Open University of Catalonia (UOC) Master's Degree in Data Science

MASTER'S THESIS

Area: 3

Segmentation of areas affected by pneumonia resulting from COVID-19 infection

---

Author: Marta Coll Pol

Tutor: Jordi de la Torre Gallart

Professor: Laia Subirats Maté

Barcelona, January 10, 2024

FINAL PROJECT RECORD

Title of the project:	Segmentation of areas affected by pneumonia resulting from COVID-19 infection
Author's name:	Marta Coll Pol
Collaborating teacher's name:	Jordi de la Torre Gallart
PRA's name:	Laia Subirats Maté
Delivery date (mm/yyyy):	01/2024
Degree or program:	Master's Degree in Data Science
Final Project area:	3
Language of the project:	English
Keywords	COVID-19, Medical Imaging, Object detection

Abstract

While the immediate urgency of the COVID-19 pandemic has evolved, the lessons we have learned during this period remain invaluable. We have seen that sometimes a timely and accurate diagnosis can be the difference between life and death.

Currently, there are two methods to diagnose COVID-19 with high fidelity, the first being the well-known PCR, which stands for polymerase chain reaction and requires a lab analysis, and the second is based on imaging techniques like chest X-rays or computed tomography scans, that provide greater precision[1]. However, PCR results can take several hours and, in some cases, more than a day, whereas radiographs can be obtained in a matter of minutes. For this reason, a tool that provides a visible bounding container around the area suspected of being affected by COVID-19 could really help speed up an accurate diagnosis, especially for non-radiologist medical personnel.

This project, inspired by the Kaggle challenge "SIIM-FISABIO-RSNA COVID-19 Detection"[2], focuses on the development of a machine learning model capable of identifying and localizing COVID-19 abnormalities in chest radiographs, more specifically, the model will categorize radiographs as negative for pneumonia or as having a typical, indeterminate or atypical appearance for pneumonia, the latter of which is associated with COVID-19, and provide a visual bounding box around the detected appearance. The main goal is to provide medical professionals with a quick and safe diagnostic tool, which could be expanded later to include the diagnosis of other lung conditions. In addition, doctors will be able to make better treatment decisions, as they will have a visual proof of the extent of the disease, which could prevent the worst possible outcomes.

Keywords: COVID-19, medical imaging, object detection, machine learning

[1]Mirsadraee S, Pourabrollah Toutkaboni M, Bakhshayeshkaram M, Rezaei M, Askari E, Haseli S, Sadraee N. Radiological and Laboratory Findings of Patients with COVID-19 Infection at the Time of Admission.[Publication date: 21/10/2020][Access date: 10/10/2023] Available at: https://ijp.iranpath.org/article_240036.html

[2]Andrew Kemp, Anna Zawacki, Chris Carr, George Shih, John Mongan, Julia Elliott, Kaiwen, ParasLakhani, Phil Culliton. SIIM-FISABIO-RSNA COVID-19 Detection [Publication date: 2021][Access date: 10/10/2023] Available at: https://kaggle.com/competitions/siim-covid19-detection

Resumen

Aunque ya hemos dejado atrás la urgencia inmediata de la pandemia de COVID-19, las lecciones que hemos aprendido durante este periodo siguen siendo inestimables. Hemos visto que a veces un diagnóstico oportuno y preciso puede ser la diferencia entre la vida y la muerte. Actualmente, existen dos métodos para diagnosticar COVID-19 con alta fidelidad: los análisis en laboratorio de la conocida prueba PCR, y el segundo se basa en técnicas de imágen como las radiografías de tórax o las tomografías computarizadas, que aportan mayor precisión[3]. Sin embargo, los resultados de la PCR pueden tardar varias horas, mientras que las radiografías pueden obtenerse en cuestión de minutos. Por esta razón, una herramienta que proporcione una imagen delimitada alrededor de la zona sospechosa de estar afectada por COVID-19 podría ayudar realmente a acelerar un diagnóstico preciso, especialmente para el personal médico no radiólogo.

Este proyecto, inspirado en el reto de Kaggle "SIIM-FISABIO-RSNA COVID-19 Detection"[4], se centra en el desarrollo de un modelo de aprendizaje automático capaz de identificar y localizar anomalías de COVID-19 en radiografías de tórax. El modelo categorizará las radiografías como negativas para neumonía o de aparencia típica, indeterminada o atípica de neumonía, siendo esta última asociada a COVID-19, y proporcionará un cuadro delimitador visual alrededor de la enfermedad. El objetivo principal es proporcionar a los profesionales médicos una herramienta de diagnóstico rápida y segura. Además, los médicos podrán tomar mejores decisiones de tratamiento, ya que dispondrán de una prueba visual del alcance de la enfermedad, lo que ayudará a prevenir los peores desenlaces.

Keywords: COVID-19, imágenes médicas, detección de objetos, aprendizaje automático

[3]Mirsadraee S, Pourabrollah Toutkaboni M, Bakhshayeshkaram M, Rezaei M, Askari E, Haseli S, Sadraee N. Radiological and Laboratory Findings of Patients with COVID-19 Infection at the Time of Admission.[Publication date: 21/10/2020][Access date: 10/10/2023] Available at: https://ijp.iranpath.org/article_240036.html

[4]Andrew Kemp, Anna Zawacki, Chris Carr, George Shih, John Mongan, Julia Elliott, Kaiwen, ParasLakhani, Phil Culliton. SIIM-FISABIO-RSNA COVID-19 Detection [Publication date: 2021][Access date: 10/10/2023] Available at: https://kaggle.com/competitions/siim-covid19-detection

Contents

Abstract iii

Resumen v

Table of Contents vii

List of Figures xi

List of Tables 1

Copyright and License 3

1 Project Introduction 5

1.1 Proposal . . . 5
1.2 Relevance of the Proposal . . . 5
1.3 Personal Motivation . . . 6
1.4 Hypothesis . . . 7
1.5 Objectives . . . 7
1.6 Project Methodology . . . 9
1.7 Plan . . . 10

2 Literature Review 13

2.1 Introduction . . . 13
2.2 Medical Image Analysis for COVID-19 Detection . . . 13
2.3 Intersection of Medical Imaging and Deep Learning for COVID-19 Detection . . . 15

3 Dataset 17

3.1 SIIM-FISABIO-RSNA COVID-19 Detection Dataset . . . 17
3.1.1 Dataset Description . . . 19
3.2 Exploratory Data Analysis . . . 19

3.2.1 Missing Values . . . 20
3.2.2 Study-level Labels . . . 20
3.2.3 Image-level Labels . . . 20
3.2.4 DICOM Metadata . . . 23
3.2.5 Conclusions . . . 27

4 Dataset Pre-processing **31**
4.1 Photometric Interpretation Differences . . . 31
4.2 Data Clean-up . . . 32
4.3 Handling of Class Imbalances . . . 32
4.3.1 Data Augmentation . . . 32
4.3.2 Excluding Data . . . 35
4.3.3 Class Balancing Strategy . . . 35
4.4 Resizing . . . 36

5 Object Detection and Classification **37**
5.1 Introduction . . . 37
5.2 Justification of Development Approaches . . . 38
5.3 Yolo Framework . . . 39
5.3.1 YOLOv5 Loss Function . . . 40
5.4 Model Selection . . . 42
5.5 Evaluation Metrics . . . 43
5.5.1 Confusion Matrix . . . 43
5.5.2 Precision (P) . . . 43
5.5.3 Recall (R) . . . 44
5.5.4 Recall-Confidence Curve . . . 44
5.5.5 Precision-Recall (PR) Curve . . . 44
5.5.6 Mean Average Precision (mAP) . . . 44
5.5.7 F1 Score Curve . . . 45

6 Experiments **47**
6.1 Experiment Overview . . . 47
6.1.1 Experimental Setup . . . 47
6.1.2 Key Experiments . . . 48
6.2 Results . . . 49
6.2.1 No Data Augmentation Experiment . . . 49
6.2.2 Background Class Experiment . . . 49

6.2.3 Individual Exploration of YOLOv5 Model Architectures 51
6.2.4 Ensemble Learning 54
6.3 Conclusions 56

7 Future Work Areas **63**

Bibliography **63**

List of Figures

1.1 Gantt diagram from October to mid November. 10
1.2 Gantt diagram from mid November to January. 11

3.1 Examples of DICOM images from SIIM-COVID19-Dataset. 18
3.2 Example of metadata found in DICOM images from SIIM-COVID19-Dataset. . 18
3.3 Distribution of target classes distribution from SIIM-COVID19-Dataset. 20
3.4 Examples of DICOM images for the Negative for Pneumonia case. 21
3.5 Examples of DICOM images for the Typical appearance case. 21
3.6 Examples of DICOM images for the Indeterminate appearance case. 22
3.7 Examples of DICOM images for the Atypical appearance case. 22
3.8 Stacked Bar Chart of counts by study-level classes and the presence of image bounding boxes labels. 23
3.9 *De-identification Method*, *Patient's Sex*, *Modality* and *Photometric Interpretation* distributions. 24
3.10 DICOM images examples for different *Patient sex* values. 25
3.11 DICOM images examples for different *Modality* values. 26
3.12 *Body Part Examined* Distribution. 26
3.13 DICOM images examples for different *Body Part Examined* values. 27
3.14 *Private Creator* Distribution. 28
3.15 DICOM images examples for different *Private Creator* values. 28

4.1 Data augmentation techniques. 33
4.2 Comparison after applying an horizontal flip, blur and noise injection. 34
4.3 Comparison after applying a translation, blur and noise injection. 34
4.4 Comparison after applying cropping, blur and noise injection. 35
4.5 Class distribution after data clean-up. 36
4.6 Class distribution after applying the class balancing strategy. 36

5.1 YOLOv5 simplified architecture[1]. 40

6.1 YOLOv5m's confusion matrix resulting from the experiment where data augmentation techniques are not applied in the dataset. The Experiment has been conducted using YOLOv5m with 100 epochs, and has been evaluated on Validation set. . . . 50
6.2 YOLOv5m's confusion matrix evaluated on Validation set for the experiment where all Negative for Pneumonia images where included in the training of the model as background images. . . . 51
6.3 YOLOv5m's confusion matrix evaluated on Validation set for the experiment where 10% of Negative for Pneumonia images where included in the training of the model as background images. . . . 52
6.4 Training dynamics of YOLOv5s trained with 500 epoch. . . . 53
6.5 Training dynamics of YOLOv5m trained with 500 epoch. . . . 53
6.6 Training dynamics of YOLOv5l trained with 500 epoch. . . . 54
6.7 YOLOv5m's confusion matrix for the Test set. . . . 57
6.8 YOLOv5m's F1-Confidence curve for the Test set. . . . 57
6.9 YOLOv5m's Recall-Confidence for the Test set. . . . 58
6.10 YOLOv5m's Precision-Recall curve for the Test set. . . . 58
6.11 YOLOv5m's ground truth for batch 0 of the Test set. . . . 59
6.12 YOLOv5m's predictions for batch 0 of the Test set. . . . 59
6.13 Ensemble learning confusion matrix for the Test set. . . . 60
6.14 Ensemble learning F1-Confidence curve for the Test set. . . . 60
6.15 Ensemble learning Recall-Confidence for the Test set. . . . 61
6.16 Ensemble learning Precision-Recall curve for the Test set. . . . 61
6.17 Ensemble learning ground truth for batch 0 of the Test set. . . . 62
6.18 Ensemble learning predictions for batch 0 of the Test set. . . . 62

