The multimodal medical data preprocessing and classification framework

Final project for Foundations of Data Science course

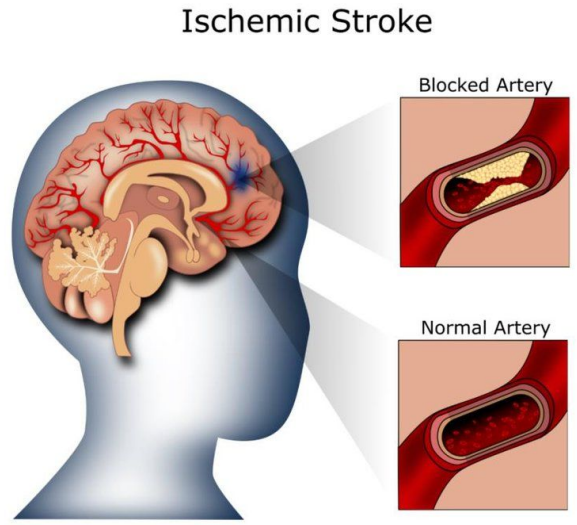
Skoltech

Ekaterina Ivanova
Nikita Khromov
Viktoria Chekalina

Problem statement

Ischemic stroke is the most socially significant disease of the nervous system

- third cause of death in the developed countries
- often connected and complicated by atrial fibrillation (AF).
- no sufficiently accurate method for atrial fibrillation detection



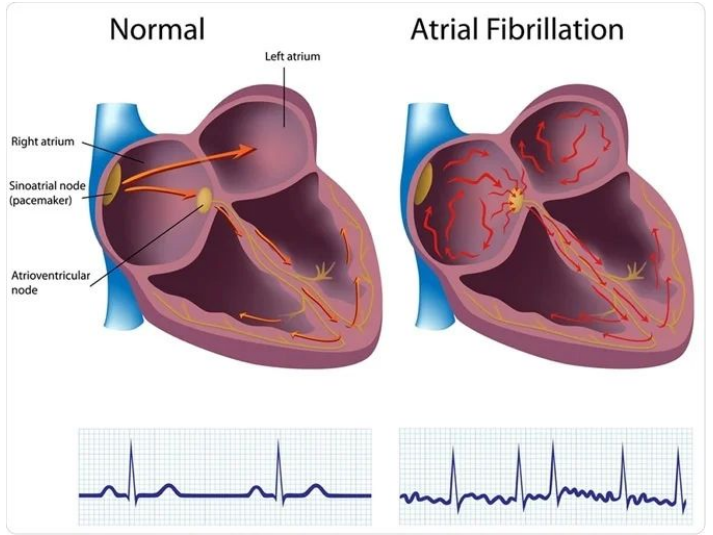
AF is asymptomatic, especially on early stages

⇩

Prophylactic therapy is not prescribed

⇩

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.

Skoltech

Problem statement

Problem: patient's data is studied and stored separately. There is no standard workflow with multi-modal datasets.

A combination of several modalities may profit in terms of more accurate disease predictions.

Aim:

- to create the first framework that can use all insights from data extracted from the text descriptions of medical records (EHR) and from the time-series recordings of the electrocardiograms in twelve leads (ECG)

- to provide a visualisation tool to simplify the recommendation process for an expert

