
- Classified wave segments are then mapped to tokens; authors do not describe how encoding of tokens look like (I assume these are random initialized embeddings, whether these are on top trainable parameters in the model is also not described in their paper)
- In addition, positional information (temporal) and CNN feature embeddings are added to tokens
- Authors reason that tokens can only provide general ECG context information, but CNN features provide refined pattern information; CNN features are extracted from raw ECG using an U-Net
- Based on this pipeline, the autors create training samples that form ECG sentences of wave segments, where each sentence contain 1-8 consecutive heartbeats (a heartbeat is composed of several wave segments)
- The sentence are then inputted to the Transformer-encoder module, where several sequences can be inputted which are splitted by a seperation “[SEP]“ token
- Training data from Physionet.org: (AFIB/arrhythmia) MIT-BIH database and Apnea-ECG database
- Model is pre-trained using 15% masking (similar to Masked Language Modeling in LLMs)
- Several models are fine-tuned (at low cost using pre-trained model) on different downstream tasks by adapting/fine-tuning the output layers to: AFIB classification (binary), heartbeat classification (4 classes, AAMI standard), ECG patient verification and sleep apnea detection

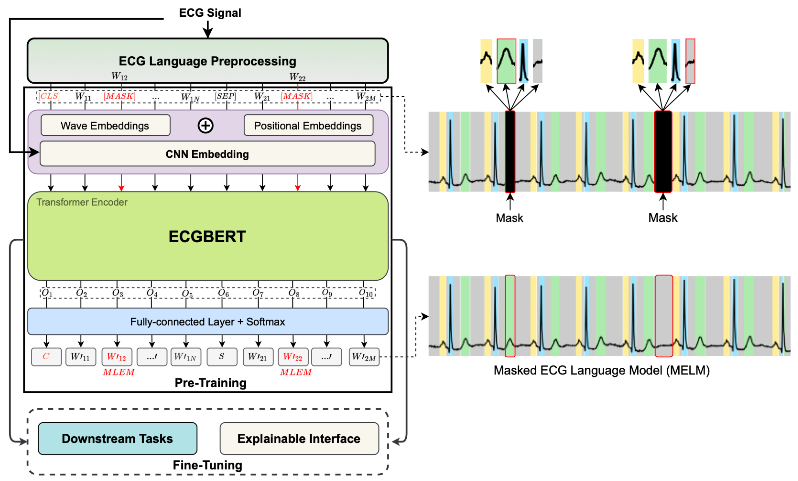


Figure 2: Illustration of the proposed ECGBERT architecture

AFIB

Model	Paradigm	Signal Length	Performance: Accuracy	Specificity	Sensitivity	PPV
Tuboly et al. [2021]	Intra-patient	60s	0.980	0.987	0.974	0.988
ResNet	Inter-patient	10s	0.884	0.951	0.846	0.969
Andersen et al. [2019]	Inter-patient	30 RRIs	**0.978**	**0.989**	0.969	0.957
Pereira and Andreão [2022]	Inter-patient	10	0.908	0.910	0.915	-
ECGBERT	Inter-patient	10s	0.973	0.976	**0.970**	**0.981**

Heartbeat classification

		Predicted: N	S	V	Q	Per-class Performance: Accuracy	Specificity	Sensitivity	PPV
True	**N**	38538	1483	1941	1119	0.86	0.45	0.89	0.94
	S	187	26	39	7	0.95	0.99	0.10	0.01
	V	1778	201	1280	277	0.91	0.94	0.36	0.38
	Q	451	77	25	2445	0.96	0.99	0.82	0.64

Research Questions

1. How well does a Transformer-based model perform on the Physionet 2021 challenge data compared to a feature-based model or a Convolutional Network?

2. Can an ensemble of Transformer model and Convolutional Network effectively capture spatio-temporal information and improve accuracy?

3. Which model performs best at discriminating SR, AF, AFL, PAC and PVC on both datasets?

4. What are the challenges in transferring the pre-trained models from the Physionet 2021 challenge data to the MyDiagnostick database? Do the models generalise well, even though different ECG devices were used?