List of Tables

5.1 Summary of YOLOv5 architecture components[2]. 41
5.2 Summary of YOLOv5 Model Variants[2]. 42

6.1 Baseline model resulting from the experiment where data augmentation techniques are not applied in the dataset. The Experiment has been conducted using YOLOv5m with 100 epochs, and has been evaluated on Validation set. . . 49
6.2 Background Class Experiment using YOLOv5m with 300 epochs, and different amount of background images, evaluated on Validation set. 50
6.3 Performance metrics for experiments with YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x evaluated on the Validation set. 55
6.4 Performance metrics for experiments with YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x evaluated on the Test set. 56
6.5 Ensemble Learning performance metrics on Test and Validation Sets 56

Copyright and License

Chapter 1

Project Introduction

1.1 Proposal

The proposal of this thesis is based on the Kaggle challenge "SIIM-FISABIO-RSNA COVID-19 Detection"[3]. This challenge was born in the context of the COVID-19 pandemic, where there was a need for innovative solutions that could speed up the accurate diagnosis of the virus. Specifically, the challenge encourages to provide solutions by applying machine learning techniques to chest X-rays (CRX) in order to detect and locate the presence of COVID-19 infection. To this end, a dataset of anonymised chest X-rays, previously annotated by expert radiologists, is provided by the institutions sponsoring the challenge. The aim of the proposed Master's thesis is to provide a data science-based solution for the automated identification and localisation of anomalies in CRX that may be a direct consequence of COVID-19 disease.

As shown in the study "Radiological and Laboratory Findings of Patients with COVID-19 Infection at the Time of Admission"[4], medical imaging techniques, in particular chest X-rays (CRX) and CT scans, are used not only to verify the presence of COVID-19 infection when it is suspected and the well-known PCR test is negative, but also to monitor the progression of the infection. While CT scans are more accurate, they are expensive and time-consuming. CRX, on the other hand, can be obtained in minutes and is also faster than the time required to obtain PCR results, which can be hours or even days in the worst-case scenario because of the need for laboratory analysis. It can be concluded that using CRX for rapid diagnosis of COVID-19 and visualising the spread of infection is an option worth investigating.

1.2 Relevance of the Proposal

The COVID-19 pandemic has had an unprecedented global impact. It has paralysed people's daily lives and overwhelmed healthcare facilities with the challenges of managing the high

volume of patients. In order to effectively control the spread of the disease and improve the management of the resources available, and therefore the impact on patients' lives, rapid and accurate detection of the disease is essential. The limitations of PCR in terms of speed in obtaining results and its accuracy of only 70%[4], make medical imaging a complementary diagnostic tool of high value.

The following social issues are addressed by the development of this project:

- **Diagnostic capabilities improvement:** Contributes to the improvement of the diagnostic capabilities of the health care system, which can have a direct impact on the services it can provide.

- **Improvement of health resource management efficiency:** By ensuring that patients who need urgent attention receive the care they need, rapid and accurate diagnosis helps to manage resources more efficiently.

- **Reducing transmission:** COVID-19 is characterised by being a disease with a high rate of transmission, so early detection has a dramatic impact on reducing the extent of the disease.

The following contributions are highlighted in terms of scientific relevance:

- **Advances in medical imaging:** The objective is to create a project that can make a meaningful contribution to the field of disease detection using medical imaging.

- **Real-world impact:** A successful project can provide real-world solutions that help make automated tools for detecting disease a new standard in the medical field, leading to an improvement on patient care.

1.3 Personal Motivation

My personal motivation is deeply rooted in my desire to contribute to the greater good of society. My ambition is to pursue a career in the pharmaceutical sector, with a particular focus on the application of data science in the research of new treatments, the development of tools for the early detection of disease, or the creation of solutions that directly or indirectly benefit the healthcare system and people's lives. What particularly excites me is working with images as data, a passion I discovered during my bachelor's degree in Audiovisual Systems Engineering, when I collaborated with the image processing group at the UPC (Universitat Politècnica de Catalunya).

Working with images not only appeals to my technical skills, but also fits perfectly with my goal of making a positive impact on healthcare and people's well-being. Images contain a wealth of information that can be used to advance medical diagnostics, discover new therapeutic interventions and ultimately improve patient outcomes. I am keen to apply my knowledge and skills to projects that have a tangible and lasting impact on society, and I believe that the convergence of data science and healthcare is where I can make a meaningful contribution.

1.4 Hypothesis

Machine learning models, trained on a dataset of chest X-ray images, can be used to detect abnormalities associated with COVID-19 while at the same time effectively locating and highlighting these regions of interest within the images. This approach aims to provide a valuable automated diagnostic tool to support accurate and rapid decision-making by medical practitioners.

1.5 Objectives

For a correct assessment of the project development, the intermediate objectives and their associated tasks are defined below, sorted according to the expected evolution of the project.

1. State of the art research:
    - Researching the state of the art in detecting disease using medical imaging. It will be of particular interest to see what kind of implementations are used and what conclusions are drawn. The aim is to set a benchmark to beat and learn from the successes or failures of similar projects.
    - Researching the state of the art in detecting detection regions of interest (ROI) within an image.
    - Draw conclusions from the investigation.
2. Prepare the development environment setup:
    - Given the high level of computing resources expected, research how to properly set up an environment.
    - Set the environment where code will be executed.
    - Set a code version control mechanism.

3. Design the coding project:
    - Design the project folder structure and script organization.
    - Decide the algorithms and techniques to explore for Multi-Label Classification on images.
    - Define the data preparation steps needed for the chosen algorithms.
    - Decide evaluation metrics to implement.
    - Decide the region of interest (ROI) detection algorithms to implement.
    - Decide an strategy to evaluate ROI detection algorithms results.
4. Exploratory data analysis (EDA):
    - Get familiar with the contents of the dataset.
    - Perform an EDA to obtain a better understanding of the data.
5. Data pre-processing:
    - Code the necessary functions for data pre-processing.
    - Apply data pre-processing to data.
6. Multi-label classification model implementation:
    - Define a baseline model.
    - Implement evaluation metrics.
    - Implement different model approaches.
    - Train, test and compare models results.
    - Select, fine-tune and optimize the best approach.
7. Development of Region of Interest (ROI) detection algorithms:
    - Implement ROI detection algorithms.
    - Implement evaluation metrics for ROI detection.
    - Train, test and compare different approaches to choose the best one.
8. Evaluation of the results:
    - Evaluate the metrics results of the different approaches used.
    - Perform a visual evaluation of the results.

- Draw conclusions of the evaluations.

9. Document the thesis:

   - Document the different sections of the thesis.
   - Review and improve the document.

1.6 Project Methodology

This project, inspired by the SIIM COVID-19 Detection Challenge[3], starts with a thorough examination of the problem statement and the objectives of the competition. As the competition dataset is readily available, the primary focus will be on understanding the unique characteristics of the dataset. The initial phase will involve extensive data exploration to gain insight into the structure, distribution and labelling scheme for COVID-19 detection. This exploratory analysis will lay the groundwork for the data pre-processing and model selection that will follow.

Python will be used as the primary programming language. Python libraries such as NumPy, Pandas and OpenCV will be used for data manipulation, image pre-processing and statistical analysis[5]. Popular frameworks such as TensorFlow or PyTorch will be used for model development, hyperparameter tuning and model interpretability for machine learning and deep learning tasks[6].

Given the readiness of the dataset for this specific task, data pre-processing steps will include tasks such as image resizing to a consistent format, pixel value normalisation and missing data handling. Data augmentation techniques will also be used, if necessary, to increase the diversity of the training dataset, possibly using libraries such as Keras for image augmentation[7].

A variety of deep learning architectures will be explored for model development. The special nature of medical image data will be taken into account. The focus will be on model fine-tuning and optimisation for maximisation of performance. Evaluation metrics for the specific task will be used to provide a rigorous assessment of the performance and robustness of the models.

Jupyter Notebook will be used for code development, experimentation and visualisation. Visualisation libraries such as Matplotlib and Seaborn will be used for the effective presentation of findings and results[8].

A free cloud environment with sufficient resources will be considered for project development, given the computational demands of deep learning tasks. This cloud environment could ensure scalability and availability of GPUs for accelerated model training. The Kaggle platform offers free access to GPU resources, with limits ranging from 20 to 30 hours per week depending on the specific GPU. This makes it a convenient and easily accessible option.

During the project development, detailed records of experiments, hyperparameters, and results will be meticulously maintained, ensuring transparency and reproducibility. Modern tools such as Weights & Biases (wandb)[9] and ClearML[10] provide valuable support for experiment tracking and management. These platforms enhance collaboration, facilitate result interpretation, and contribute to the overall efficiency of the development process. Search engines such as Google Scholar will be used for literature searches and to be up to date with the latest developments in the field. This comprehensive methodology in combination with the use of Python and relevant libraries will facilitate the development of an effective solution for COVID-19 detection in medical images.

1.7 Plan

The thesis development plan has been defined using a Gantt diagram.

