The multimodal medical data preprocessing and classification framework

This project is dedicated to creating an analytic tool for preprocessing and visualization of medical data of several modalities such as text medical records and electrocardiograms. The obtained framework will allow getting as input multi-modal data and handling it in a unified manner. The final goal is to provide feature engineering and explanatory data analysis tool for further identification of cardiac diseases by solving machine learning classification problem.

Problem statement

Ischemic stroke is the most socially significant disease of the nervous system, it is the third cause of death in the developed countries after myocardial infarction and cancer. A significant portion of ischemic stroke cases is complicated by atrial fibrillation (AF). Despite the importance and prevalence of AF, there is still no sufficiently accurate method for its detection, especially in the early stages. Anticoagulants are prescribed to prevent cardioembolic stroke in patients diagnosed with AF and showing traits of additional risk factors. However, often, the AF is asymptomatic, the prophylactic therapy is not prescribed, and the disease immediately manifests itself as ischemic stroke. In some cases, the first attack of the obvious AF is complicated by the development of a stroke.

This project focuses on the discovery of additional predictors of cardioembolic stroke after or without AF manifestation. For this purpose, we intend to analyze the data extracted from the text descriptions of medical records (EHR) and from the time-series recordings of the electrocardiograms in twelve leads (ECG). We also aim to provide a visualisation tool to simplify the recommendation process for an expert. This will also help to identify a cohort of patients who would be prescribed anticoagulants to prevent stroke before the AF is confirmed and to develop personalized recommendations for screening patients at risk in the future.

Usually, patient’s data is studied and stored separately. Meanwhile, a combination of several modalities may profit in terms of more accurate disease predictions. There is no standard workflow with multi-modal datasets, so our aim is to create the first framework that can use all insights from such data.

Main challenges

The data storage culture affected medical organisations later than many other organisations so there are only several years of recording available to process. And there hasn’t been any strict policies regarding the storage and quality of data. So that, there are a lot of problems with data quality such as cases of broken external keys, making data incompatible or decreasing the amount of data to work with; human-made mistakes which increase preprocessing time. Medical data could be biased or changed. Bias may appear beginning with the patients that provide wrong or incomplete information and a doctor that could have a specific interpretation, ending with intentional changes in the medical data, for example, to increase profit from an insurance company. Also, patient monitoring is not constant and data

represented at discrete time points. So if some dramatic change happened after the last examination, it won't be shown in the data but will affect research data.
In order to create a feature generation pipeline for both modalities, domain expertise is required.

Baseline solution

According to privacy regulations, the medical data requires several steps of anonymisation and can't be provided to a second organisation without a set of agreements. So there are still no publically available datasets combining both ECG and EHR. And since datasets are not available, there are no benchmarks for both modalities together.
However, there are statistical approaches (HAVOC, cha2ds2-vasc [1, 2]) addressing this task for a single EHR modality. In the case of EHR, the ability to compare different datasets is questionable because of possible medical and language biases.
ECG is a more unified and independent modality, so there are state-of-the-art models that solve AF classification problems [3]. The shown results are good enough, however, most of them are the black-box solutions. In order to be useful to doctors for interpretation of the results, preliminary feature engineering and analysis is required. There are existing libraries that segment physiological signals [4], including delineating ECG signals into peaks and segments. These frameworks may be improved by adding generation physiologically meaningful features which are characteristic of a particular disease.

Roles for the participants

Ekaterina Ivanova

- To create a pipeline for ECG feature generation
- Feature analysis
- To create a visualisation tool
- To review project documentation

Nikita Khromov

- To create unified feature storage for the features of both modalities.
- To create a pipeline for text feature generation including a tool for annotation and a tool for creating additional context-free grammar-based features.
- Feature analysis
- AutoML for classification
- To review code documentation