ECG Data

ECG data

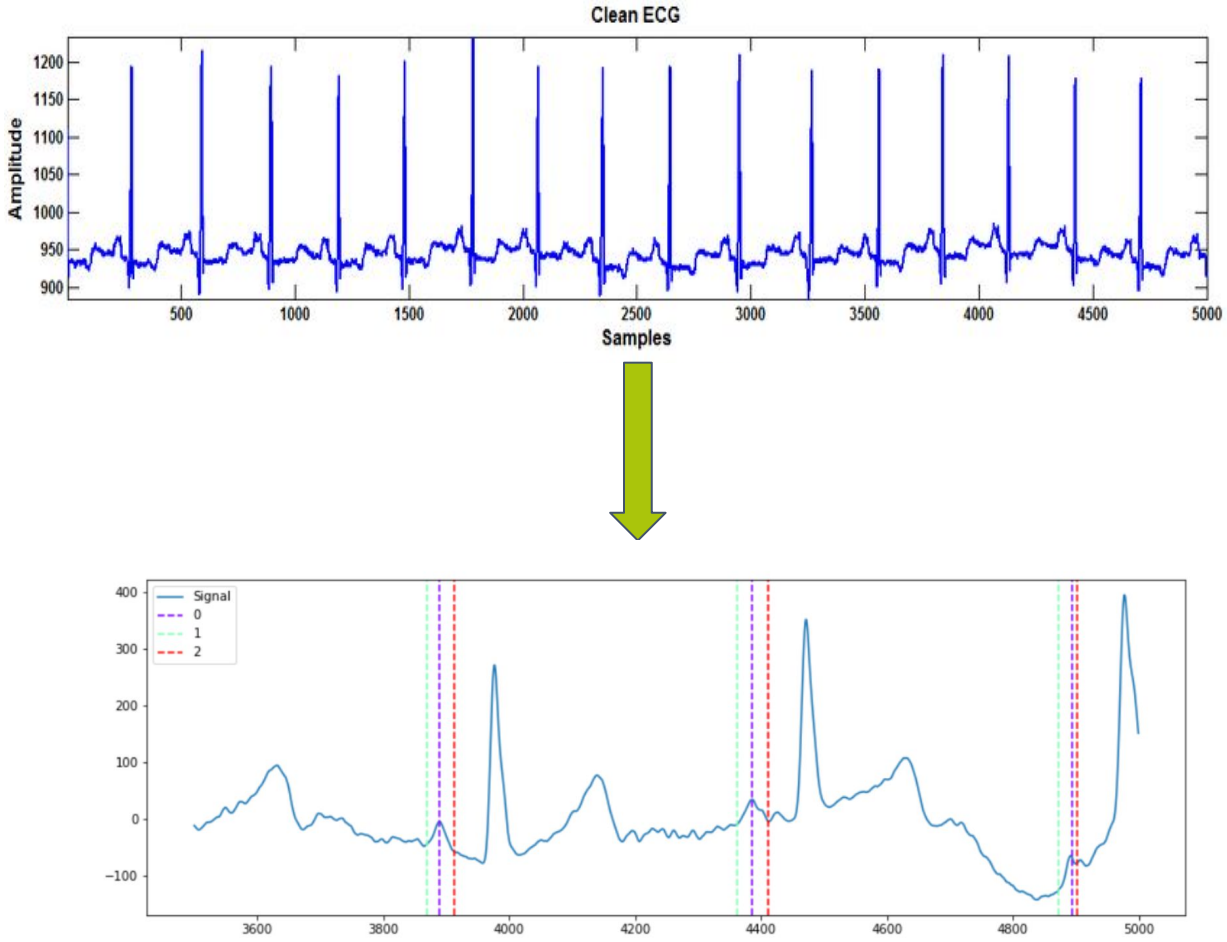
Provided for the PhysioNet/Computing in Cardiology Challenge 2020

China 12-Lead ECG Challenge Database

- 3,453 (male: 1,843; female: 1,610)
- 12-lead ECG recordings
- lasting from 6 seconds to 60 seconds.
- sampled at 500 Hz