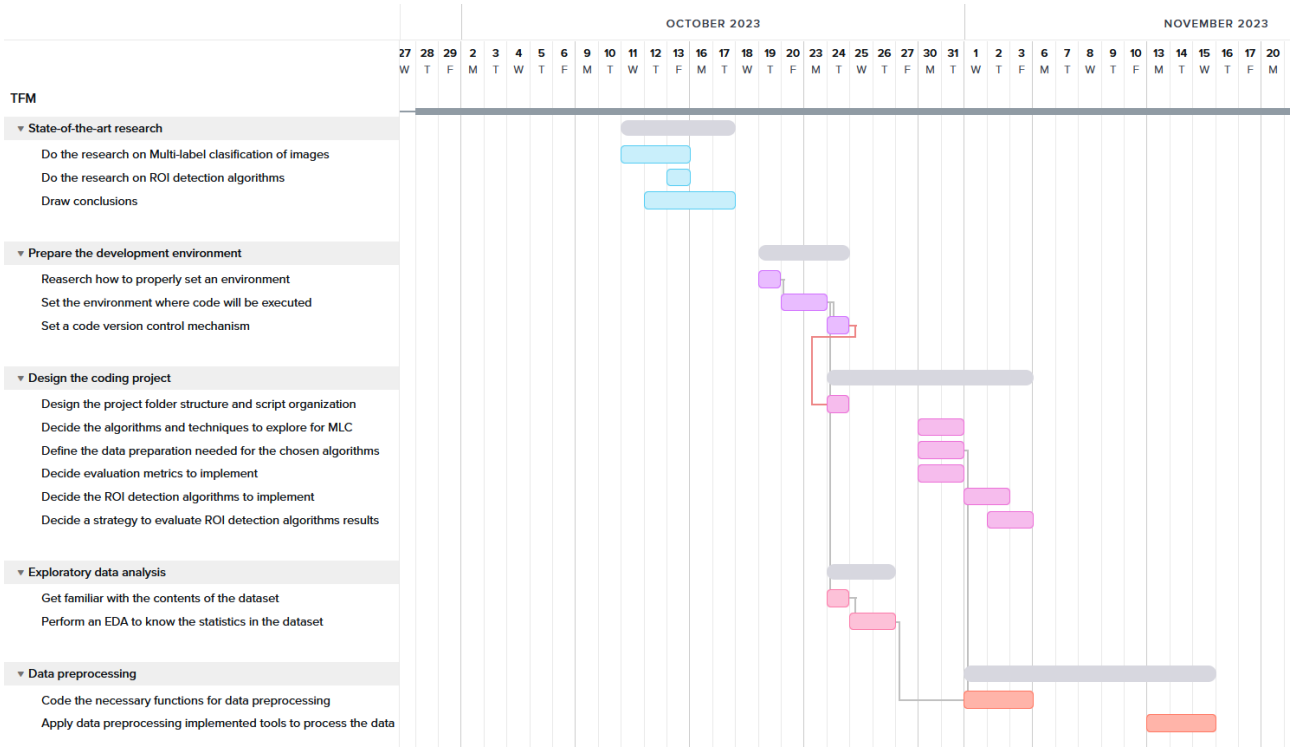

Figure 1.1: Gantt diagram from October to mid November.

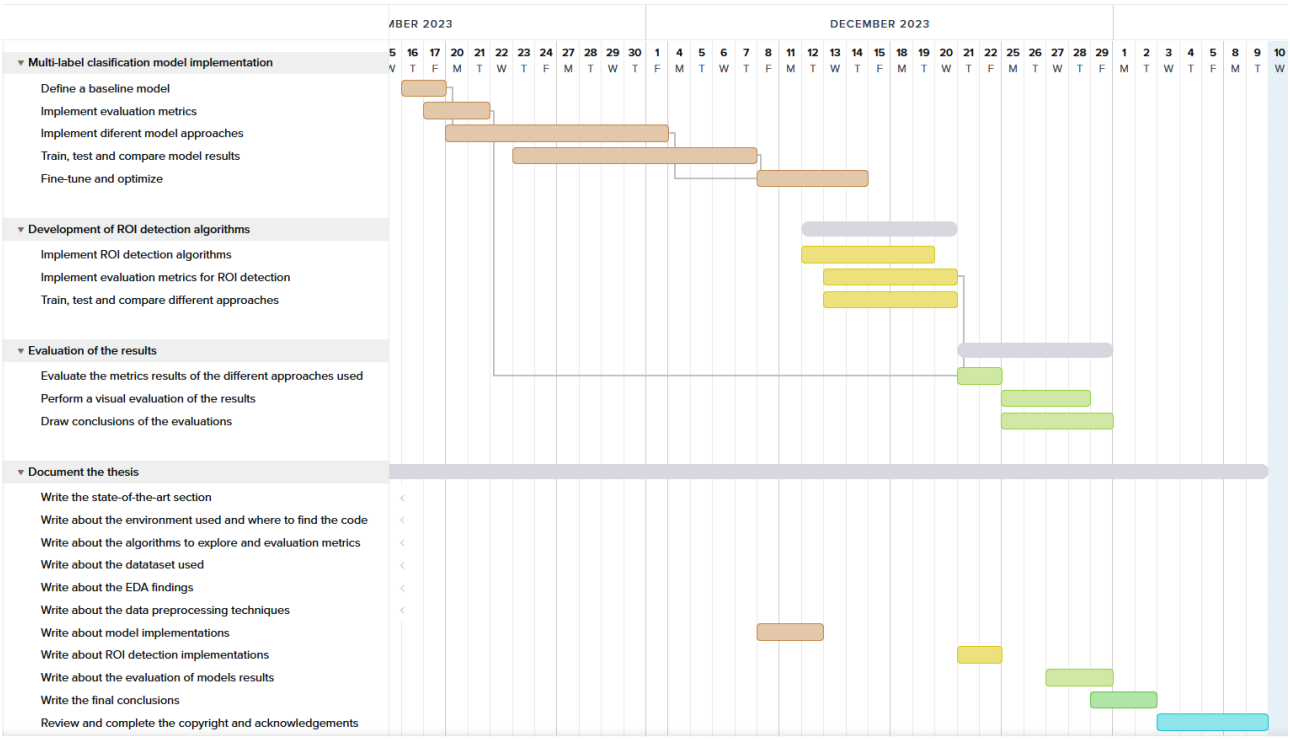

Figure 1.2: Gantt diagram from mid November to January.

Chapter 2

Literature Review

2.1 Introduction

In late 2019, a new coronavirus called SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), or COVID-19, was discovered in a group of patients from Wuhan, China [11]. Due to its highly infectious nature, it spread rapidly to the rest of the world, collapsing healthcare on a global scale. The virus infected mainly the lungs and caused fever and respiratory symptoms. During the early stages of the pandemic, a significant number of patients developed pneumonia, from which a few developed acute respiratory distress syndrome and, of these, some soon worsened and died of multi-organ failure[12].
Reverse transcription polymerase chain reaction (RT-PCR) became the standard for diagnosis of COVID-19. However, this serological test had a high number of false-negative results, especially in the early stages of infection, which delayed accurate diagnosis and potentially contributed to the spread of the virus, in addition to the inconvenient delay from sample collection to laboratory results. For these reasons, medical imaging often was an adjunct test in the diagnosis of patients suspected of having COVID-19 infection and in monitoring the course of the disease[13][14].

This literature review will outline the main areas of research in medical image analysis related to COVID 19 and explore the pivotal studies, methodologies, and trends at the intersection of medical imaging and deep learning for COVID-19 detection.

2.2 Medical Image Analysis for COVID-19 Detection

In the fight against COVID-19, various imaging techniques were used, each of which is important in its own context.

Magnetic resonance imaging (MRI) is an imaging technique that allows soft tissues to be visualised in great detail. Although MRI is not suitable for detecting COVID-19 due to the long scan time and high cost of its use, it is useful for assessing the effects of COVID-19 in children and pregnant women as it does not contain radiation. Due to the excellent visualisation of structural and functional information of different soft tissues, MRI could be used to study the vulnerability of different organs for a better understanding of the effects of COVID-19 infection[13].

Chest computed tomography (CT) played an important role not only in the detection but also in the follow-up of COVID-19 infection, providing detailed 3D images of the lungs, allowing rapid and accurate diagnosis even in the early days of infection when RT-PCR is more likely to produce false-negative results[15]. Characteristic signs of COVID-19, such as ground-glass opacities, multi-focal patchy consolidation and/or interstitial changes with peripheral distribution, are easily observed on CT. However, it is important to note that CT scans use ionising radiation, which can damage DNA with prolonged exposure, increasing the risk of cancer and causing direct tissue damage. Therefore, diagnosis and follow-up cannot be based on CT scans alone[16].

Chest radiography (CRX) is a widely used diagnostic tool that is often used in the initial assessment of people suspected of having COVID-19. Chest X-rays provide a two-dimensional view of the chest, allowing experts to examine the lungs and surrounding structures for signs of abnormalities, such as the presence of areas where lung tissue is denser, often due to fluid accumulation, or ground-glass opacities, which often indicate areas of lung inflammation[17]. Although CRX has limitations in the detection of subtle lung changes, in contrast to detailed CT scans, CRX radiation levels have a relatively low impact on patient safety, are easy to use and are widely available in healthcare facilities, making them a practical option for quick preliminary assessments[14].

Lung ultrasound (LSU) has emerged as a radiation-free imaging modality for the diagnosis of COVID-19. LSU uses ultrasound waves to produce real-time images of the lungs, providing insight into indicators such as B-lines, pleural thickening and pleural effusion, all of which are associated with COVID-19 pneumonia[18]. This approach is non-invasive, cost-effective and can be performed conveniently at the patient's bedside, offering significant advantages, particularly where minimising radiation exposure is a priority. However, it's important to note that the availability of both trained personnel and the necessary equipment for LSU can vary from one healthcare facility to another, which may affect its widespread adoption[19].

2.3 Intersection of Medical Imaging and Deep Learning for COVID-19 Detection

The intersection of medical imaging and artificial intelligence (AI), emerged as a potent approach in the fight against the COVID-19 pandemic, offering the potential for rapid and accurate detection of the disease through the application of deep learning models.

A deep learning model consists of interconnected layers, which serve as computational building blocks. These layers operate collectively to emulate the cognitive processes observed in human thought. For the particular case of solving image-related tasks, a type of deep learning methodology known as Convolutional Neural Network (CNN) is commonly used. CNNs leverage the potential of convolutional operations, a mathematical technique, to deconstruct complex images into smaller, manageable parts. This segmentation process allows the network to analyze these parts and extract patterns and features that can be used to recognize the contents of the image[20].
The ability of these algorithms to automatically extract relevant features from images has proven to be fundamental in the processing of CRX, CT scans and LSU for the identification of specific COVID-19 patterns. Studies such as "A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis"[21] have demonstrated the effectiveness of deep learning models in distinguishing COVID-19 cases from other respiratory diseases, facilitating early intervention and containment.
In the mentioned study by Wang et al. (2020)[21], the system presented comprises two distinct deep learning models built upon Convolutional Neural Networks (CNNs). The initial model, called DenseNet121-FPN, takes a CT scan as input and is designed to identify and extract the specific region of interest (ROI) within the lung, suspected of being impacted by COVID-19. Subsequently, a process is employed to eliminate non-lung tissues or organs from the extracted ROI, after which the refined data is forwarded to the second model, called COVID-19Net, for a comprehensive prognostic and diagnostic analysis. The system achieved an accuracy of approximately 80% on two validation sets for the diagnostic task. In addition, the framework produced heat maps indicating suspected areas affected by COVID-19, on top of CT scans, contributing to the trend of showcasing the potential of AI-driven solutions not only in diagnosis but also in guiding treatment and resource allocation decisions.

In the study called "Detection and analysis of COVID-19 in medical images using deep learning techniques" by Yang et al. (2021)[22], the authors explore the effectiveness of deep learning techniques for the detection of COVID-19 in medical images, specifically X-rays and CT scans.

