The multimodal medical data preprocessing and classification framework

Ekaterina Ivanova [1] **Nikita Khromov** [1] **Viktoria Chekalina** [1]

Abstract

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.

1. Introduction

1.1. 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.

2. Baseline solutions and main challenges

2.1. 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.

2.2. 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 publicly 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 (Kwong et al., 2017), (Olesen et al., 2012) 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 (Clifford et al., 2017). The shown results are good enough, however, most of them are the black-box solu-

tions. 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 (Makowski et al., 2020), 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.

3. Methods

3.1. 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)." (Kwong et al., 2017). Secondly, it's reported in (Olesen et al., 2012) 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"

	Entity	precision	recall	f1	ACCURACY
0	diabetes	0.983784	0.989130	0.986450	0.993197
1	stroke	0.975904	0.964286	0.970060	0.993197
2	valvular_dis	0.983333	0.921875	0.951613	0.991071
3	hypertension	0.990581	0.992138	0.991359	0.985075
4	heart_failure	0.992308	0.865772	0.924731	0.971429
5	peripheral_artery_dis	0.972603	0.959459	0.965986	0.945726
6	coronary_artery_dis	0.985748	0.995204	0.990453	0.989116
7	fibrillation	0.986957	1.000000	0.993435	0.995918
8	tia	1.000000	0.892857	0.943396	0.995185
9	sys_thromb_emb	0.500000	1.000000	0.666667	0.966667
10	parox_fibril	1.000000	0.938596	0.968326	0.990476
11	sinus_rythm	0.993258	0.827715	0.902962	0.871099
12	AV1	0.861111	0.584906	0.696629	0.938497
13	tyroid	0.912281	0.912281	0.912281	0.977221
14	dia_dysf	0.859206	0.995816	0.922481	0.909297
15	LVH	0.804878	0.474820	0.597285	0.798186

Figure 1. Metrics for text features extraction

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 (Stenetorp et al., 2012) 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.

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.

3.2. Text Feature Benchmarking

After text feature extraction we tested its quality on a validation dataset which consists of 750 annotated electronic health records. The results are presented in the following table.

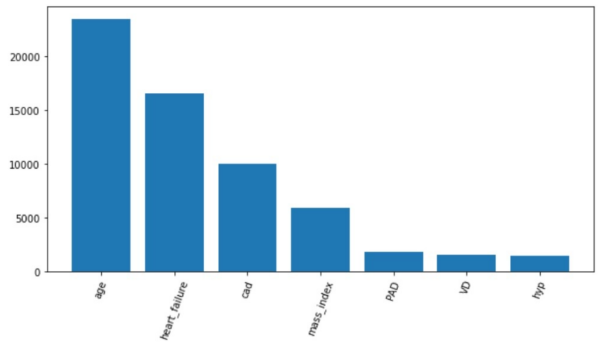

Figure 2. Feature importance for 7 most important text features

It can be observed from the table that the quality of target features are above 0.97. It's enough to proceed to the next step of analyzing the information. Some features with low F1 are not included to the main classification analysis task. We also tried Deep Pavlov Bert model (the best BERT model available for Russian language)

Entity	Rule-based extraction F1	Model extraction F1
Hypertension	0.98	0.97
PAD	0.94	0.73
Diabetes	0.98	0.90
CAD	0.99	0.88

These numbers require some explanation. The first issue is that this BERT for Russian language was trained on Russian Wikipedia and articles database. It's not as good as kind of the BERT trained on the medical data. Some additional fine tuning or full model retraining on medical corpus could improve model quality. Unfortunately, there are no available medical datasets in Russian for retraining and transfer learning. For the purpose of this task only features extracted by context free grammar based rules because their F1 quality is significantly better.

For building initial 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 initial prediction power of havoc is 0.65(ROC-AUC scale). The feature importance is provided here: Overall, the detailed analysis for initial benchmarking is provided in the automl jupyter notebook with havoc target calculations.