Viktoria Chekalina

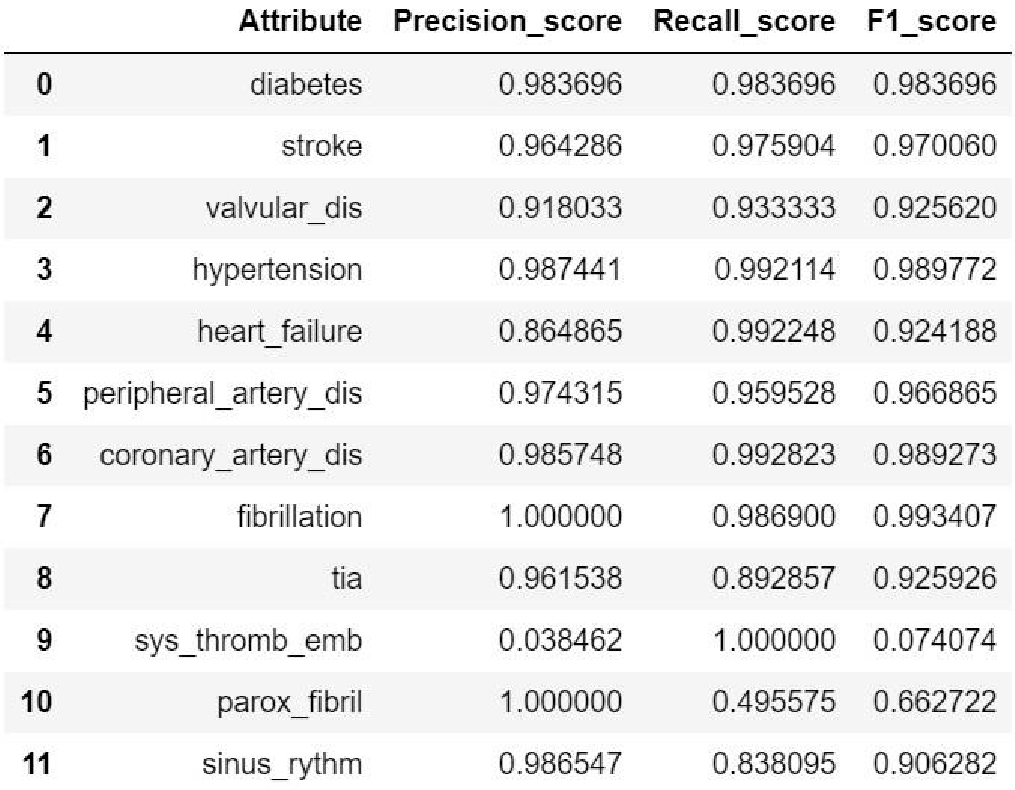
- To add SOTA NER models to the general text feature generation pipeline and an evaluation tool for them
- To train classification models
- Feature analysis
- To review project code

Second report

Text Feature Extraction

The second report is devoted to benchmarking. As it’s previously stated, there are some statistical scales like HAVOC or CHA2SK. It’s firstly reported that “A risk scoring system, the HAVOC score, was constructed using these 7 clinical variables that successfully stratifies patients into 3 risk groups, with good model discrimination (AUC=0.77).” [1]. Secondly, it’s reported in [2] that the risk of stroke in patients with atrial fibrillation (AF) can be assessed by use of the CHADS2 and the CHA2DS2-VASc score system. The main focus of this study is “that these risk scores and their individual components could also be applied to patients paced for sick sinus syndrome (SSS) to evaluate risk of stroke and death”

So, in order to create the proposed benchmarks two tasks have to be solved. The first task is to create a validation corpus to score feature extraction. The second task is to extract these features using context free grammar rules and/or utilize state of the art neural networks. There are both tools (Inception, Brat, etc.) and platforms(Lionbridge, Tagtog, Dataturks Text Annotation Tools, etc.) for annotation. Since the data of this project cannot be publicly distributed, any cloud based solution doesn’t meet the criteria. The brat annotation tool [3] was deployed and hosted at Skoltech local server with some randomly selected files for test and validation. These files were annotated by an expert later on. A link with instructions both on how to deploy a docker with brat provided in brat ipython notebook in a text features folder. This docker is deployed on a sholtech server. A tool for writing features based on context free grammars is provided. The current scoring for feature extraction is provided.