First of all, four well-established pre-trained CNN models, VGG16, DenseNet121, ResNet50, and ResNet152, are explored for the task of classifying COVID-19 in CT-scan images. To optimize the models, they introduce the Fast.AI ResNet framework, automating the selection of architecture, pre-processing steps, and training parameters. They also tackle the challenge of limited data by applying transfer learning techniques to enhance training efficiency. Remarkably, their approach achieves both high accuracy and F1-scores, exceeding 96% in diagnosing COVID-19. Secondly, for detecting COVID-19/pneumonia from X-ray images, an enhanced VGG16 model is presented, which impressively attains a 99% accuracy.

The approach by Gaur et al. (2023) called "Medical image-based detection of COVID-19 using Deep Convolution Neural Networks"[23], presents a practical solution that employs Deep Convolutional Neural Networks (CNN) to detect COVID-19 from chest X-rays and differentiate between normal cases and those impacted by Viral Pneumonia. Three pre-trained CNN models, EfficientNetB0, VGG16, and InceptionV3, are assessed using transfer learning due to their balance of accuracy and efficiency with fewer parameters, making them suitable for mobile applications. From the models explored, the EfficientNetB0 model reported the highest accuracy of 94.79% in detecting and classifying COVID-19 chest X-rays from other categories of chest abnormalities and an overall accuracy of 92.93%.

In summary, combining medical imaging with deep learning is a good approach to detect, diagnose and predict COVID-19 more quickly, accurately and affordably. The studies that have been analysed show significant advances in the use of deep learning in X-rays and CT scans, which can help healthcare professionals make better decisions in resource and patient management. As this field of research advances, it is becoming increasingly clear that the intersection between medical imaging and artificial intelligence will become a crucial part of modern healthcare for the detection and treatment of infectious diseases.

Chapter 3

Dataset

3.1 SIIM-FISABIO-RSNA COVID-19 Detection Dataset

The SIIM-FISABIO-RSNA COVID-19 Detection dataset is a curated collection designed for a Kaggle competition hosted by FISABIO, The Foundation for the Promotion of Health and Biomedical Research of Valencia Region and the Radiological Society of North America (RSNA). The challenge is to detect and localise COVID-19 abnormalities in chest radiographs, a problem that involves both object detection and classification tasks. The dataset provides a set of studies, and each study contains a variable set of chest radiographs for a single patient. In addition, study-level and image-level metadata files are provided for the training set. The data from this dataset originates from BIMCV-COVID19[24] and MIDRC-RICORD[25], each with its respective usage agreements[26][27], both respected during the development of this project, and distributed under the CC BY-NC 4.0 DEED license[28], that allows sharing and adapting of the material. Material has been adapted to fulfill the development needs of the project. Data has been re-annotated for this competition as described in the paper "The 2021 SIIM-FISABIO-RSNA Machine Learning COVID-19 Challenge: Annotation and Standard Exam Classification of COVID-19 Chest Radiographs"[29].

Participants in the associated competition are required to predict bounding boxes and corresponding classes for image-level findings and a class for study-level findings.

The chest x-rays consist of DICOM images, similar to the example images from Figure 3.1. DICOM is a format used for medical imaging that includes metadata to give context to the image. From information about how the image was taken to information about the patient. An example of the given metadata in a DICOM image, can be seen in Figure 3.2. It can be observed that de-identification methods have been applied to remove or anonymize patient-specific information, such as names, dates of birth, and other identifiers, while retaining the necessary clinical information for research or analysis purposes. These methods help balance

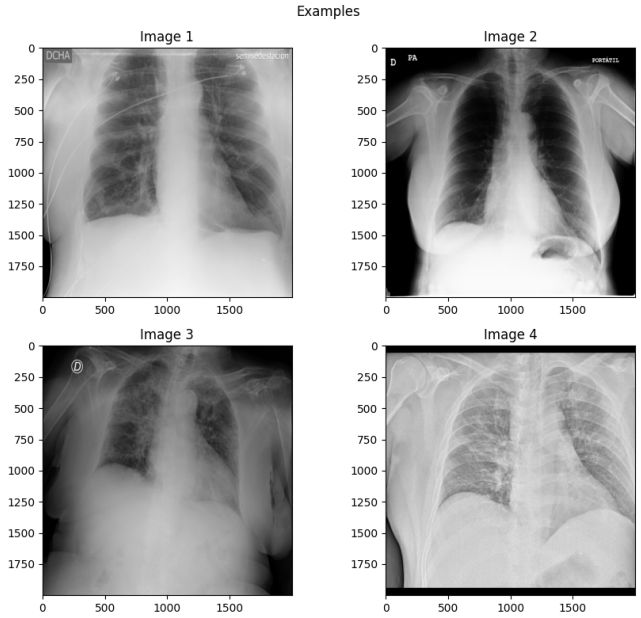

Figure 3.1: Examples of DICOM images from SIIM-COVID19-Dataset.


```
{'Specific Character Set': 'ISO_IR 100',
 'Image Type': "['ORIGINAL', 'PRIMARY']",
 'SOP Class UID': "'03a65300fa41'",
 'SOP Instance UID': "'000c3a3f293f'",
 'Study Date': 'd09eda152722',
 'Study Time': '543adb46f494',
 'Accession Number': '1c2708371bc6',
 'Modality': 'CR',
 "Patient's Name": "'ef8c31f8dfdd'",
 'Patient ID': 'f09ff9b7dab3',
 "Patient's Sex": 'M',
 'De-identification Method': 'CTP Default:  based on DICOM PS3.15 AnnexE. Details in 0012,0064',
 'De-identification Method Code Sequence': '<Sequence, length 6>',
 'Body Part Examined': 'CHEST',
 'Imager Pixel Spacing': '[0.15, 0.15]',
 'Study Instance UID': "'ff0879eb20ed'",
 'Series Instance UID': "'d8a644cc4f93'",
 'Study ID': '55625fb42f3f',
 'Series Number': 1,
 'Instance Number': 1,
 'Samples per Pixel': 1,
 'Photometric Interpretation': 'MONOCHROME2',
 'Rows': 2320,
 'Columns': 2832,
 'Bits Allocated': 8,
 'Bits Stored': 8,
 'High Bit': 7}
```


Figure 3.2: Example of metadata found in DICOM images from SIIM-COVID19-Dataset.

the need for sharing and using data for research or analysis while protecting individuals' privacy and complying with legal and ethical standards.

3.1.1 Dataset Description

3.1.1.1 Training Dataset:

The training dataset comprises 6054 studies with a total of 6334 chest scans in DICOM format, meticulously de-identified to protect patient privacy. Expert radiologists carefully labelled each image, indicating the presence of opacities, defining a bounding box around them, and assigning a class to the study based on overall appearances observed in chest x-rays. Organised in a hierarchical structure with paths following the *study/series/image* format, the study ID directly correlates with study-level predictions, while the image ID facilitates image-level predictions.

The study-level metadata provided for the training set consists of a link between the study ID and the class assigned to the study by the experts. There are four different classes:

- **Negative for Pneumonia:** There are no signs of pneumonia on the study.
- **Typical Appearance:** The overall appearance is consistent with typical signs of pneumonia.
- **Indeterminate Appearance:** The overall appearance is indeterminate.
- **Atypical Appearance:** The overall appearance is atypical. This appearance is associated with COVID-19 infection.

The metadata provided at the image level for the training set contains one row for each bounding box that delimits an opacity in an image, which is expressed using the format *xmin, ymin, xmax, ymax*. The image is referenced by the image ID and the study ID to which the image belongs.

3.1.1.2 Hidden Test Dataset:

The hidden test set reflects the size of the training set, maintaining a balance of complexity and diversity. Labels and bounding boxes are not facilitated for the test set, as predictions on this set are intended to be evaluated using the competition submission platform. An example of a submission file is provided, which contains all image and study-level IDs for the test set.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the SIIM-FISABIO-RSNA COVID-19 Detection dataset. This process involves examining and summarizing key characteristics, patterns, and distributions within the dataset. The development of the EDA has been performed in the Kaggle notebook titled "EDA SIIM-COVID19-DETECTION"[30].

3.2.1 Missing Values

The data provided has been analyzed and there are no missing values on relevant columns.

3.2.2 Study-level Labels

As outlined in the dataset description, four distinct classes are identified at the study level: **Negative for Pneumonia, Typical Appearance, Indeterminate Appearance,** and **Atypical Appearance**. The distribution of these classes, illustrated in Figure 3.3, reveals an inherent imbalance within the dataset. The category *Typical Appearance* emerges as significantly more prevalent, comprising a total of 3007 observations. In contrast, the *Atypical Appearance* category represents the minority with only 483 observations. Occupying the middle ground are *Negative for Pneumonia* and *Indeterminate Appearance*, with 1736 and 1108 observations, respectively. This distribution underscores the imperative to employ balancing techniques to rectify the inherent class imbalance in the dataset.

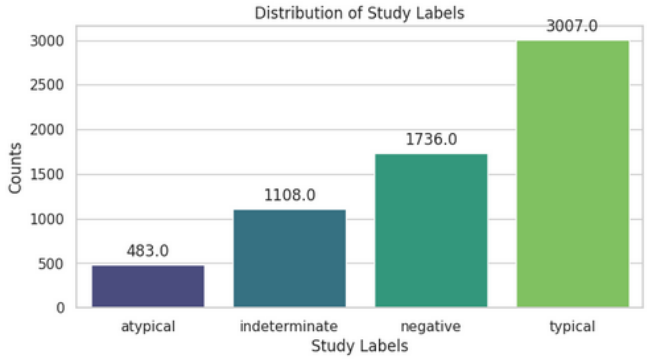

Figure 3.3: Distribution of target classes distribution from SIIM-COVID19-Dataset.

The different classes have been visually examined and representative examples are shown in Figures 3.4, 3.5, 3.6 and 3.7. From the visual examination it can be concluded that there are no discernible differences in the presentation format of the radiographs across these classes. Regarding the presence of opacities, the class *Negative for pneumonia* shows an absence of opacities as expected.

