Predicting respiratory diseases from lung sounds using machine learning

Richard Annilo, Computer Science Bachelor, year III, University of Tartu, Institute of Computer Science
Supervisor: Dmytro Fishman

The purpose of this project was to create a freely accessible codebase that would make it easier for machine learning researchers to create innovation in the medical field. Combining machine learning with the medical field would decrease false diagnoses and **save lives**.

This project focused on **predicting lung diseases** from a breathing sounds dataset. 8 experiments were conducted on 5 different machine learning models. In addition, a novel data augmentation algorithm was performed and tested. This in total took 223 hours.

All of the results are packaged in a **freely accessible codebase**.

Implementation

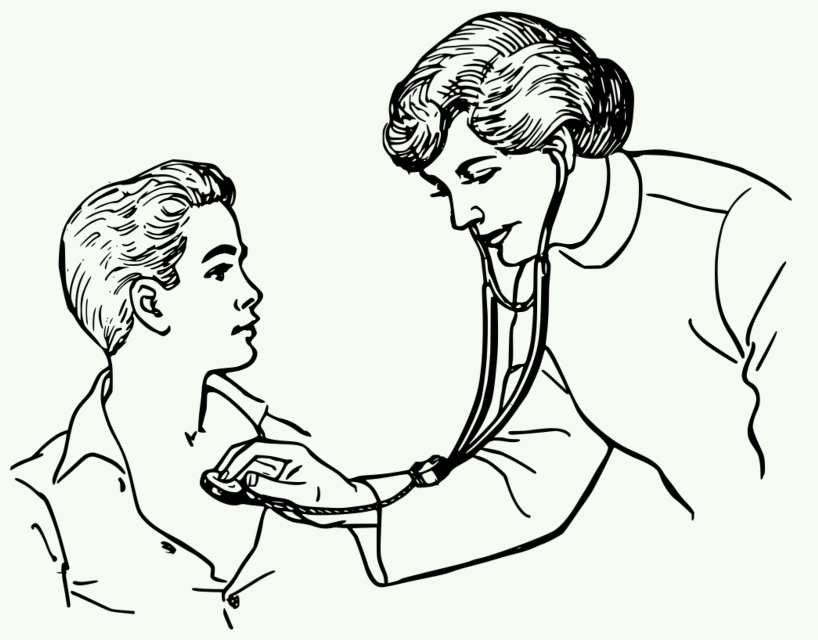
The following experiments were conducted on 4 classical machine learning methods:

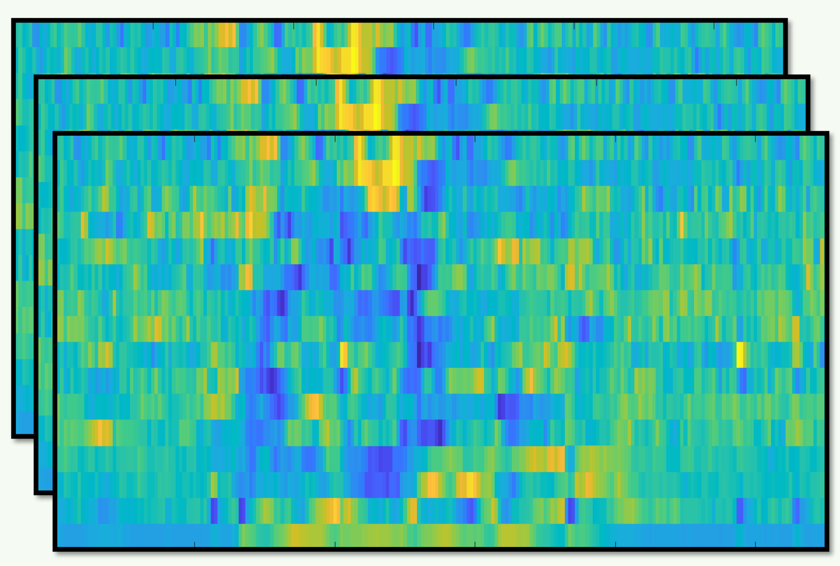
- using **all features** to train the models,
- using **less complex** models to decrease overfitting,
- using **class weights** to counter dataset unbalancedness,
- using **fewer features** to decrease noise in the data.

Experiments on the deep learning model:

- using **all features** to train the model,
- using **class weights** to counter dataset unbalancedness,
- using a novel **data augmentation algorithm**.

The following models were used: decision tree classifier, random forest, support vector machine, XGBoost and CNN.





	Decision tree	Random forest	XGBoost	SVM
All features	0.4956 +/- 0.12	0.4195 +/- 0.13	0.4461 +/- 0.18	0.1140 +/- 0.003
Less complex	0.4875 +/- 0.11	0.4175 +/- 0.13	N/A	0.1129 +/- 0.002
Class weights	**0.5716 +/- 0.10**	0.4338 +/- 0.18	N/A	0.1423 +/- 0.08
Fewer features	**0.5749 +/- 0.12**	**0.4499 +/- 0.15**	**0.4693 +/- 0.21**	**0.2629 +/- 0.06**

	CNN
All features	0.2637 +/- 0.06
Class weights	**0.3123 +/- 0.08**
Data augmentation	0.2357 +/- 0.1

MFCC

Mel-frequency cepstral coefficients (MFCCS) were used as inputs for the deep learning model. These are representations of sound, which give information about **frequencies** and their **intensities** in time. MFCCs are designed to closely imitate the functioning of the human ear: lower frequency ranges increase linearly while higher ranges increase logarithmically.

Data augmentation algorithm

Because the diagnosis distribution for sounds was especially unbalanced, it was chosen to try data augmentation to improve this. The algorithm goes as follows:
For the majority class patients, MFCCs are extracted as normal.
For others, the sound was split in pieces based on the respiratory cycles.
After that, the cycles are shuffled and a new sound file is created. MFCCs are extracted from that sound.

Results

The results show that on the particular dataset, which was highly unbalanced (6 times more COPD patients than pneumonia patients), the simplest classical machine learning model – the decision tree – came out ahead. In addition, decreasing the number of features and implementing class weights helped.

The data augmentation algorithm did not work as intended, perhaps because it was used too excessively.

All the scores are the macro-average of f1-scores, meaning it includes each diagnosis class equally when calculating the f1-score.

UNIVERSITAS TARTUENSIS

Project website:
https://lingid.ee/Cg660

IT Akadeemia