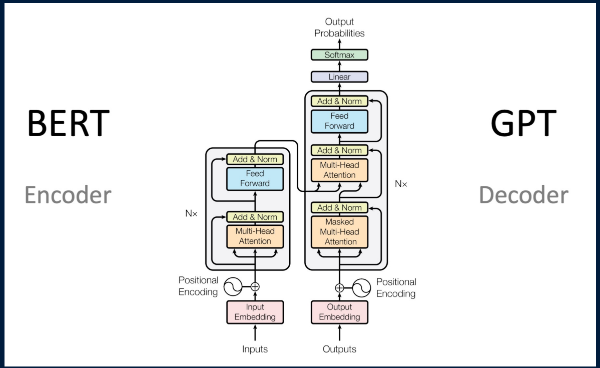
Methodology

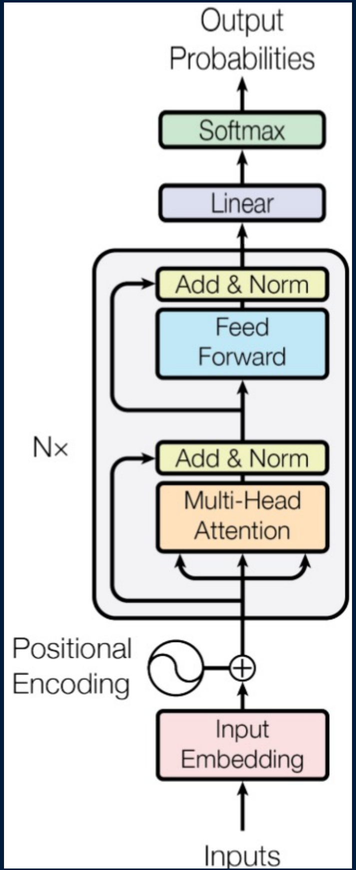
Several models are compared:

1. Feature-based classifier (Biobss features)
2. Residual Convolutional Network
3. Standard Transformer Encoder
4. Convolutional Network + Attention
5. Ensemble of various models (use multiple models or AdaBoost)

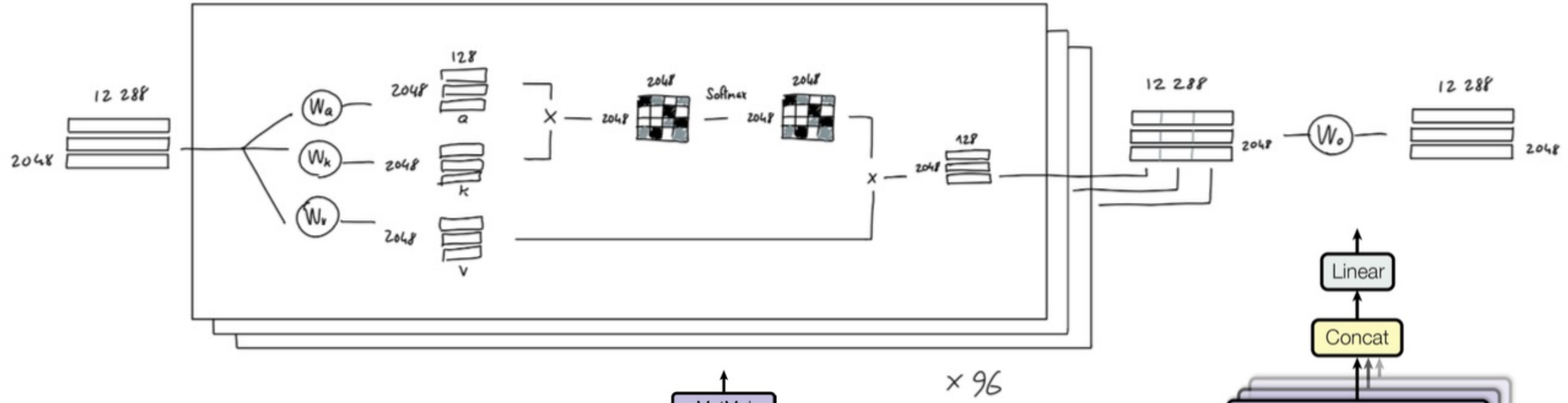
Methodology: Transformer

- Developed model uses Multi-head attention block from Transformer encoder, since it is appropiate for classification tasks
- Decoder objective is generation, which uses cross-attention and adds masking to map input sequence to output sequence
- Multi-head attention block outputs same number of inputted embeddings as outputs, but enrichted with attention / contextualized
- Outputted embeddings can then be feeded to a dense layer for classification tasks





Mathematical background: Transformer



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

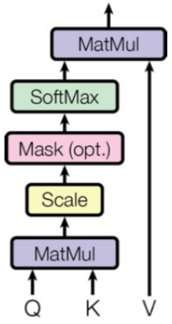


Figure 3.2: Scaled Dot-Product Attention

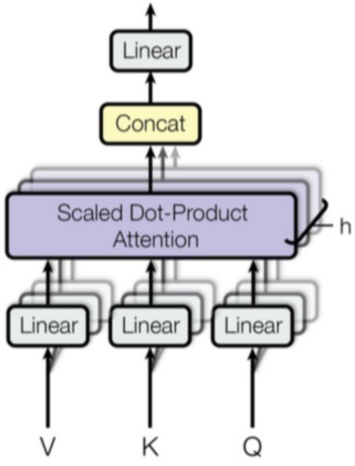
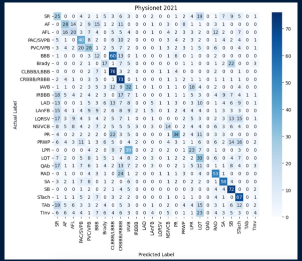