3.2.3 Image-level Labels

As previously explained, image level predictions, essentially consist on looking for signs of opacities in chest X-rays and predict a bounding box around them. To better grasp how often there are bounding boxes in x-rays from the dataset based on study-level class, take a look at the stacked bar chart in Figure 3.8. Upon visually inspecting different classes earlier, it had

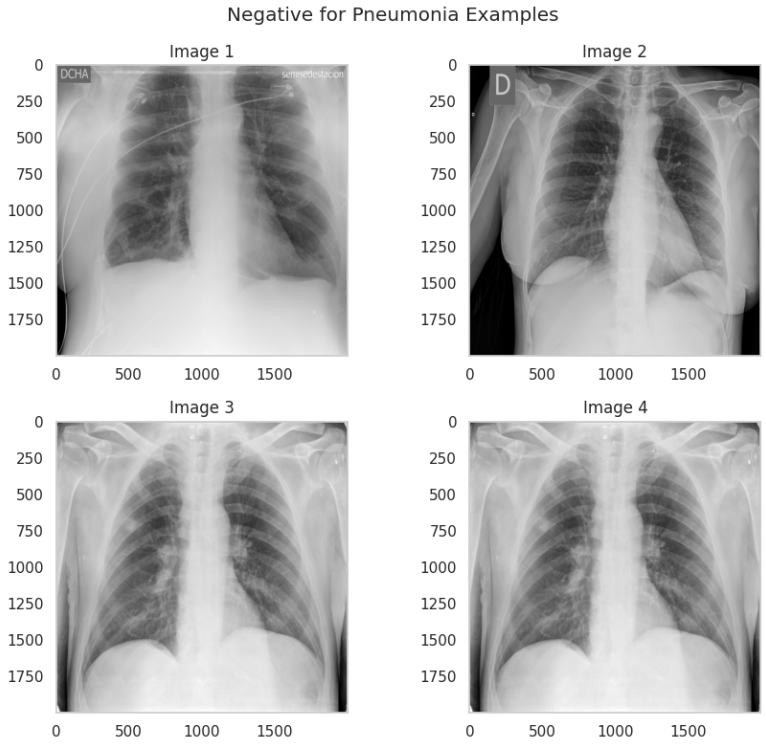

Figure 3.4: Examples of DICOM images for the Negative for Pneumonia case.

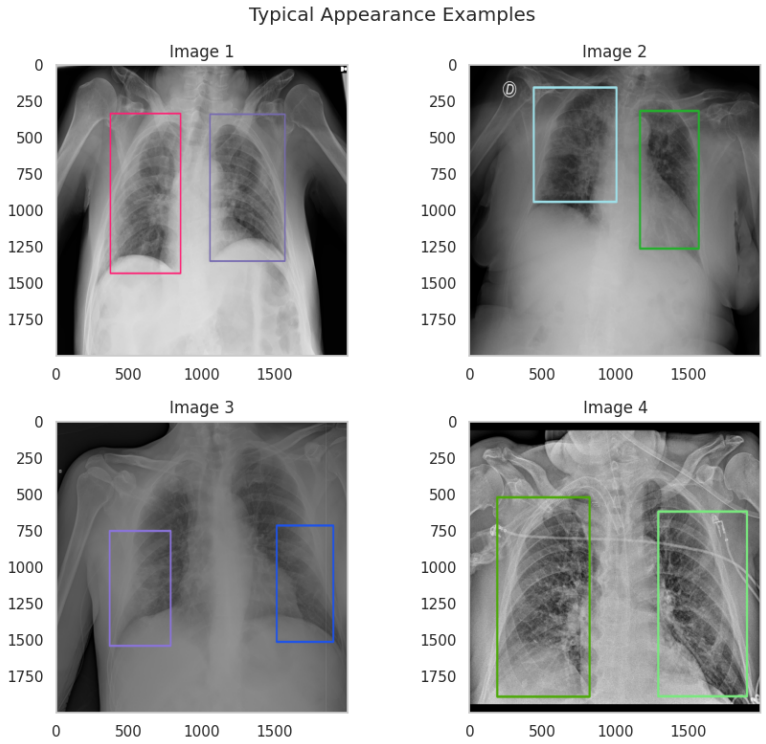

Figure 3.5: Examples of DICOM images for the Typical appearance case.

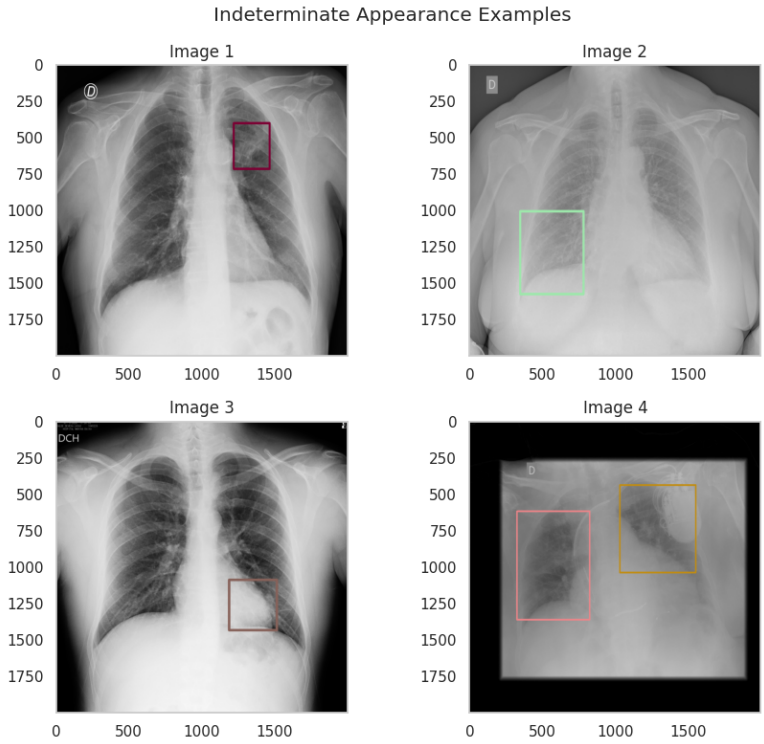

Figure 3.6: Examples of DICOM images for the Indeterminate appearance case.

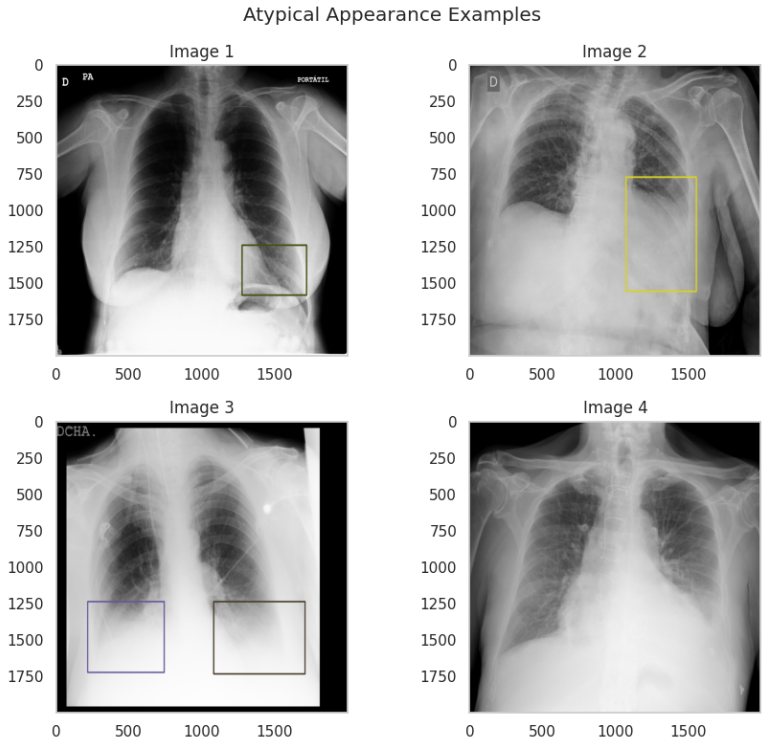

Figure 3.7: Examples of DICOM images for the Atypical appearance case.

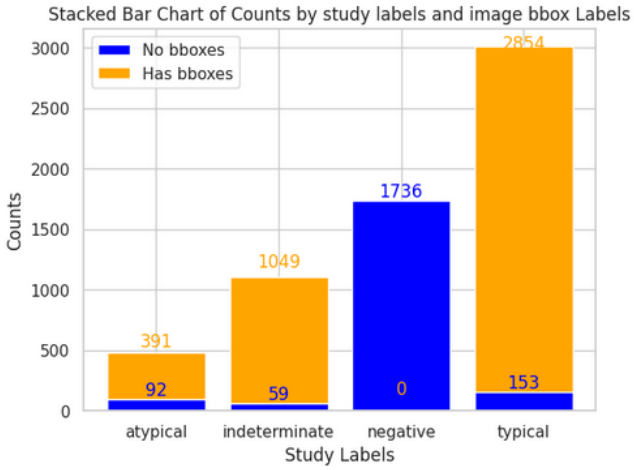

Figure 3.8: Stacked Bar Chart of counts by study-level classes and the presence of image bounding boxes labels.

become evident that X-rays without signs of pneumonia don't show opacities, therefore have no bounding boxes. This observation is further supported by looking at the graph, where it can be seen that while the *Negative for pneumonia* class has no bounding boxes present, almost all observations in the other classes do.

This finding is crucial because in future classification tasks, radiographs without bounding boxes for classes different from *Negative for pneumonia*, can be considered as outliers and be excluded from the study. It also ensures that the identification and characteristics of opacities play a key role in determining the study-level category, highlighting their importance in the classification process.

3.2.4 DICOM Metadata

To gain insights into the data earmarked for the study, understand its preparation, and pinpoint images that might deviate from the norm—potentially outliers or those exhibiting substantial differences from the majority—DICOM files with varied metadata values will undergo exploration. A reference to Figure 3.2 can be made to recollect potential metadata fields present in DICOM files.

The uncensored metadata fields studied are:

- **De-identification Method:** Method used to apply censorship to patient data that could help identify the subject.
- **Patient's Sex:** Biological sex of the patient.
- **Modality:** Imaging device or method used to acquire medical images[31].

- **Photometric Interpretation:** Representation of pixel data in an image and how it should be interpreted. It provides information on how the pixel values relate to the actual colors and brightness levels in the image[31].
- **Body Part Examined:** Label to identify the body part show in the DICOM image.
- **Private Creator:** Mechanism that allows vendors or institutions to define their own private data elements within the DICOM standard. The Private Creator tag provides a way for these entities to create custom data elements without conflicting with other standardized elements[31].

3.2.4.1 De-identification Method

Examining the distribution for **De-identification Method**, shown in Figure 3.9, various methods are employed to censor patient data. The predominant method, utilized in 92.5% of cases, is the *CTP Default method based on DICOM PS3.15 AnnexE*. Conversely, the least employed method, accounting for 2.8%, falls under the category *CTP Default*, which potentially refers to the same de-identification method as the majoritarian one. Lastly, the *RSNA Covid-19 Dataset Default* method is observed in 4.7% of cases.