Skoltech

ECG Signal Segmentation

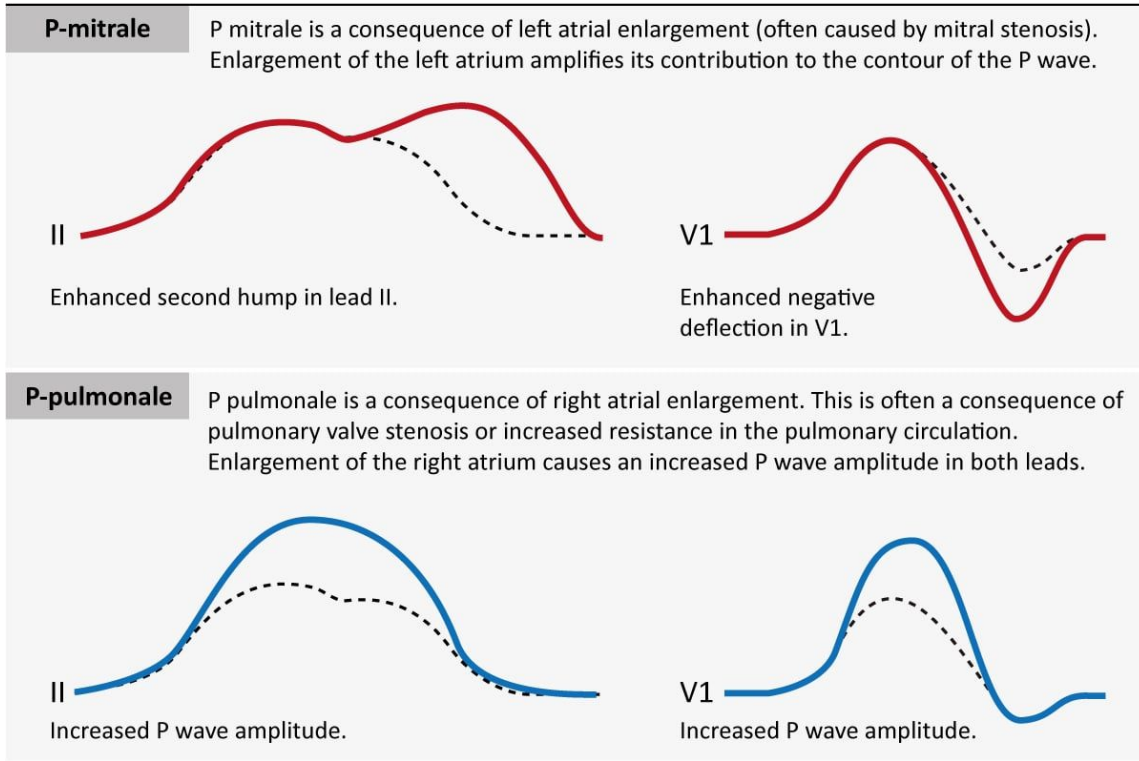


- P wave are represent atria conduction
- Segmentation of ECG is a separate hard challenge
- In our project for segmentation we use Neurokit python package

Skoltech

ECG feature generation

- 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 и 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



Skoltech

EHR Textual Data

Main challenges:

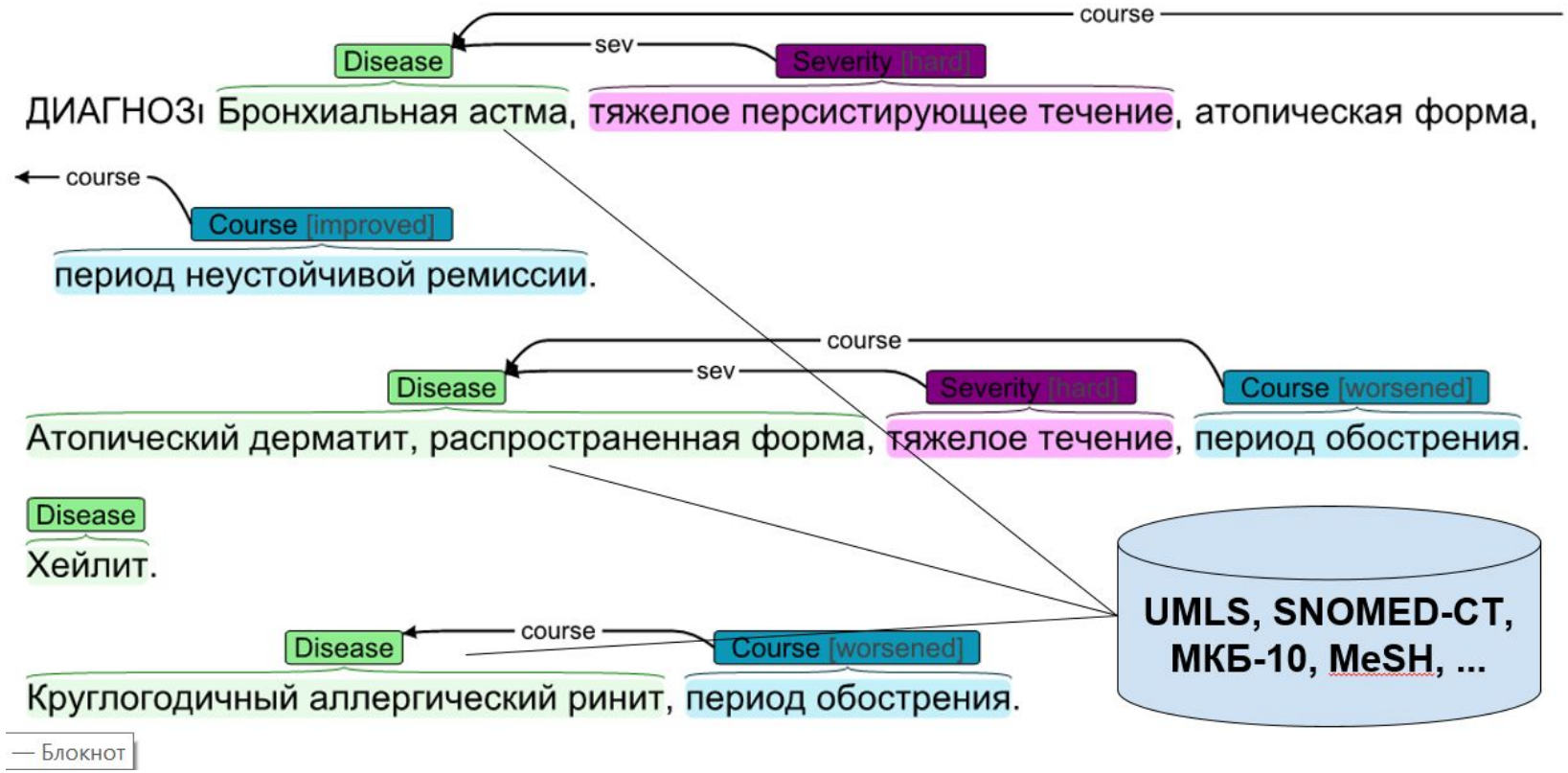
- No state of the art datasets medical textual data in Russian
- No trained models for medical data in Russian
- Only statistical approaches exist for evaluation

Our dataset:

- 164 000 EHR
- Noisy with many typos and errors
- Some keys are broken

Skoltech

EHR Diagnosis Sample



Skoltech

Text Feature Extraction

Pipeline:

- Restoring keys
- Restoring broken htmls
- Rules for some common typos

Approaches:

- Yargy parser
- BERT based models for Russian

Skoltech

Text Feature Extraction Quality Comparison

Attribute To extract	Rules F1	Bert Models F1(avg)
Hypertension	0.98	0.96
PAD	0.94	0.67
Diabetes	0.98	0.90
CAD	0.99	0.88

Tested on:

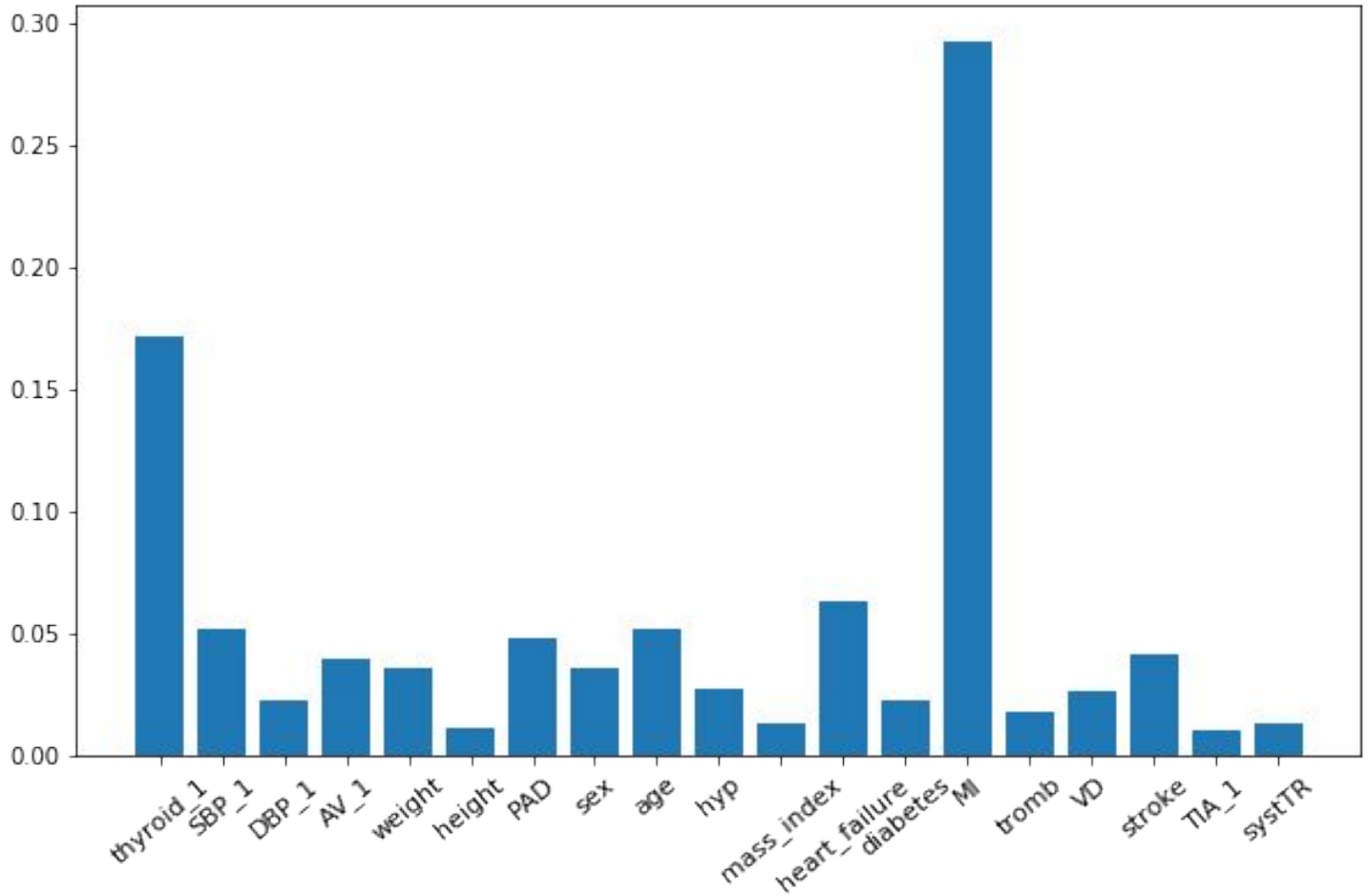
- Dataset labeled by expert (750 fully annotated EHR)

Reasons:

- BERT based models were trained on non medical data
- Not enough samples in the training corpus

