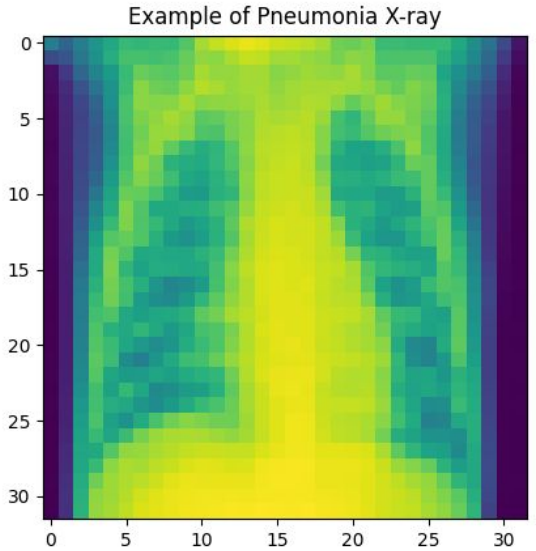
Model search for X-ray data 

Prepared by Grigorii Antonov (me)

Motivation

- High amount of medical images
- Low amount of medical specialists
- **Result**: delays in declaration of diagnosis and start of curing



About data

- **Chest X-ray images**; were selected from retrospective cohorts of pediatric patients of one to five years old.
- The dataset is organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal)
- **Unbalanced labels**: 70% of the "Normal" class

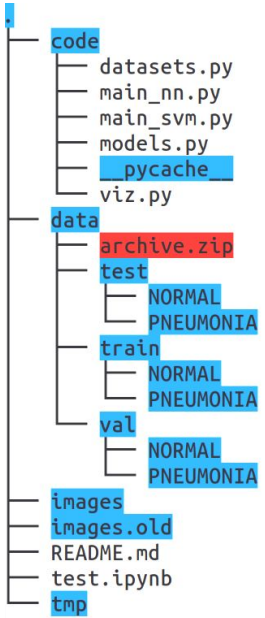
Data: https://data.mendeley.com/datasets/rscbjbr9sj/2

Project structure

Separate files for models, datasets and visualization operated by main file

Separate folders for images, code and data

Separate virtual environment to run the code without conflicts



Aim and methods

Aim: compare different lightweight models to find the most efficient one for the selected data; find the optimal classificator to distinguish normal condition from pathological

Methods:

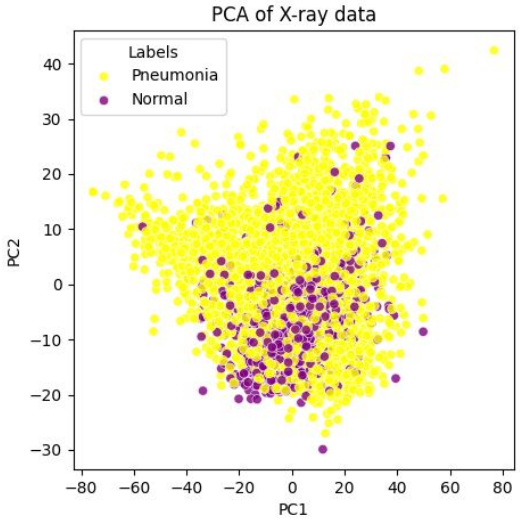
Dimension reduction methods: PCA, tSNE, UMAP