Figure 3.3: Multi-head attention

Methodology: Ensemble of models

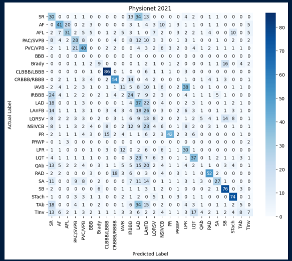
Improved classification performance

Model 1



+

Model 2



+ …

- Can be different models
- Can be same model using different validation splits(+ AdaBoost?)

Experiments and Evaluation

- For all model training the Physionet 2021 challenge database will be used: https://physionet.org/content/challenge-2021/1.0.3/

- Experiments & evaluation will focus on two tasks:

  1. Physionet 2021 challenge database for training & evaluation (contains 26 classes annotations); models will be compared with other pariticipant models (https://moody-challenge.physionet.org/2021/results/) using the evaluation metrics (https://github.com/physionetchallenges/evaluation-2021)

  2. MyDiagnostick database (annotations contain 5 classes: SR, AF, AFL, PAC, PVC), is used for evaluation (not training) to evaluate the generalization of the transfered models from pre-training on the Physionet data

  – At the moment only single-lead (I) ECGs are used for training and evaluation, because the MyDiagnostick data only provides single-lead ECGs (I am not sure if I will be able to focus on multiple leads experiments in time)