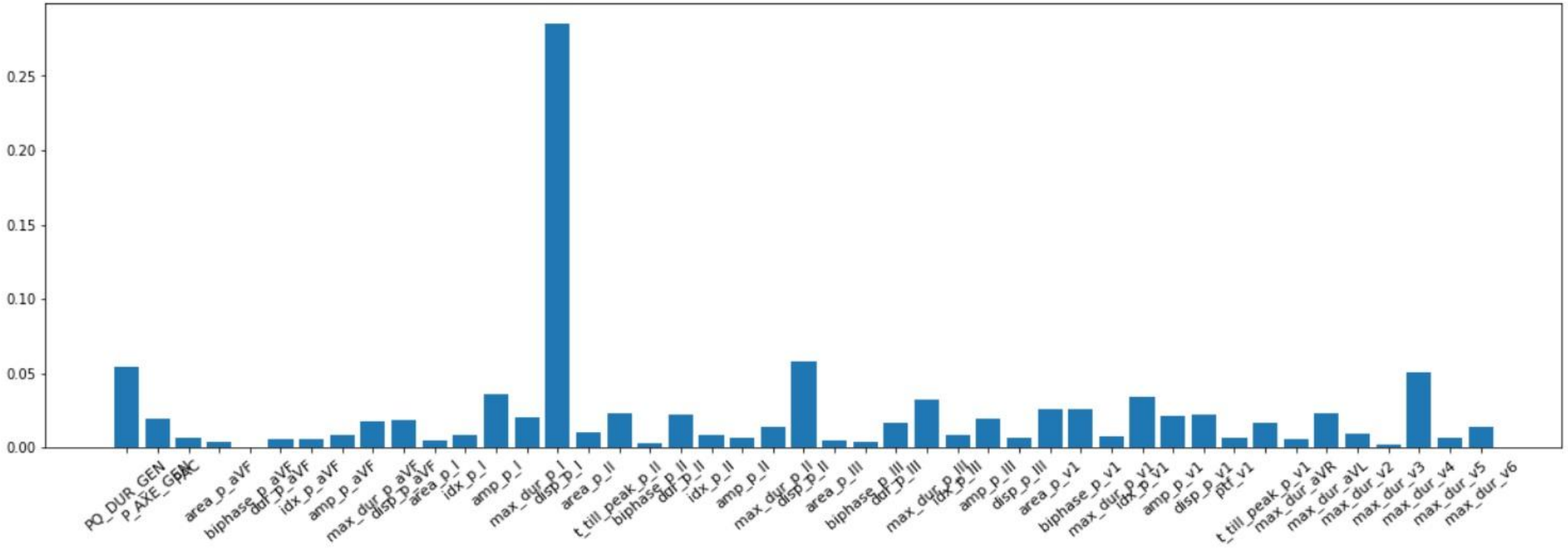
Skoltech

Results
Text Feature Importance



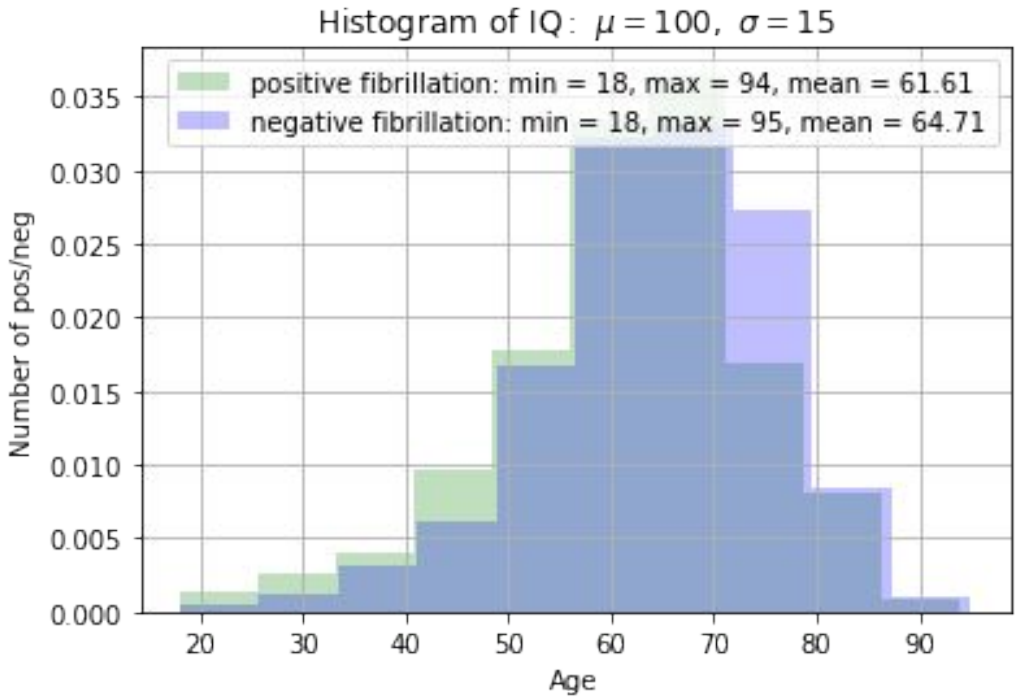
Skoltech

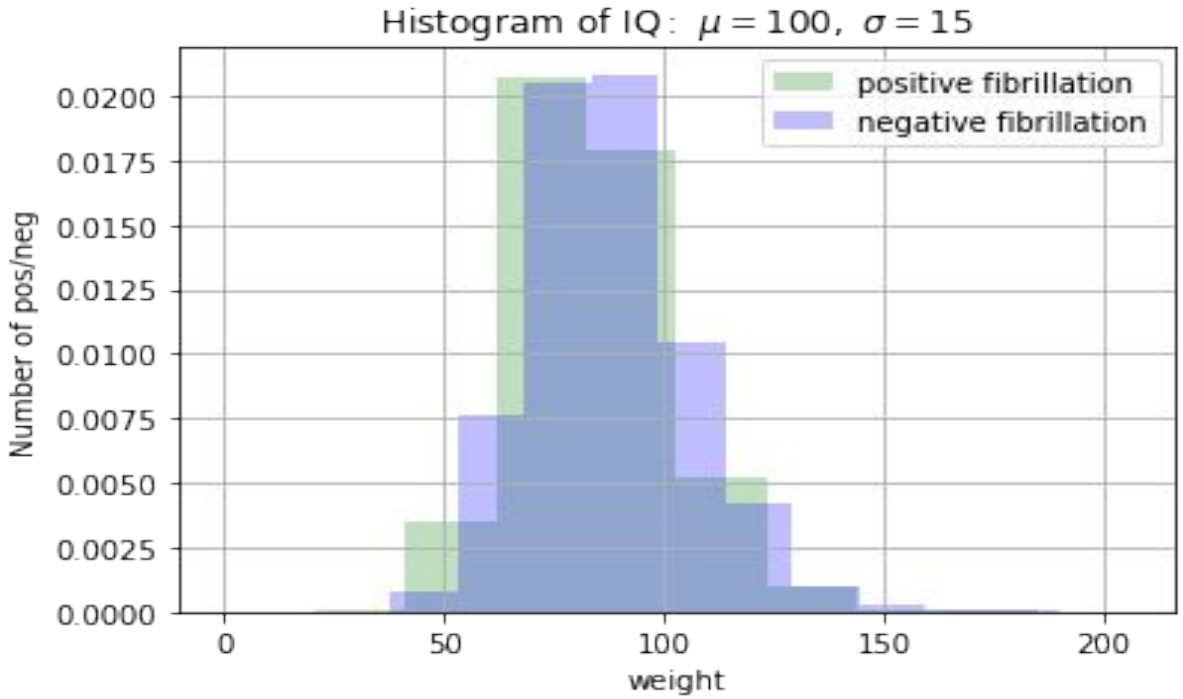
Results
ECG Feature Importance



