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 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 problem [3]. The showed 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

A link to the GitHub repository:
https://github.com/adasegroup/MMDF-multimodal-medical-features

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*.