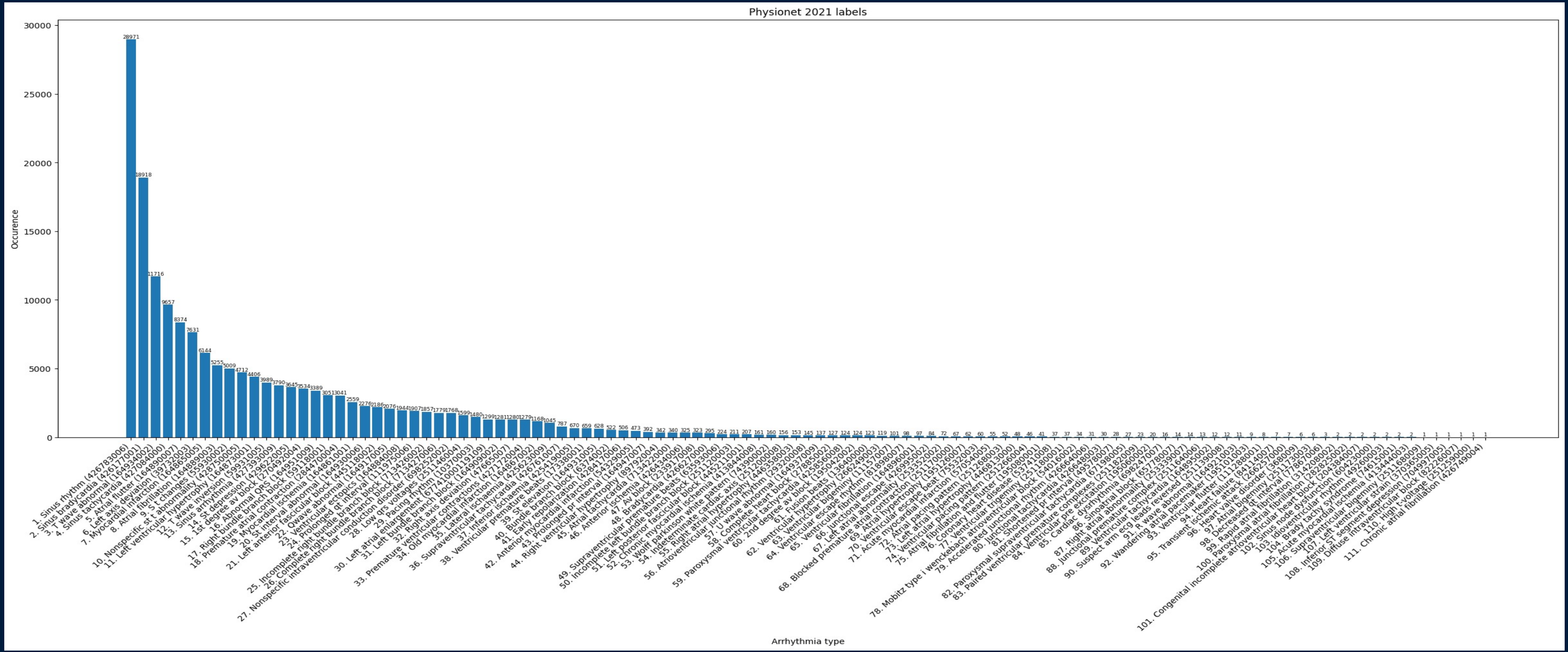
Data: Physionet 2021 Challenge database

Dataset source	Average ECG length (seconds)	Data samples
Ningbo database	10s	34,905
PTB-XL database	10s	21,837
Chapman-Shaoxing database	10s	10,247
Georgia 12-lead challenge data	9s	10,344
CPSC database	15s	6. 877
CPSC-extra database	15s	3,453
PTB database	110s	516
INCART database	1800s	74

- database is composed of several datasets

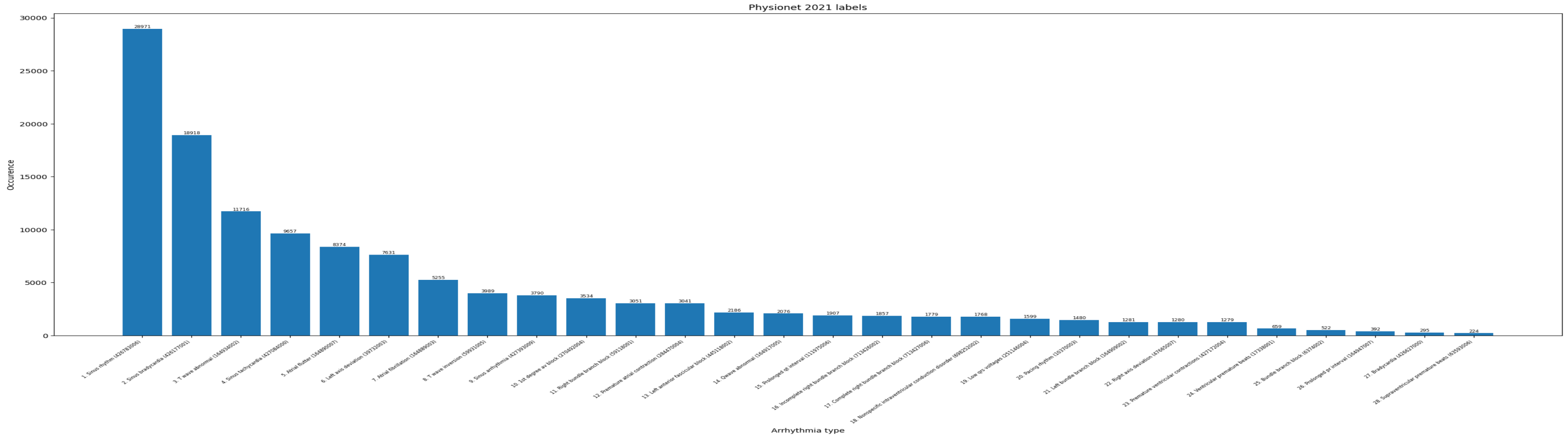
- contains about 89.000 12-lead ECGS with more than 100 arrhythmia type annotations

Physionet 2021 data distribution



Maastricht University

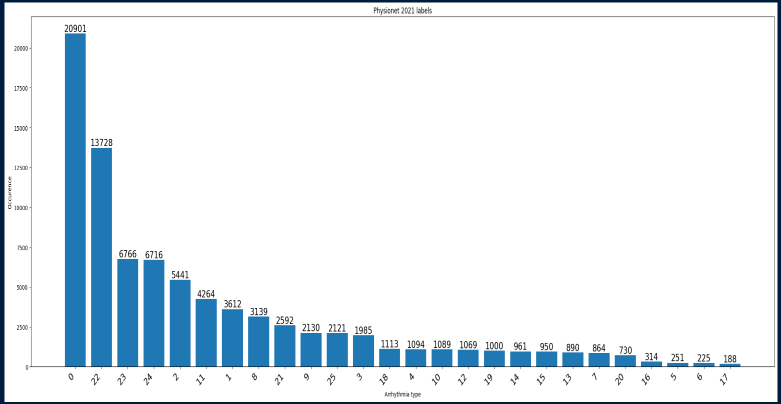
Physionet 2021 challenge data distribution (subset)