3.3. 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 (fig.3), 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

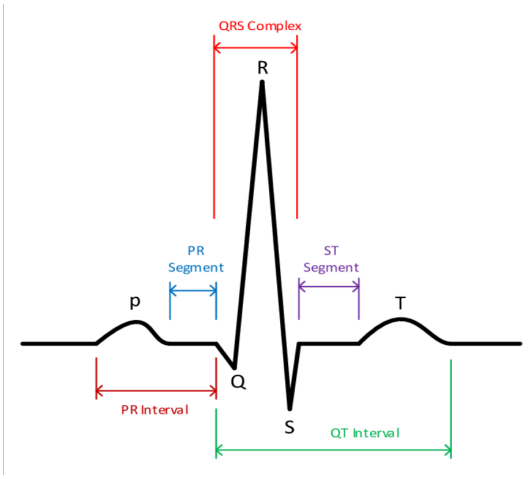

Figure 3. Example of standard cardiocycle in lead II. All peaks and segments are labelled.

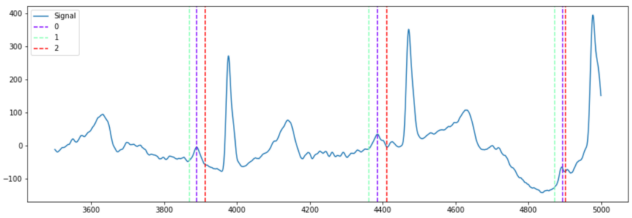

Figure 4. Segmentation of P wave, that is responsible of atrial contraction. Atrial fibrillation is manifested by changes in this particular area of ECG signal

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.

- Duration of P wave
- Maximum of P wave duration (within all leads)
- Difference between minimum and maximum duration of P wave in each lead
- Biphase P wave in leads II, III, aVF
- P wave amplitude

- Time from P wave peak start till its absolute maximum value in leads V1 and II
- Area under P wave curve
- Index area under P wave curve/Duration
- PQ interval duration
- Terminal index: production of amplitude on duration of negative phase of P wave in V1 lead

4. Results

Within the framework of this project was implemented following steps. The first one is a feature generation pipeline that transform multi-lead ECG time series to physiological meaningful features. The list of the features are provided by the clinicians and represented atrial fibrillation abnormalities in ECG signal. Moreover, any other deceases related to malfunction of human atria may be catch by these features as well. As the next step, we created a feature generation pipeline that extract important information from medical electronic health records such as age, diagnosis, date of diagnosis.

The result of our project is described in Visualization section. We present a histogram distribution for person feature values (age, sex, weight) for people who have and have not heart disease. We attempt to find a correlation between age/weight and disease presence, but it turned out that there is no correlation.

We also observe a correlation between the feature ”sinus rhythm” and disease presence and in this case the correlation is turned around.

We create a function to illustrate the changes in one of the electrocardiogram parameters for a patient with fixed ID during the time. It can be useful both in monitoring the patient's conditions during the time and in finding the dependencies of the ECG parameters on the presence / absence of a symptom (provided that at different time steps a symptom may appear or not).

We also provide a functionality that extracts from initial set of features ones that correspond to one group (for example, group of features can made up by records about drug dose or value of ECG features). For every person we have a feature vector of features and map this vector to 2D dimension using either PCA (model finding or TSNe(finding k principal component by changing basis) or t-SNE (model finding two-dimensional representation by minimizing the Kullback-Leibler divergence between the joint probabilities of coordinates).

We also use XGBoost classifier to select the most important features from text group of features and ECG group of features. The result of it are on fig.5, fig.6. As we can see in the figures, the most important features are *thyroid_1* - overactive thyroid and *MI* - Myocardial infarction. For ECG the most important feature is *max_dur_p_1*

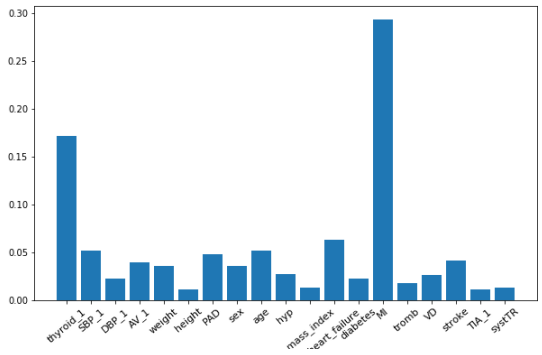

Figure 5. Feature importance for text features.

All code and documentation can be found at our github repository.