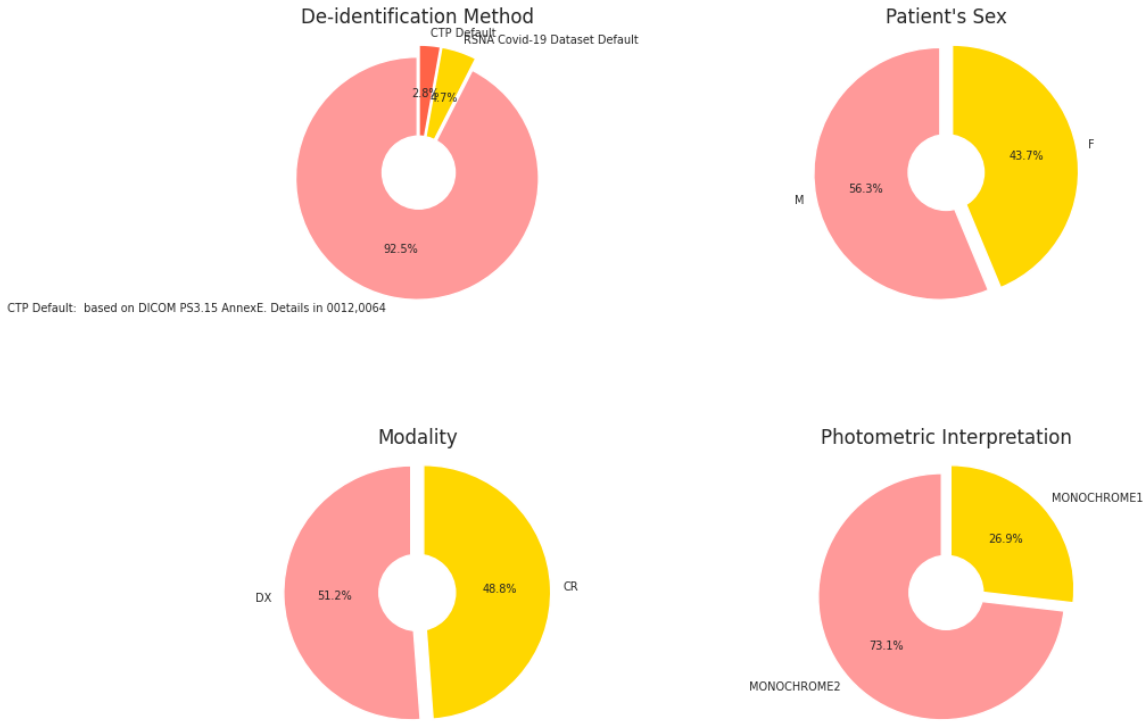

Figure 3.9: *De-identification Method*, *Patient's Sex*, *Modality* and *Photometric Interpretation* distributions.

3.2.4.2 Patient's Sex

Looking back at Figure 3.9, the balance of data representation for **Patient's Sex** is evident, with both males and females well represented. Maintaining balanced data in this category is critical, as there are physical differences between the two biological sexes that, if not balanced, could introduce bias into machine learning solutions. This physical differences can be appreciated in the visual exploration example from Figure 3.10.

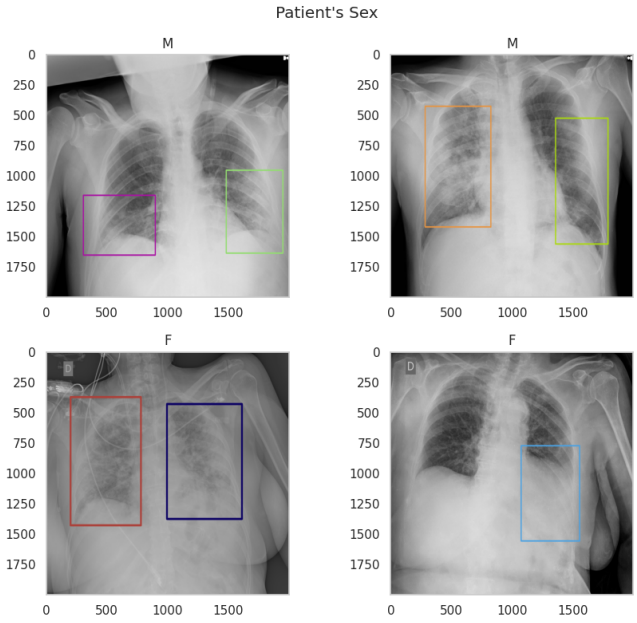

Figure 3.10: DICOM images examples for different *Patient sex* values.

3.2.4.3 Modality

Regarding **Modality**, a balanced representation is observed, in Figure 3.9, between two acquisition methods. CR (Computed Radiography), a method tat involves the use of a cassette-based system with a photostimulable phosphor plate, and DX (Digital Radiography), which employs digital detectors for direct X-ray image capture, eliminating the need for film and cassette systems[31]. No relevant visual differences were found between the two categories in the visual inspection carried out. An example of this is shown in Figure 3.11.

3.2.4.4 Photometric Interpretation

In the case of **Photometric Interpretation**, which can also be seen in Figure 3.9, there is an imbalance between the two possible interpretations; *MONOCHROME1* and *MONOCHROME2*, with the latter predominant at 73.1%. *MONOCHROME1* indicates that higher pixel values

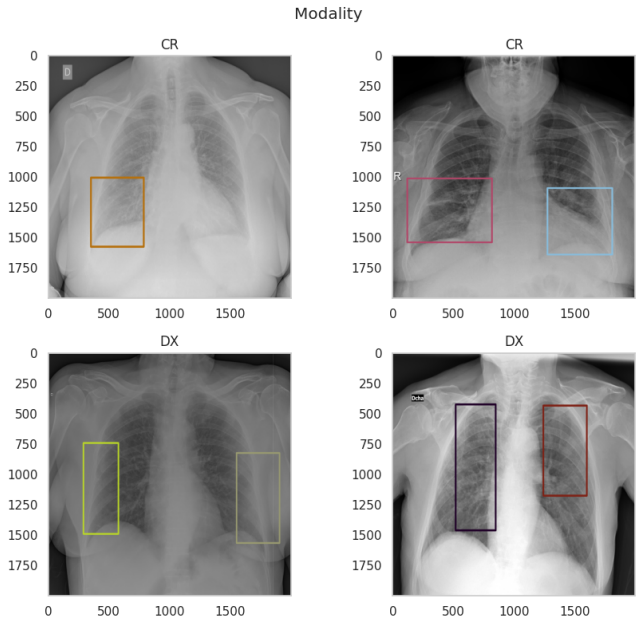

Figure 3.11: DICOM images examples for different *Modality* values.

correspond to darker shades, whereas *MONOCHROME2* interprets higher pixel values as lighter shades[31]. Given the significant representation of both categories in the data, this difference in pixel data interpretation highlights the need for data processing to ensure consistent pixel interpretation across all images.

3.2.4.5 Body Part Examined

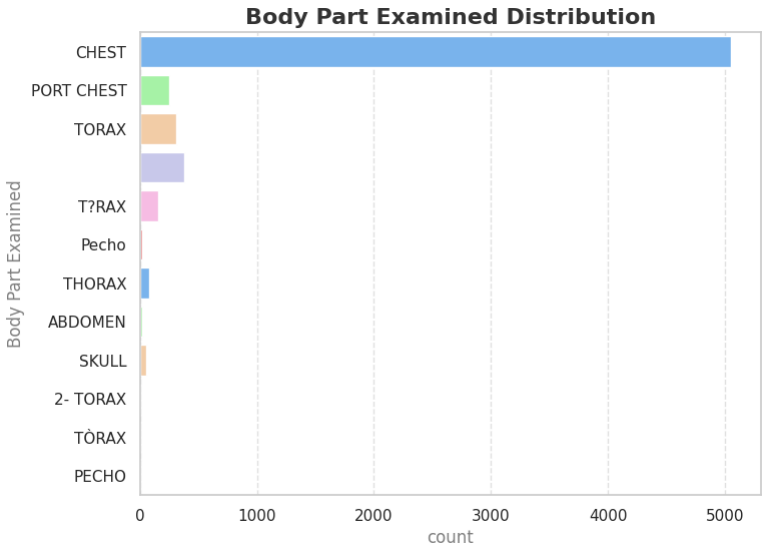

Figure 3.12: *Body Part Examined* Distribution.

Figure 3.12 shows the distribution of **Body Part Examined**. Most images use the tag *CHEST* to identify the body part shown in the image, or equivalent tags such as *TORAX* or *PORT CHEST*. Of relevance to this category, some DICOM files use the *SKULL* tag, which could indicate possible outliers. After a visual exploration of the data, an example of which can be seen in Figure 3.13, no relevant visual differences were found between images with different tags.

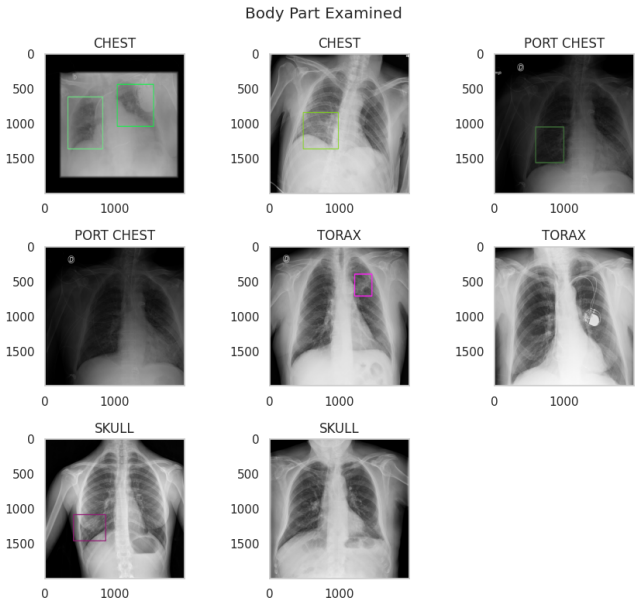

Figure 3.13: DICOM images examples for different *Body Part Examined* values.

3.2.4.6 Private Creator

The **Private Creator** distribution shown in 3.14 indicates that almost all DICOM files use the *GEIIS* value, there is a small amount of images using the *Philips RAD Imaging DD 097* value, and an insignificant amount using the *GEMS_GDXE_ATHENAV2_INTERNAL_USE* value. As shown in Figure 3.15, there were no relevant visual differences between different categories of **Private Creator**.

3.2.5 Conclusions

A number of key observations and considerations emerged from the exploratory data analysis.

The dataset shows a fair representation of both male and female sexes in the context of *Patient's Sex*. This balanced representation is crucial due to the physical differences between the two biological sexes. Failure to maintain this balance could introduce bias into subsequent machine learning solutions.

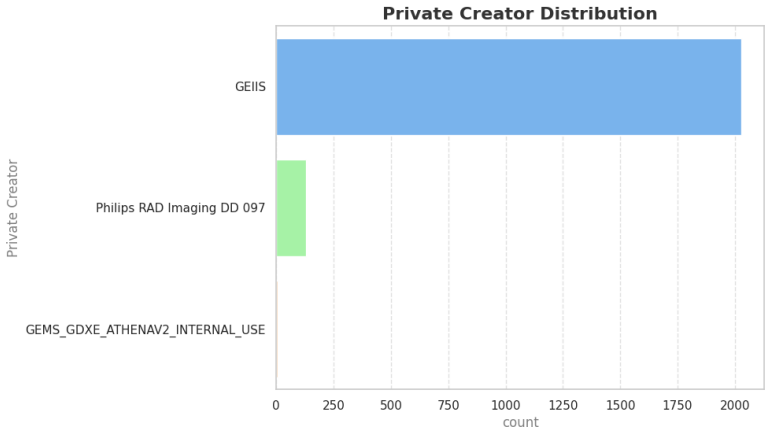

Figure 3.14: *Private Creator* Distribution.

