

# ANALYSIS OF COVID-19 CHEST X-RAYS:

## Report 3: Final Report

Saniya Arfin, Yvonne Breitenbach, Alexandru Buzgan

June 2025

### Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Dataset Overview - Description of the "Chest-X-Ray" Dataset</b>	<b>2</b>
<b>3</b>	<b>Modelling</b>	<b>2</b>
3.1	Introduction to modelling . . . . .	2
3.2	Step 1: Machine Learning Techniques . . . . .	2
3.2.1	Logistic Regression . . . . .	3
3.2.2	Random Forest . . . . .	3
3.2.3	Support Vector Machines (SVM) . . . . .	3
3.3	Step 2: Transition to Deep Learning . . . . .	3
3.4	Comparing the Results . . . . .	4
<b>4</b>	<b>Conclusion and Future Work</b>	<b>4</b>
<b>5</b>	<b>Discussion - Reflecting on our project</b>	<b>4</b>
<b>6</b>	<b>References</b>	<b>5</b>
<b>7</b>	<b>Appendices</b>	<b>5</b>

# 1 Introduction

Some ideas:

Background and Project Motivation

Describe the context of the project, the public health importance of chest X-ray analysis, and the impetus for the study.

Outlines the background, motivation, objectives, scope, and structural layout of the report.

# 2 Dataset Overview - Description of the "Chest-X-Ray" Dataset

Some ideas:

Describes the "Chest-X-Ray" dataset in detail, discussing its composition, class imbalance issues, and data quality challenges, along with the preprocessing techniques applied.

source of the data

# 3 Modelling

Some ideas

Summary of the modelling. Must be content from our second report, but more filtered. Maybe only the best try from each model. And some graphical overview of how good theses "best" candidates are. So that the reader get's an overview. Don't include all the details!!!

Findings from traditional machine learning, findings from deep learning, and transfer learning.

## 3.1 Introduction to modelling

In this project we used a two-step approach to solve this classification task. Initially, we applied machine learning techniques to extract meaningful insights. The models used were "Logistic Regression", "Random Forest", "Support Vector Machines" and "XGBoost". As a transition to deep learning, we also tried out "MLP-Classifer". For the tests with the machine learning models, we prepared the input data in different ways so that we had different data sets that we could try out. As a second step, we tried to solve this classification task with deep learning models. On the one hand, we used a self-build lightweight Convolutional Networks (CNN). On the other hand, we applied transfer learning. Therefore, the following pretrained models from KERAS were used: VGG16, EfficientNet, InceptionV3 and DenseNet121. We tried to apply these models with and without fine-tuning to our data.

## 3.2 Step 1: Machine Learning Techniques

We started with traditional machine learning algorithms. We already had in mind that these models might not be the best choice to solve a classification task with X-ray images. However, it was worth a try to test them because they have a big advantage over deep learning models: They are computationally inexpensive and we can quickly iterate over different versions, features and preprocessing methods. The models we employed included:

- Logistic Regression – A well-known method for classification problems, helping us understand the probability of an event occurring.
- Random Forest – An ensemble method that builds multiple decision trees to improve accuracy and handle complex relationships in the data.
- Support Vector Machines (SVM) – A technique that finds the optimal boundary between different categories, performing well in situations with clear separability.

- XGBoost – A powerful gradient boosting algorithm that refines predictions by learning from errors iteratively.
- MLP Classifier – A simple neural network that captures non-linear relationships but remains limited compared to deeper architectures.

### 3.2.1 Logistic Regression

With the logistic regression model 6 different model runs were performed. On the one hand, the images themselves were used as input data. For performance reasons, however, they had to be significantly reduced in size (only 20x20 pixels). For another run, filtering methods were applied to these X-ray images. In another run, additional masks were added to the images to obscure the areas outside the lungs. This is intended to focus the model on the important areas. For another model run the HOG-technique (Histogram of Oriented Gradients) had been applied to extract structural patterns from the grayscale X-ray images. HOG captures from X-rays

- Edges of lung boundaries, rib cage, and lesions.
- Orientation and distribution of densities i.e. darker/lighter areas correspond to tissue and opacity differences.
- Shape and spread of abnormalities, like how diffuse or sharp an opacity is.