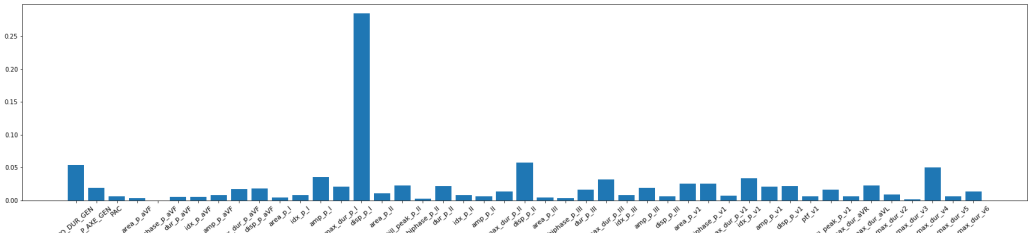

Figure 6. Feature importance for the EGG features.

References

Clifford, G. D., Liu, C., Moody, B., Li-wei, H. L., Silva, I., Li, Q., Johnson, A., and Mark, R. G. Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pp. 1–4. IEEE, 2017.

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

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Schoelzel, C., and Chen, S. A. Neurokit2: A python toolbox for neurophysiological signal processing. 2020.

Olesen, J. B., Torp-Pedersen, C., Hansen, M. L., and Lip, G. Y. 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, 2012.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. 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, 2012.

A. Team member's contributions

Ekaterina Ivanova (% of work)

- ECG dataloader implementation
- Segmentation and signal preprocessing
- Creation of a pipeline for ECG feature generation
- Providing project documentation
- Creation of a SQL based feature storage

Nikita Khromov (% of work)

- Creation of a pipeline for text feature generation including a tool for annotation and a tool for creating additional context-free grammar-based features.
- Feature analysis
- Provide review of code documentation
- AutoML for classification

Viktoria Chekalina (% of work)

- Deployment boost classifier and feature selection to ECG-Text features.
- Creation analysing tool
- Creation visualization pipeline
- Provide review of code documentation
- Providing project documentation

B. Reproducibility checklist

Answer the questions of following reproducibility checklist. If necessary, you may leave a comment.

1. A ready code was used in this project, e.g. for replication project the code from the corresponding paper was used.

   ☐ Yes.
   ☑ No.
   ☐ Not applicable.

   General comment: If the answer is **yes**, students must explicitly clarify to which extent (e.g. which percentage of your code did you write on your own?) and which code was used.

   Students' comment: None

2. A clear description of the mathematical setting, algorithm, and/or model is included in the report.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

3. A link to a downloadable source code, with specification of all dependencies, including external libraries is included in the report.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: Link are provided above in the report body.

4. A complete description of the data collection process, including sample size, is included in the report.

   ☐ Yes.
   ☐ No.
   ☑ Not applicable.

   Students' comment: Medical dataset is private, so we did not provided original text dataset. ECG features are calculated on publicly available data. More detail in github documentation.

5. A link to a downloadable version of the dataset or simulation environment is included in the report.

   ☐ Yes.
   ☐ No.
   ☑ Not applicable.

   Students' comment: Same comment as for previous question.

6. An explanation of any data that were excluded, description of any pre-processing step are included in the report.

   ☑ Yes.
   ☐ No.
   ☐ Not applicable.

   Students' comment: None

7. An explanation of how samples were allocated for training, validation and testing is included in the report.

   ☐ Yes.
   ☐ No.
   ☑ Not applicable.

   Students' comment: None

8. The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results are included in the report.

   ☐ Yes.
   ☐ No.
   ☑ Not applicable.

   Students' comment: None

9. The exact number of evaluation runs is included.

   ☐ Yes.
   ☐ No.
   ☑ Not applicable.

   Students' comment: None

10. A description of how experiments have been conducted is included.

    ☑ Yes.
    ☐ No.
    ☐ Not applicable.

    Students' comment: None

11. A clear definition of the specific measure or statistics used to report results is included in the report.

    ☐ Yes.
    ☐ No.
    ☑ Not applicable.

    Students' comment: None

12. Clearly defined error bars are included in the report.

- ☐ Yes.
- ☐ No.
- ☑ Not applicable.

Students' comment: None

13. A description of the computing infrastructure used is included in the report.

- ☑ Yes.
- ☐ No.
- ☐ Not applicable.

Students' comment: None