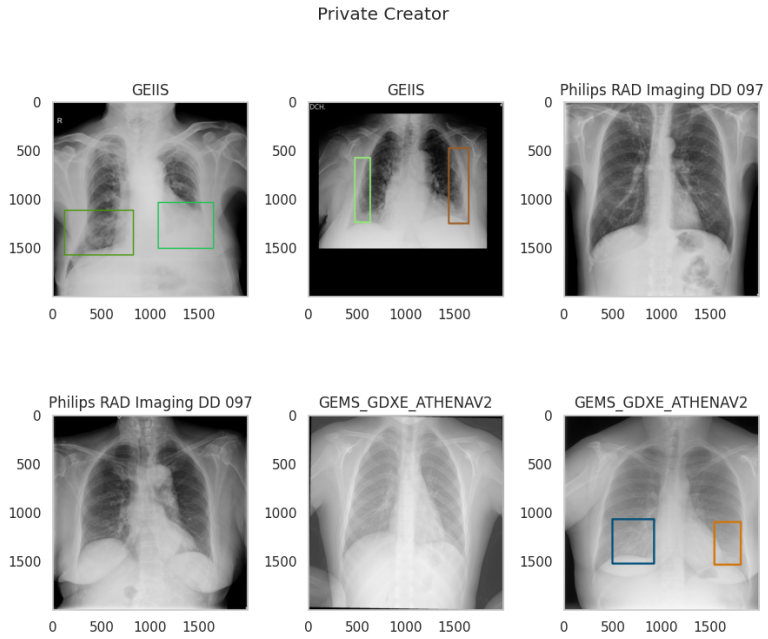

Figure 3.15: DICOM images examples for different *Private Creator* values.

Visual analysis of DICOM images with different metadata did not reveal any visual differences that could identify possible outliers.

The following observations highlight issues that need to be addressed during pre-processing of the dataset:

- **Study level class imbalance**: The training data has an imbalance in study-level classes. Balancing classes improves the model's ability to generalise across different study categories, contributing to robust and unbiased results.
- **Treating outliers at the image-level**: Localised outliers, particularly in images other

than the *Negative for pneumonia* class where no bounding boxes are present, should be excluded during pre-processing. This will ensure the integrity of subsequent analyses and promote a more accurate representation of relevant data.

- **Handling of photometric interpretation**: Given the substantial representation of both photometric interpretation categories, *MONOCHROME1* and *MONOCHROME2*, addressing this difference in pixel data interpretation is paramount. Applying data processing measures could ensure consistent pixel interpretation across all images, improving the overall quality and reliability of the dataset.

Chapter 4

Dataset Pre-processing

This chapter focuses on the crucial phase of dataset pre-processing; a fundamental step in preparing raw data for subsequent analysis. The primary objectives of dataset pre-processing are to ensure data quality, address potential biases, and establish a well-structured and standardised basis for machine learning tasks. The methods and considerations used to refine the SIIM-FISABIO-RSNA COVID-19 detection dataset are outlined, optimising it for effective model training and evaluation. The development of dataset pre-processing can be consulted in Kaggle notebook "DATA PREPROCESSING SIIM-COVID19-DETECTION"[32].

4.1 Photometric Interpretation Differences

The dataset exhibits significant representation in both *MONOCHROME1* and *MONOCHROME2* photometric interpretation categories, each with its own pixel data interpretation nuances. The MONOCHROME1 category implies that higher pixel values correspond to darker shades, with the darkest shade represented by the maximum pixel value. In contrast, MONOCHROME2 maintains a similar interpretation, but higher pixel values correspond to lighter shades, with the brightest shade represented by the maximum pixel value[31].

Addressing this divergence in pixel data interpretation becomes crucial during dataset pre-processing. By fostering consistency in how pixel values are interpreted, potential biases and inconsistencies are mitigated, contributing to more robust and accurate analyses in subsequent stages of the machine learning pipeline.

Read DICOM image array pixel values are inverted if *MONOCHROME1* is used, using the formula in 4.1 to match the photometric interpretation of *MONOCHROME2*.

$$\text{data} = \text{np.amax(data)} - \text{data} \tag{4.1}$$

4.2 Data Clean-up

Throughout the EDA, certain outliers were pinpointed—specifically, a small number of images that do not fall under the **Negative for Pneumonia** class and lack bounding boxes. As part of the data clean-up process, these outliers are excluded from the dataset. This meticulous step holds a great importance in preserving the accuracy and relevance of the dataset, mitigating potential distortions in subsequent analyses. The exclusion of outliers contributes to cultivating a cleaner and more reliable dataset

4.3 Handling of Class Imbalances

The exploratory data analysis has revealed an imbalance in study-level classes in training data, prompting the need for strategic interventions. To address the imbalance, data augmentation and exclusion strategies have been applied. Balancing classes in the dataset is crucial as it enhances the model's capacity to generalize across various study categories, fostering robust and unbiased results.

4.3.1 Data Augmentation

In the context of X-ray image data augmentation, careful consideration of the medical context and specific characteristics of X-ray imaging is essential. The following operations were studied[33].

- **Rotation:** Rotating X-ray images by small angles to simulate different perspectives, aiding the model's robustness to variations in positioning during imaging.
- **Flip (Horizontal and Vertical):** Applying horizontal and vertical flips to simulate X-ray images taken from different orientations, enhancing the model's ability to generalize.
- **Translation:** Shifting X-ray images horizontally or vertically to simulate slight changes in the patient's position or the X-ray machine's alignment.
- **Brightness and Contrast Adjustments:** Modifying the brightness and contrast of X-ray images, simulates variations in imaging conditions. It is important to ensure adjustments do not compromise diagnostic quality.
- **Noise Injection:** Adding different types of noise (e.g., Gaussian noise) to X-ray images to make the model more robust to noisy data.

- **Blur:** Applying slight blurring to simulate inherent blurriness present in X-ray images due to factors like motion or equipment limitations.

- **Cropping** Randomly cropping X-ray images to focus on specific regions of interest, helping the model learn to detect features in different parts of the image.

- **Gamma Correction**: Adjusting the gamma of X-ray images to simulate variations in the intensity of X-ray radiation.

A visual example of each operation can be seen in Figure 4.1.

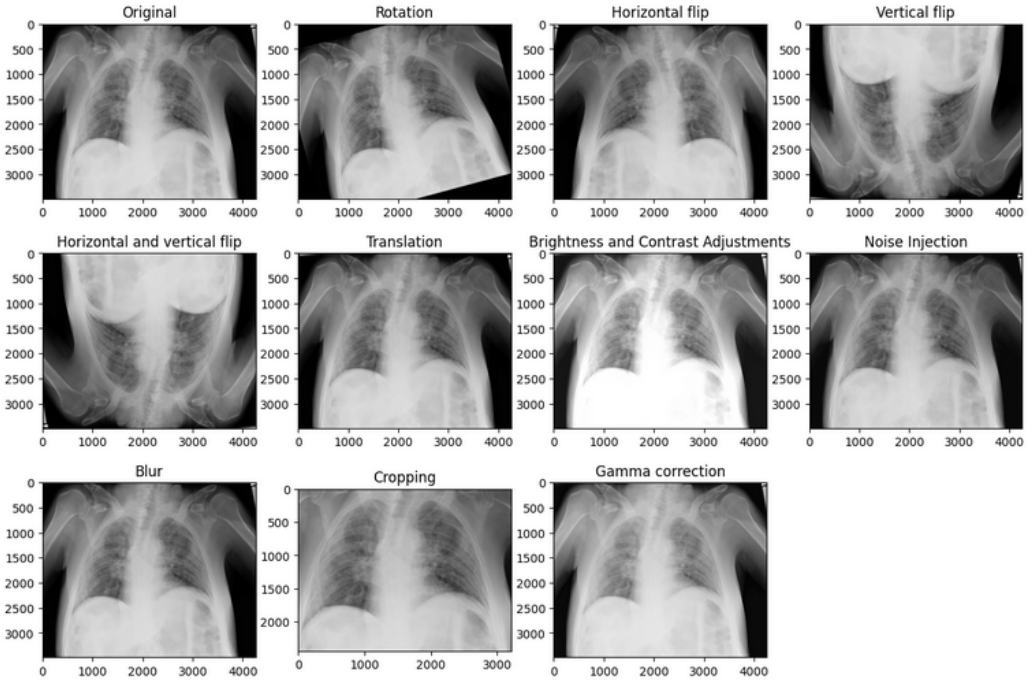

Figure 4.1: Data augmentation techniques.

It is crucial to approach data augmentation in medical imaging cautiously, considering patient safety and the potential impact on diagnostic accuracy.

After careful consideration, cropping, translation, and horizontal flip were selected as the chosen augmentation techniques, each of them to be applied with an small amount of blur and noise injection to avoid compromising clinically relevant data. Bounding boxes were adjusted according to each transformation. A comparison between original images and transformed images for each data augmentation strategy used, can be appreciated in Figures 4.2, 4.3 and 4.4.

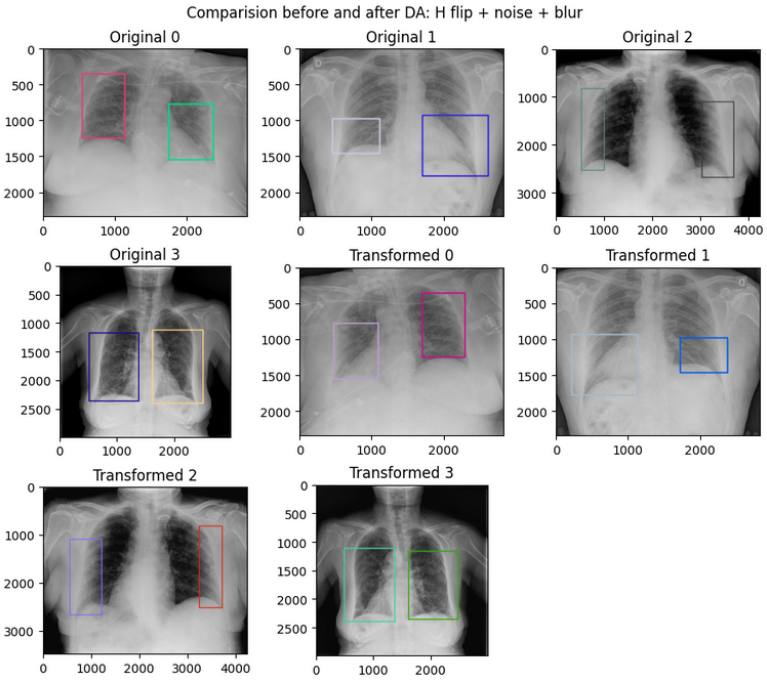

Figure 4.2: Comparison after applying an horizontal flip, blur and noise injection.