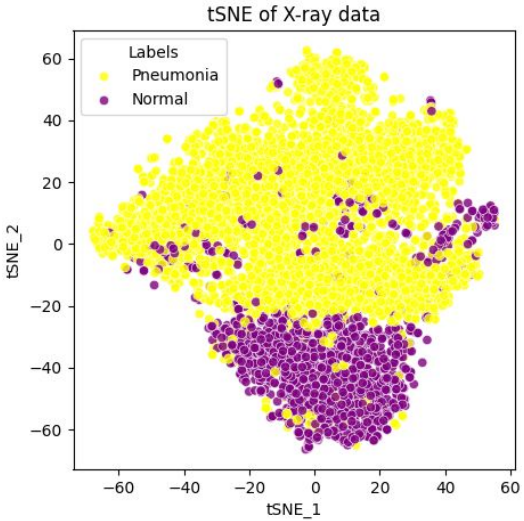
Support Vector Classification

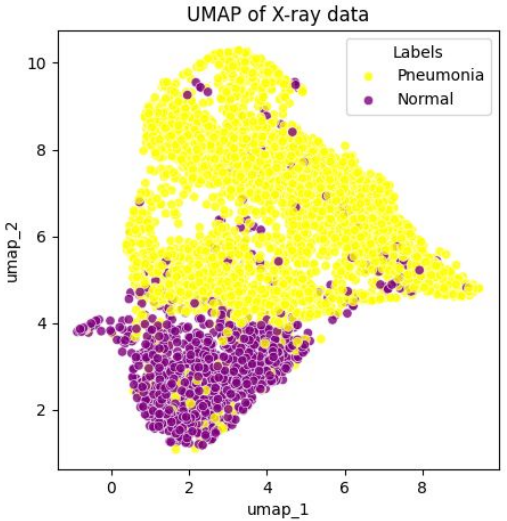
Convolutional Neural Networks

Transfer Learning

Different dimensional reduction method

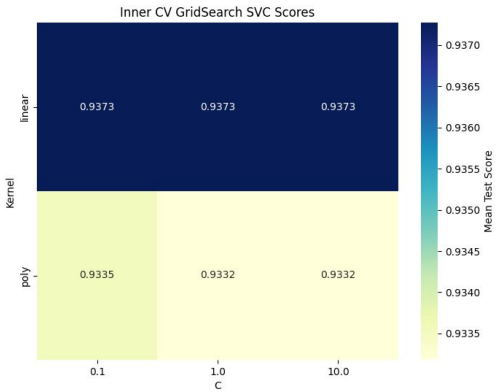


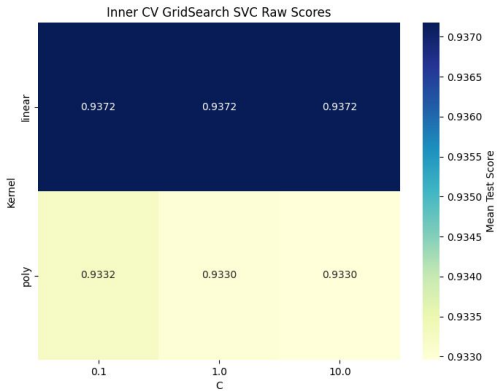


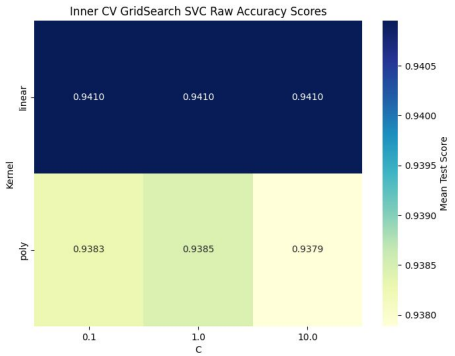


Good separable clusters for tSNE and UMAP even on low-dimentional data

Nested Grid Search for optimal SVC Model







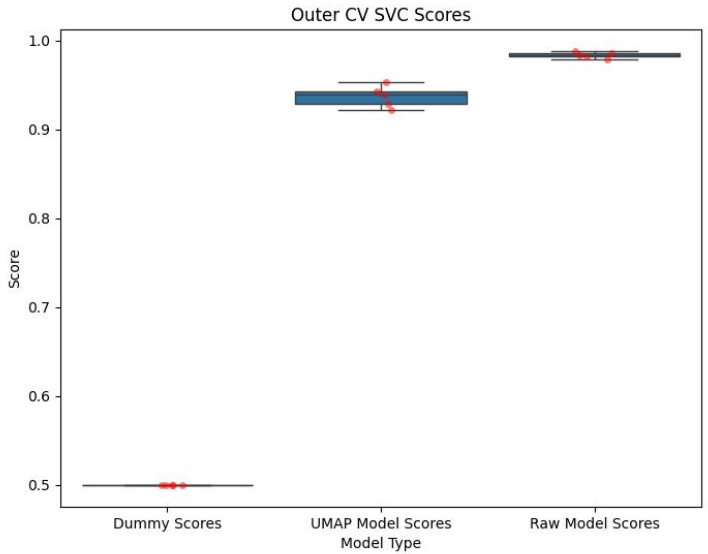
Best C for inner CV: 0.1 **Best kernel for inner CV**: linear **Best score for outer CV**: 0.961