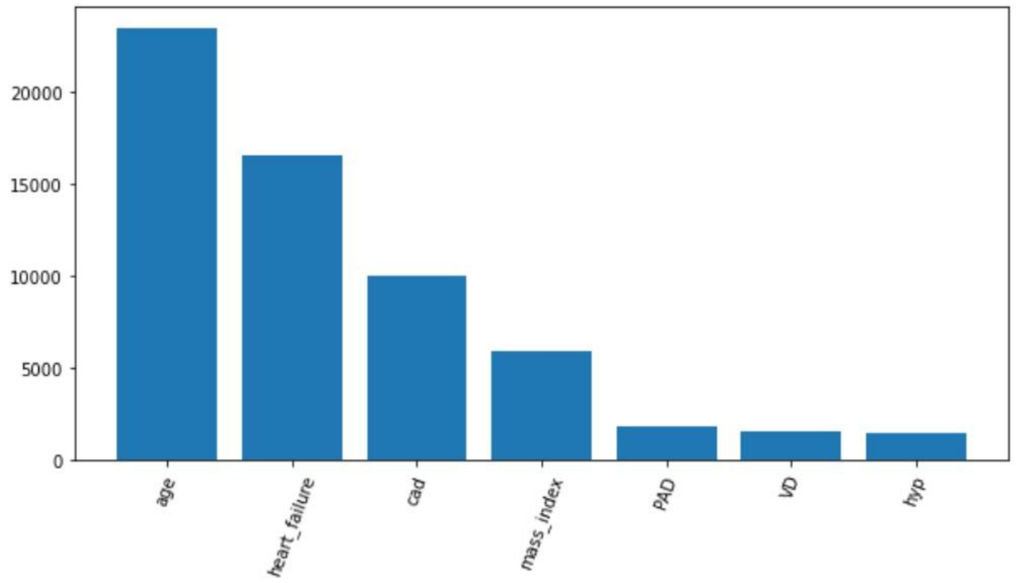


	Attribute	Precision_score	Recall_score	F1_score
0	diabetes	0.983696	0.983696	0.983696
1	stroke	0.964286	0.975904	0.970060
2	valvular_dis	0.918033	0.933333	0.925620
3	hypertension	0.987441	0.992114	0.989772
4	heart_failure	0.864865	0.992248	0.924188
5	peripheral_artery_dis	0.974315	0.959528	0.966865
6	coronary_artery_dis	0.985748	0.992823	0.989273
7	fibrillation	1.000000	0.986900	0.993407
8	tia	0.961538	0.892857	0.925926
9	sys_thromb_emb	0.038462	1.000000	0.074074
10	parox_fibril	1.000000	0.495575	0.662722
11	sinus_rythm	0.986547	0.838095	0.906282

According to this table some additional improvements are required for the parox_fibril entity. Sys_throm_emb is fine because there are only two entities overall in the validation corpus.

Text Feature Benchmarking

For building inicial benchmarks we use havoc and test its predictive power. The code is provided in the github repo and the outcome is the following: Using all the features it's possible to state, that approximate inicial prediction power of havoc is 0.65(ROC-AUC scale). The feature importance is provided here:



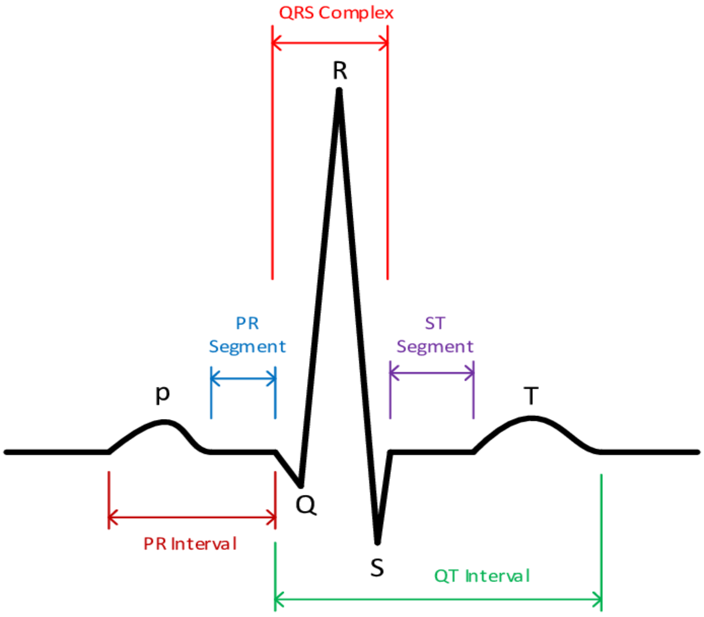
Overall, the detailed analysis for inicial benchmarking is provided in the automl jupyter notebook with havoc target calculations.

ECG Features Extraction

The ECG is a noninvasive representation of the electrical activity of the heart that is measured using electrodes placed on the torso. The standard 12-lead ECG is a key diagnostic tool used to assess the health conditions of the heart and it is widely used to diagnose a variety of cardiac arrhythmias such as atrial fibrillation. Recently there have been a lot of machine learning and deep learning methods for classifying ECGs. For clinicians one of the most important factors is the interpretability of the results of the classification algorithm. That is why in our work we focus on preliminary feature extraction.