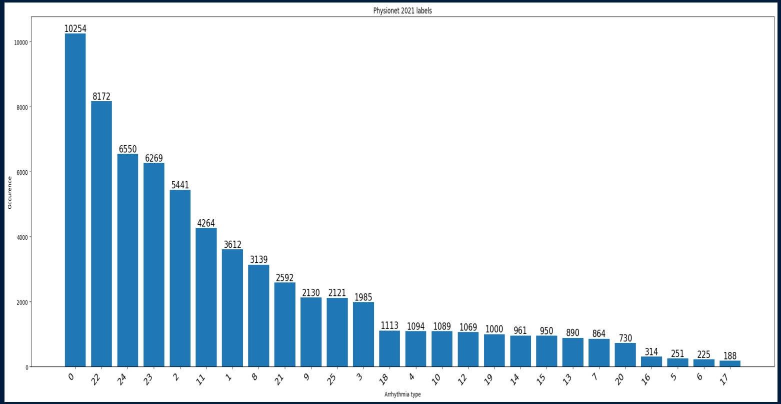


- Official Physionet 2021 challenge uses a subset of 26 arrhythmia classes (4 classes are grouped, e.g. premature ventricular contractions PVC & ventricular premature beats VPB are treated as same class, PAC & SVPB as same, …)

Evaluation: Training data

- Class balance not entirely possible, because samples are multi-label annotated
- Many Sinus Rhythm cases present, which are classified as both Sinus Rhythm and any minor class
- However, class balancing to some extend possible
- Unbalanced (image up), balanced (image down)



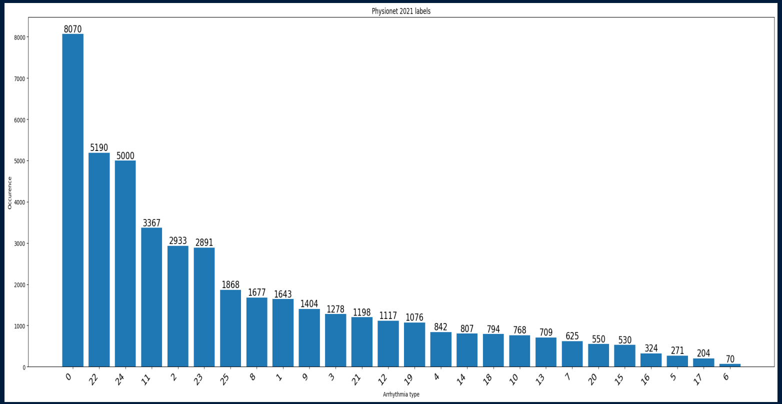


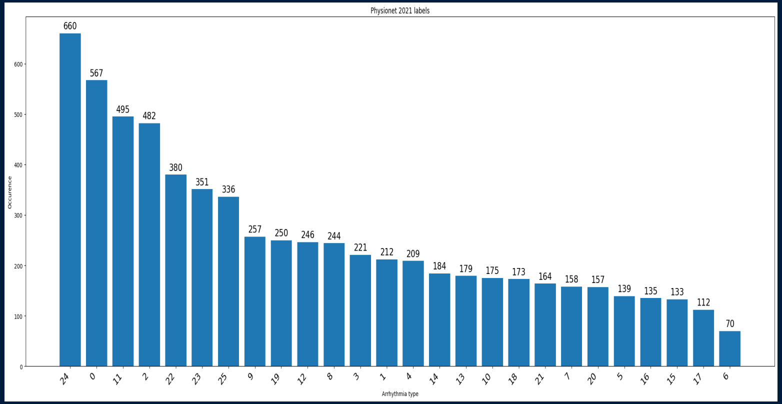
Evaluation: Test data

- I noticed that the challenge evaluation scores are highly depend on test data distribution, e.g. if Sinus Rhythm is downsampled in testset it highly affects the calculated challenge score

- Example: Above testset yielded for the same model a challenge score of 0.30, while below for same model (with less SR) only 0.05

- Reason: Challenge scores affected by metric weights and proportion of Sinus Rhythm samples





Evaluation: Test data

- Problem: Official Physionet 2021 challenge test dataset is private/withhold

- I tried to reach out to the Physionet 2021 challenge organisers to be able to access/evaluate the models on the official Physionet data

Master's Thesis evaluation: Application of Transformer models on the Physionet 2021 challenge database

KK Kramer, Kilian (Stud. DACS / Alumni DACS)
An: contact@physionet.org
Mo, 24.06.2024 11:42

Diese Nachricht befindet sich in Englisch | Übersetzen nach Deutsch | Nie aus Englisch übersetzen

Dear organisers of the CinC Physionet 2021 Challenge,

first of all, a big thank you for the provided Physionet 2021 Challenge database, which is an important part of my Master's thesis. Regarding my current Master's thesis on the application of Transformer and other models (i.e. ensemble of CNN, Transformer and feature-based model) to multi-label arrhythmia classification, I need your help to evaluate my models. I have used the Physionet 2021 Challenge database to train my models and would like to compare my results with the official Challenge metrics and participant team scores (https://moody-challenge.physionet.org/2021/results/), but the official test data is private. Could you please run my best validated model on the test data so that I can successfully draw a proven conclusion from my evaluation?