Comparison of Models AUC scores

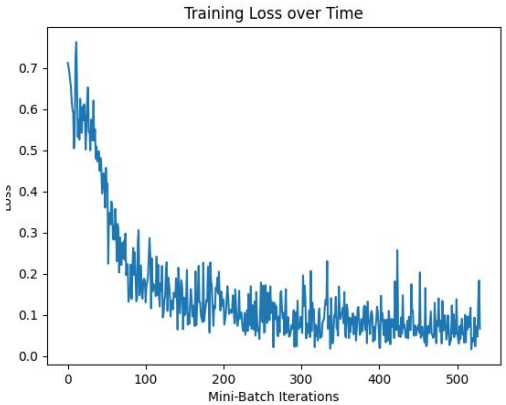
Baseline is represented by the "dummy model"

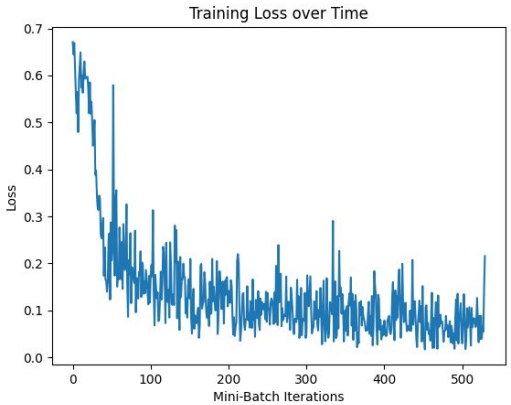
- as a dummy model sklearn module was used with most frequent option

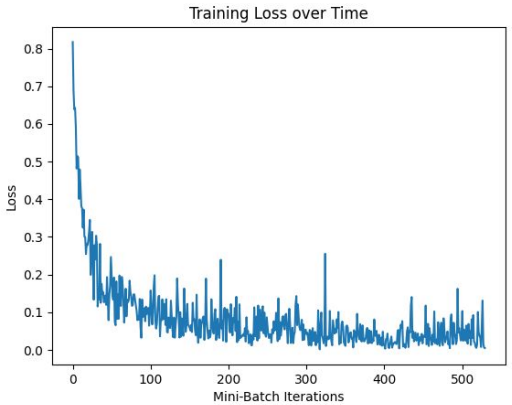
Best AUC was shown with SVM (linear, C = 1) on raw data



Testing different RNN Models



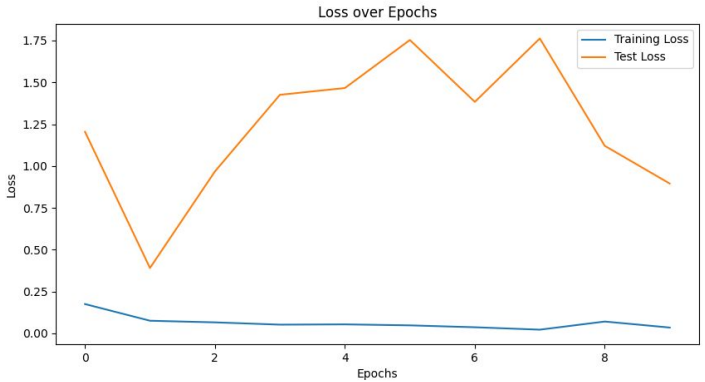


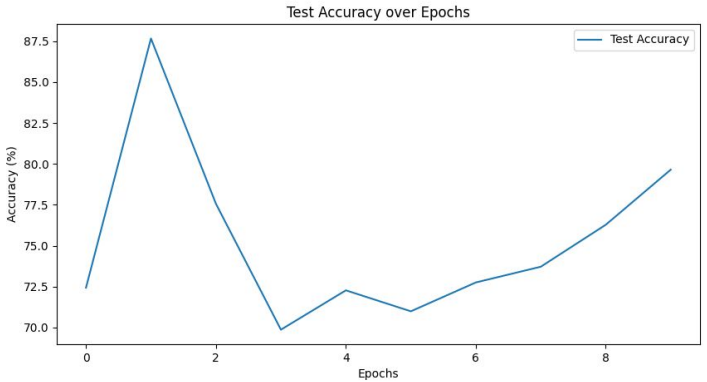


The best accuracy I achieved: 75% (refined RNN with 2 conv and 3 fc layers)

Techniques: dropout, batch norm, Adam optimizer, data augmentation (a bit)

Transfer Learning with ResNet18





Either **refining** all parameters of the models as well as training only the last FC layer and **freezing** the other ones

Best accuracy achieved: 87% **Problem** - fast overfitting

Conclusion

Sometimes simple decision are better that fancy ones > do not overcomplicate