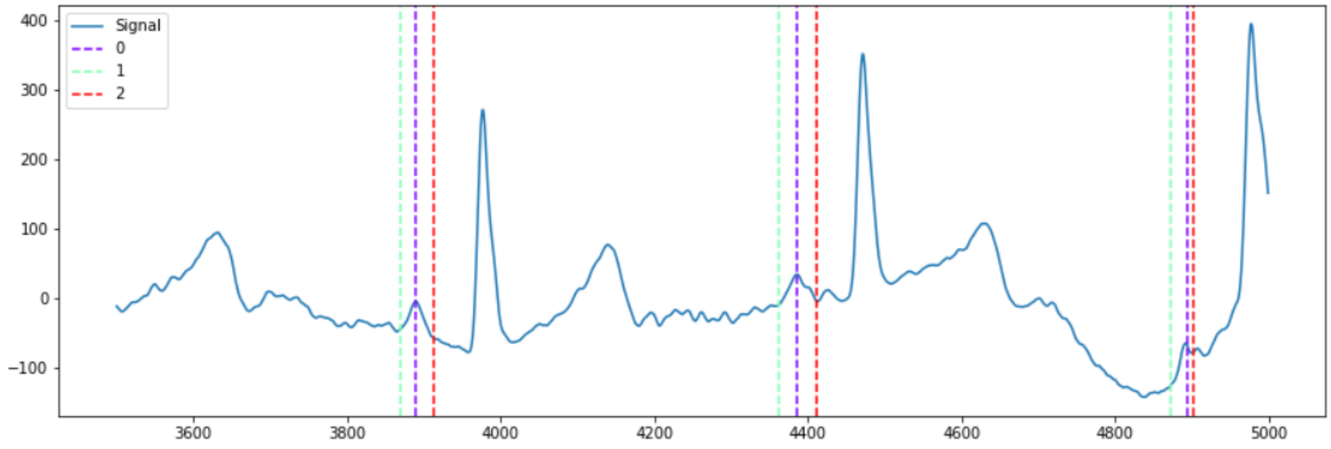
In a typical ECG cycle, its characteristic waveforms include: P wave, QRS wave (Q wave, R wave, S wave), T wave and U wave. Each segment is responsible for a certain contraction of the heart. Atrial fibrillation is characterized by the absence of a P wave, the presence of fibrillation waves and an aperiodic rhythm.

In general P wave is related to the condition of atrial activity. The P waveform is round and blunt, with small amplitude, and the shape of different leads is different, which is obvious on lead II and lead VF. The time width of P wave is usually between 0.08s and 0.11s, and the voltage (height) is usually between 0.22mV and 0.25mV. Deviations from this norm can be considered as abnormal associated with the atria.



According to the time of onset, AF can be divided into paroxysmal AF, persistent AF and permanent AF. A considerable proportion of patients with AF are asymptomatic and paroxysmal, making a timely diagnosis difficult to achieve. Although there may be no typical signs for atrial fibrillation, changes in atrial physiology can be traced in the characteristics of the P wave.

We extracted P wave using a discrete wavelet method and considered various P wave characteristics such as duration, amplitude, form, area under the wave and its derivatives. A complete list of signs characteristic of atrial fibrillation was compiled by a clinician.



Most updates for this report are presented here https://github.com/adasegroup/MMDF-multimodal-medical-features/tree/master/Text%20features

References

[1] Kwong, C., Ling, A. Y., Crawford, M. H., Zhao, S. X., & Shah, N. H. (2017). A clinical score for predicting atrial fibrillation in patients with cryptogenic stroke or transient ischemic attack. *Cardiology*, *138*(3), 133-140.]

[2] Olesen, J. B., Torp-Pedersen, C., Hansen, M. L., & Lip, G. Y. (2012). The value of the CHA2DS2-VASc score for refining stroke risk stratification in patients with atrial fibrillation with a CHADS2 score 0–1: a nationwide cohort study. *Thrombosis and haemostasis*, *107*(06), 1172-1179.

[3] Gari Clifford, Chengyu Liu, Benjamin Moody, Li-wei H. Lehman, Ikaro Silva, Qiao Li, Alistair Johnson, Roger G. Mark. AF Classification from a Short Single Lead ECG Recording: the PhysioNet Computing in Cardiology Challenge 2017. *Computing in Cardiology (Rennes: IEEE), Vol 44, 2017 (In Press).*

[4] Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*.

[5] Svendsen, J. H., Nielsen, J. C., Darkner, S., Jensen, G. V. H., Mortensen, L. S., Andersen, H. R., & DANPACE Investigators. (2013). CHADS2 and CHA2DS2-VASc score to assess risk of stroke and death in patients paced for sick sinus syndrome. *Heart*, *99*(12), 843-848.

[6] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012, April). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102-107).