Yours sincerely,

Kilian Kramer from Maastricht University, Netherlands

- At the moment I created a testset from the training data

Evaluation: Test data

- I do not expect to get an answer from the Physionet organisers
- I expect a similar amount of Sinus Rhythm cases to be in the test set: 89k are public, (29k from 89k are Sinus Rhythm) - and 36k (amount of Sinus Rhythm?) are privat samples
- TODO: Implement cross-validation to make challenge metrics to some extend comparable

Experimental results: Transformer

Input shape	Positional encoding	Encoder blocks	Heads	qkv dim	ff dim	Dropout	Trainable param.	Accuracy	Precision	Recall	F1
(40, 50)	True	1	1	25	24	0.1	155.375	0.096	0.514	0.120	0.194
(40, 50)	True	1	1	25	24	0.4	155.375	0.065	0.418	0.083	0.139
(40, 50)	True	8	1	25	24	0.1	878.818	0.061	0.579	0.079	0.139
(40, 50)	True	8	1	25	24	0.4	878.818	0.098	0.592	0.122	0.203
(40, 50)	False	1	1	25	24	0.1	155.375	0.227	0.747	0.25	0.374
(40, 50)	False	1	1	25	24	0.4	155.375	0.224	0.742	0.253	0.378
(40, 50)	False	8	1	25	24	0.1	878.818	0.228	**0.765**	0.261	0.389
(40, 50)	False	8	1	25	24	0.4	878.818	0.226	0.737	0.255	0.379
(10, 200)	False	8	1	25	24	0.1	1.004.818	0.160	0.744	0.174	0.283
(10, 200)	False	8	8	25	24	0.1	2.129.018	0.177	0.709	0.197	0.308
(40, 50)	False	8	8	400	24	0.1	6.035.018	0.223	0.762	0.247	0.374
(40, 50)	False	8	8	25	2048	0.1	65.947.210	0.219	0.740	0.257	0.382
(40, 50)	False	8	8	400	2048	0.1	70.819.210	0.226	0.751	0.257	0.383
(40, 50)	False	8	8	400	2048	0.4	70.819.210	**0.245**	0.739	**0.286**	**0.413**

Table 4.2: Physionet 2021 train/test split model comparison

- Testing standard Transformer encoder model on the Physionet data (gridsearch)

Experimental results: Physionet 2021 metrics

Model	Accuracy	Precision	Recall	F-measure
Random Forest (Biobss features)	0.272	0.800	0.291	0.427
Residual CNN	**0.392**	**0.839**	**0.505**	**0.63**
Standard Transformer Encoder*	0.238	0.733	0.280	0.406
CNN + Attention	0.272	0.768	0.330	0.462
Wavelet + CNN + Attention	0	0	0	0
Spectogram + CNN + Attention	0	0	0	0
Ensemble of Transformer (AdaBoost)	0	0	0	0
Ensemble of various Models	0	0	0	0

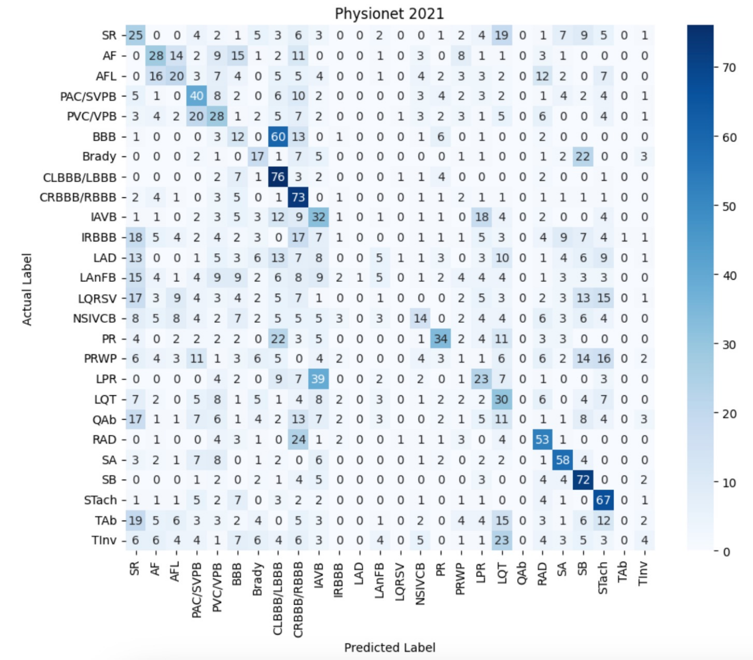
Table 4.3: Physionet 2021 train/test split model comparison