Skoltech

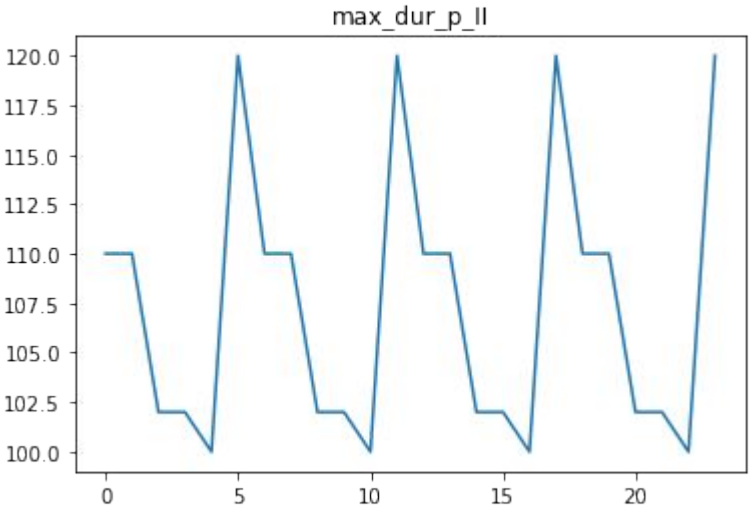
Correlation Analysis for Age and Weight

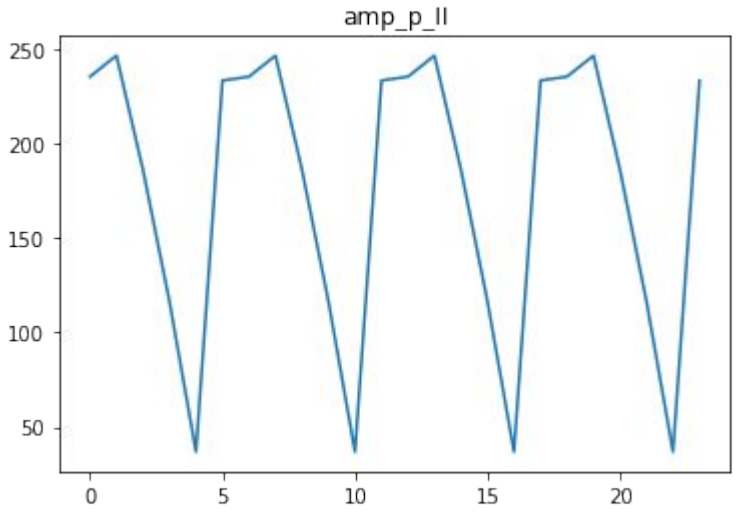


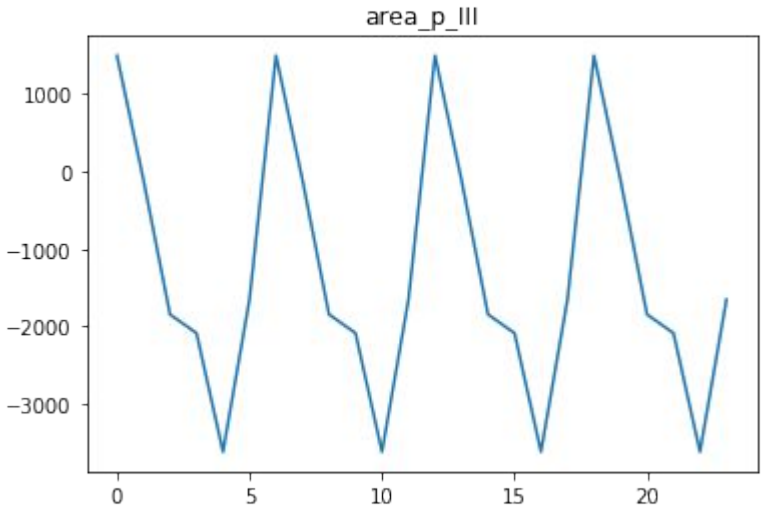


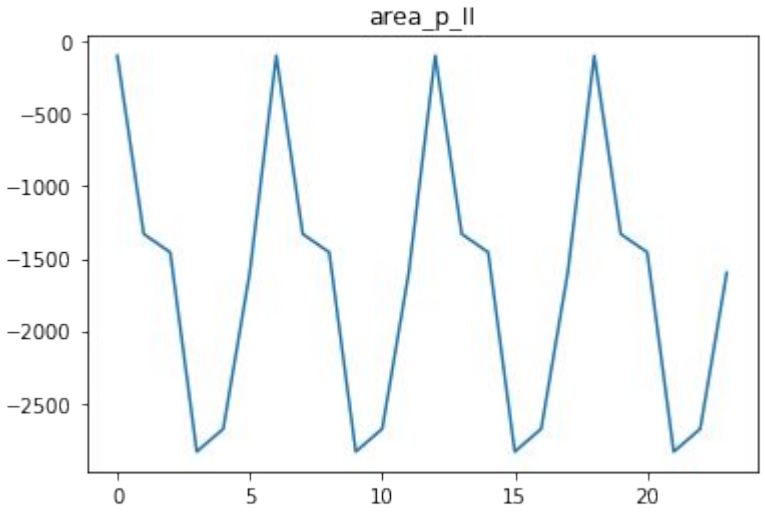
Skoltech

Time Dynamics









Skoltech

Tools Implemented

1. ECG feature generation
2. Text feature generation
3. Visualization and data analysis
4. SQL tool for data storage

thx.

Skoltech