Although HOG is handcrafted, not learned and so it may miss subtle texture or high-level patterns, it is a fast, interpretable, and memory-efficient way to extract features from chest X-rays. Additionally we compressed the HOG features by a PCA (Principal Component Analysis) and finally we tried to optimize the model's hyperparameters.

Using this HOG-extracted features as input data achieved an overall accuracy of 73.7% which is the best accuracy we could reach using a logistic regression model. This model run achieved also the highest scores for the COVID class we could achieve with a logistic regression model (precision: 52%, recall: 56% and f1-score: 54%). This result leaves room for improvement. Logistic regression might not be the best choice for such a high-dimensional feature set.

### 3.2.2 Random Forest

From the various experiments with the logistic regression model, we have already seen that the use of the HOG features as input data were the most promising model run. Therefore, the tests with the random forest model started directly with the HOG-features as input data. With the random forest model 2 different model runs were performed: One with the model's default hyperparameters. And the second run used optimized hyperparameters found by using the GridSearch-technique. This optimization of the hyperparameters brought a slight improvement. We thus achieved an overall accuracy of 76% with the random forest model. Also, the prediction of the COVID class could be slightly improved (precision: 62%, recall: 66% and f1-score: 62%) compared to our best logistic regression model.

### 3.2.3 Support Vector Machines (SVM)

These models provided valuable insights, allowing us to identify correlations and make initial predictions. However, despite their reliability, they faced limitations in capturing highly complex interactions within the data.

## 3.3 Step 2: Transition to Deep Learning

Recognizing the need for improved accuracy, we moved to deep learning—a subset of machine learning that excels in handling vast amounts of data and uncovering intricate patterns without human-defined features.

Deep learning models differ from traditional machine learning by using multiple layers of artificial neurons to process information. The key advantages we observed included:

- Better Feature Extraction – While machine learning relies on manually selected features, deep learning autonomously discovers patterns within the data.
- Improved Generalization – The model became more adaptable, performing well even with new and unseen data.
- Higher Accuracy – By leveraging deeper architectures, we saw a notable boost in predictive performance.

As a result, our deep learning models consistently outperformed traditional machine learning techniques. They identified patterns we hadn't anticipated, providing more refined and actionable insights.

### 3.4 Comparing the Results

To quantify the improvement, we evaluated the performance of both approaches using metrics such as accuracy, precision, recall, and F1-score. Here's a simplified comparison:

## 4 Conclusion and Future Work

Some ideas:

Draw a Conclusion from the previous chapter:

Which is the model we chose? Why do we chose it?

How can this model be used? What are the limitatios?

What has someone to keep in mind, when using the model?

Some notes taken during writing chapter 3 about modelling:

- Using full size images (299 x 299 pixels) and even reduced size (128 x 128 pixels) was not possible with (some?/ all?) machine learning models due to performance reasons.
- Performance issues: used a lot of RAM or even too much RAM so that the system crashed.
- we had to use very small images (20x20 pixel) to get the machine learning models running. A disadvantage is that the greatly reduced pixel number leads to a loss of information in the X-ray images.
- Our tests have shown that machine learning models perform better when using extracted features as input rather than the images themselves.

## 5 Discussion - Reflecting on our project

some ideas:

Reflection on our work.

Interprets the experimental results, identifies strengths and limitations, and discusses potential clinical implications.

Questions we could answer, FROM DS: Report methodology:

Difficulties encountered during the project:

- What was the main scientific obstacle encountered during this project?
- For each of the following points, if you encountered difficulties, detail how they slowed you down in setting up your project.
- Forecast: tasks that took longer than expected, etc.
- Datasets: acquisition, volumetry, processing, aggregation, etc.

- Technical/theoretical skills: timing of skill acquisition, skill not offered in training, etc.
- Relevance: of the approach, model, data, etc.
- IT: storage power, computational power, etc.
- Other

## 6 References

Lists all the sourced literature and references cited throughout the report.

## 7 Appendices

???

Gantt diagram.

Description of code files.

You can find the GitHub repository belonging to this project here: [https://github.com/DataScientest-Studio/mar25-bds\\_analysis-of-covid-19-chest-x-rays](https://github.com/DataScientest-Studio/mar25-bds_analysis-of-covid-19-chest-x-rays)