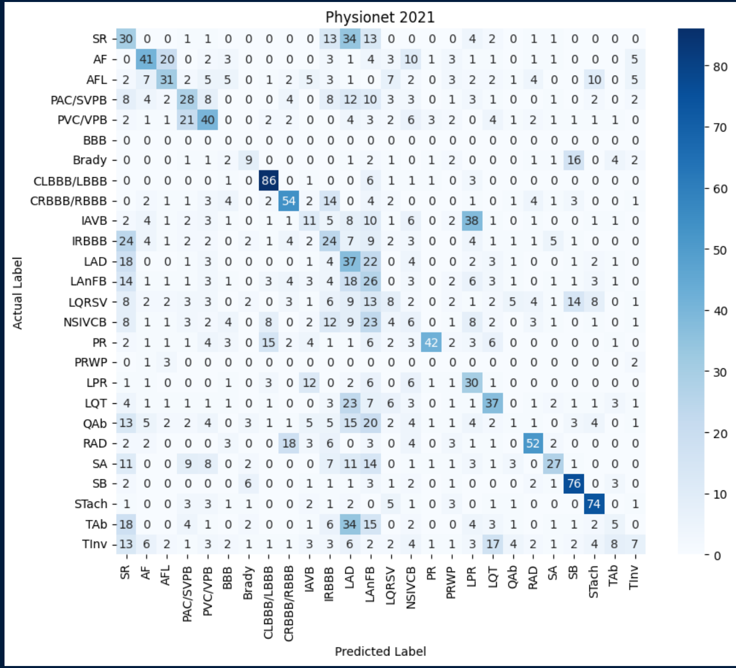
Model	AUROC	AUPRC	Accuracy	F-measure	Challenge metric
Random Forest (Biobss features)	0.554	0.146	0.272	0.137	0.073
Residual CNN	**0.895**	**0.477**	**0.392**	**0.359**	**0.376**
Standard Transformer Encoder*	0.740	0.246	0.238	0.163	0.055
CNN + Attention	0.797	0.278	0.272	0.162	0.140
Wavelet + CNN + Attention	0	0	0	0	0
Spectogram + CNN + Attention	0	0	0	0	0
Ensemble of Transformer (AdaBoost)	0	0	0	0	0
Ensemble of various Models	0	0	0	0	0

Table 4.4: Physionet 2021 challenge metric scores model comparison

- * Input dimension: (40, 50), Positional encoding: False, Encoder blocks: 8, Heads: 1, qkv_dim: 25, ff_dim: 24 (~1mio.t.p.)
- Different models compared on Physionet data (26 classes) using train-test split (no cross-validation yet)
- Averages from all classes: accuracy, precision, recall f-measure (left)
- Physionets 2021 offical challenge metric evaluaton (right)

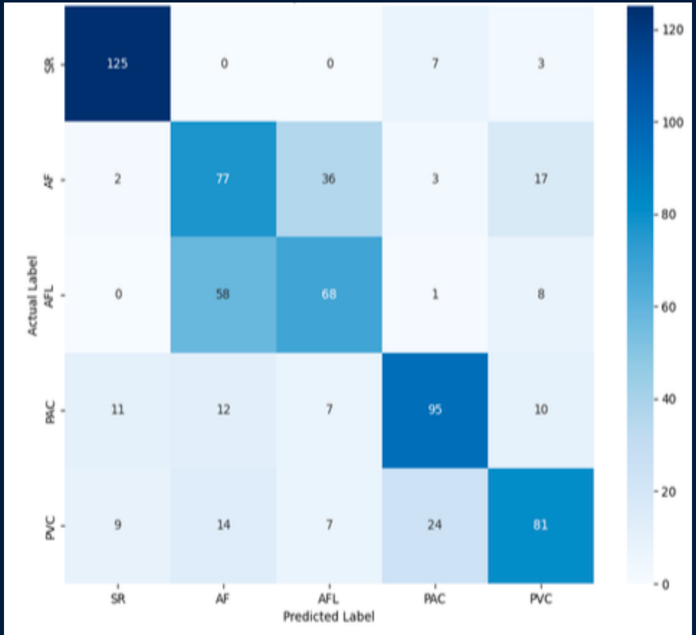
Experimental results: Residual CNN confusion matrix





- Multi-label classification problem turned into multiclassification problem
- Residual CNN from different runs, predicted on 100 test samples for each class

Experimental results: Residual CNN confusion matrix



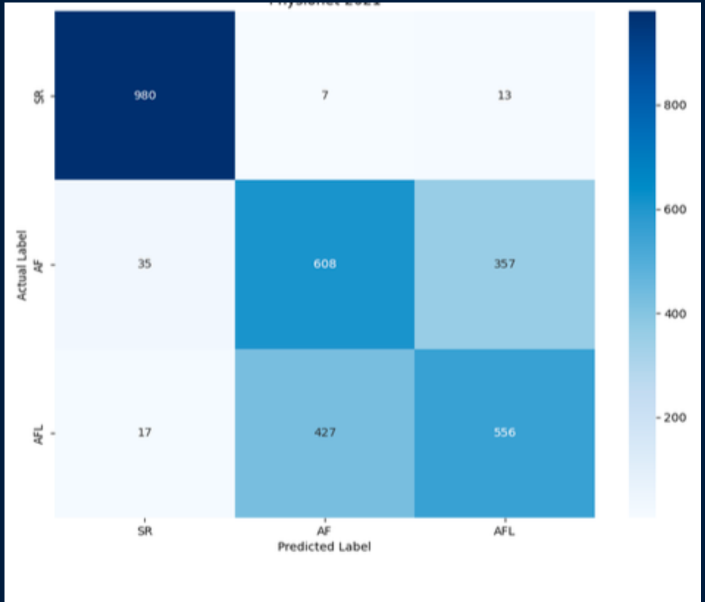


Figure 4.7: Confusion matrix - Residual CNN: SR, AF and AFL

Experimental results: MyDiagnostick data (todo)

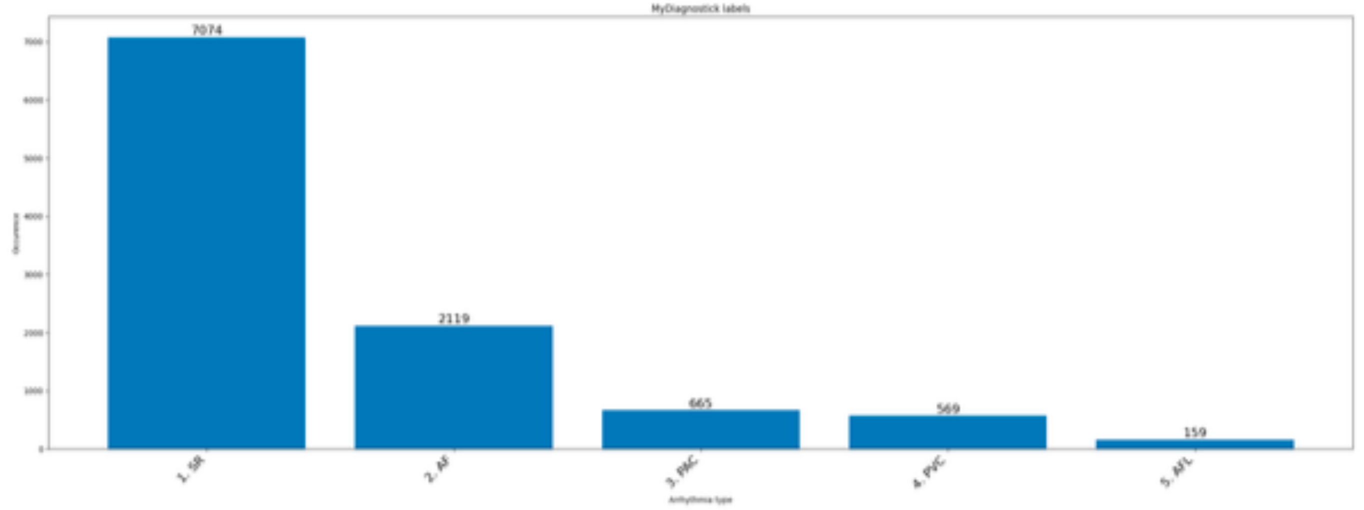


Figure 4.4: MyDiagnostick class distribution - SR, AF, AFL, PAC and PVC

Experimental results: MyDiagnostick data (todo)

Next

- Continue on thesis writing, evaluation and answer research questions in thesis
- Work on own approach (including own graphs)
- Stick to plan