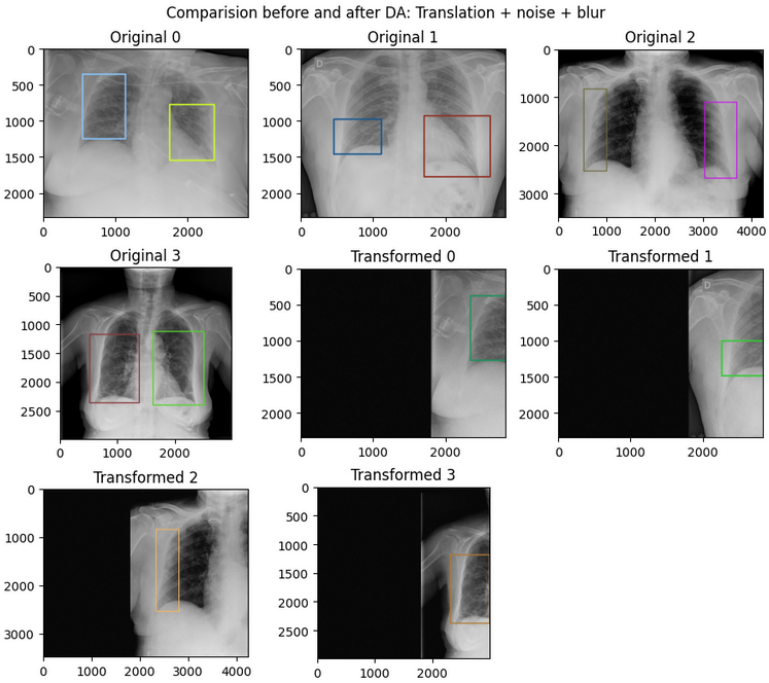

Figure 4.3: Comparison after applying a translation, blur and noise injection.

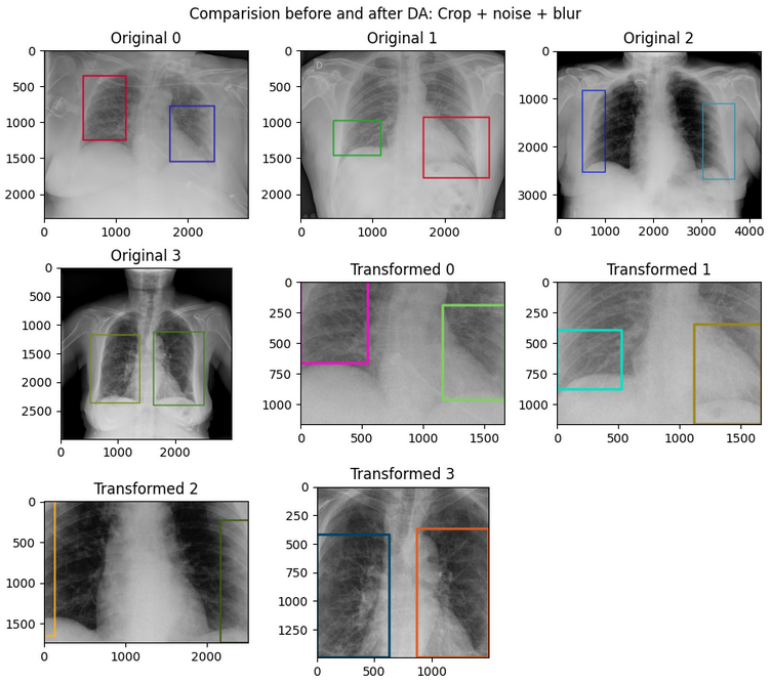

Figure 4.4: Comparison after applying cropping, blur and noise injection.

4.3.2 Excluding Data

To avoid over-generation of data, certain observations from the dominant class, *Typical Appearance*, were excluded to facilitate class balancing.

4.3.3 Class Balancing Strategy

The distribution of classes post data clean-up is depicted in Figure 4.5, comprising 2854 observations of **Typical Appearance**, 1736 of **Negative for pneumonia**, 1049 of **Indeterminate Appearance** and 391 of **Atypical Appearance**.

To address class imbalance, the **Atypical Appearance** class underwent augmentation through the three different strategies illustrated in Figures 4.2, 4.3 and 4.4. The **Indeterminate Appearance** class was augmented once using cropping with blurring and noise injection (Figure 4.4). The **Negative for pneumonia** class remained unaltered. Finally, a random sample of 800 observations was excluded from the **Typical Appearance** class to achieve balance. The resulting distribution, following the application of the balancing strategy, is visualized in Figure 4.6.

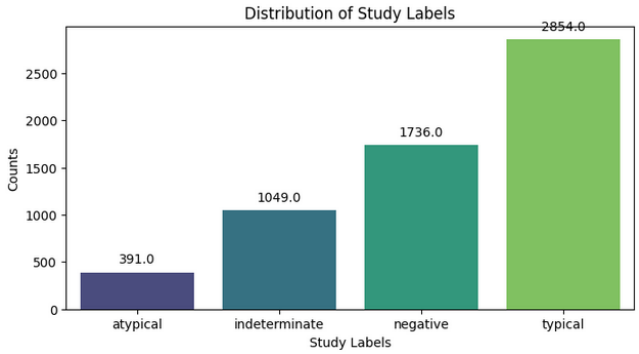

Figure 4.5: Class distribution after data clean-up.

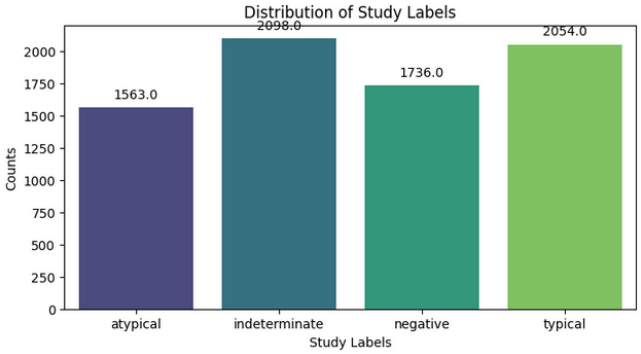

Figure 4.6: Class distribution after applying the class balancing strategy.

4.4 Resizing

Given the constraints of limited GPU free resources in Kaggle notebooks, the developing environment selected, resizing the data becomes a crucial element in the pre-processing pipeline. This purposeful adjustment seeks to optimize resource utilization, enhancing the efficiency of model training and evaluation within the confines of the Kaggle environment. Adapting the data size to the available GPU capacity contributes to a more streamlined and seamless machine learning development experience on Kaggle. However, it's essential to acknowledge that resizing may introduce a trade-off in terms of potential loss of finer details in the images. Consequently, images were resized to 256x256 pixels and ultimately stored in PNG format, facilitating easier future manipulation of the dataset.

Chapter 5

Object Detection and Classification

5.1 Introduction

Object detection and classification are fundamental tasks in Computer Vision with wide-ranging applications, particularly in the medical field, where the ability to identify and locate objects in images can be used to help diagnose diseases.

Object detection involves not only recognising the presence of objects in an image, but also determining their precise location using bounding boxes. Object classification, on the other hand, focuses on assigning one or more labels to an image, indicating the specific objects presence.

In the context of this study, the application of object detection and multi-label classification to chest radiographs for the identification of abnormalities related to COVID-19 is addressed. To tackle this complex problem, the You Only Look Once (YOLO) framework, in its fifth version, is employed, recognized as a state-of-the-art approach with real-time object detection capabilities.

The YOLO framework proves particularly well-suited to the task as it offers a holistic solution by simultaneously predicting bounding boxes and class labels for all objects in an image. This efficiency is deemed essential in medical imaging, where fast and accurate diagnosis is of critical importance.

This chapter provides the reasoning behind the development approach taken, delves into the YOLO framework, discusses the model selection, and outlines the evaluation metrics used to assess the performance of the object detection and classification system.

5.2 Justification of Development Approaches

In the pursuit of developing an effective object detection and classification system while considering available resources, several key considerations and development approaches have been adopted, each with specific justifications:

- **Ensuring Resource Efficiency:** Given limitations in free access to Kaggle resources, experiments are strategically designed for resource efficiency. The development approaches align with the constraints of the Kaggle environment, ensuring practical implementation.
- **Complex Problem Demands Complex Models:** In addressing the intricate nature of the medical imaging challenge, there is a need for sophisticated model architectures. Balancing the trade-off between extended training times and higher resource demands, the focus lies on identifying models that can be efficiently fine-tuned to reconcile both complexity and resource limitations. The YOLO framework has been selected for this purpose. Fine-tuning enables the adaptation of pre-trained object detection models to the specific nuances of medical imaging, optimizing resource utilization and simplifying the training process.
- **Risk of Overfitting:** Inherent in the project is the risk of overfitting, especially given the original imbalanced dataset. To address the imbalance, data augmentation techniques have been applied to certain classes, introducing a higher risk for overfitting for the specific classes. Overfitting happens when a model learns the training data too closely, capturing noise instead of underlying patterns. It leads to good performance on training data but poor generalization to new data. Overfit models are too complex and lack accuracy on unfamiliar data. The exploration of different model architectures aims to mitigate this risk and enhance generalization.
- **Introduction of the Negative for Pneumonia Class as Background:** Acknowledging the unique characteristics of the negative class, which is the only class with no objects associated, it has been introduced into training as background images. Adhering to YOLOv5 best practices, the background class should be limited to 10% of total images to avoid compromising the model's ability to detect objects. While evaluation metrics primarily address object-related classes, the inclusion of the negative class as background images ensures its indirect evaluation. This strategic inclusion contributes to the overall robustness of the model.
- **Performance Evaluation:** Decision-making for new experiments relies on insights from loss training curves and evaluation metrics. These metrics guide model selection and

inform the course of action based on observed performance trends during training and validation. Evaluation of model performance spans across training, validation, and test sets, derived from the pre-processed dataset, distributed in proportions of 80%, 10%, and 10%, respectively. This comprehensive evaluation ensures scrutiny across various subsets for a well-rounded assessment of efficacy.

5.3 Yolo Framework

The You Only Look Once (YOLO) framework is a pioneering deep learning open-source architecture renowned for its effectiveness in real-time object detection tasks. Developed to address the limitations of traditional two-step detection and classification pipelines, it takes an unified approach by simultaneously predicting bounding boxes and class labels in a single pass through the neural network[2]. YOLOv5 open-source repository is currently maintained by Ultralytics on GitHub[34].

One of its distinctive features is the ability to divide an input image into a grid and predict bounding boxes and class probabilities within each cell. This grid-based approach efficiently captures objects of various sizes and aspect ratios, providing a comprehensive understanding of the entire image.

A key strength lies in its versatility, offering various model architectures tailored to different needs. These architectures differ in terms of complexity, precision, and speed, allowing practitioners to choose the most suitable variant based on specific requirements.

Moreover, the framework supports fine-tuning, enabling adaptation to new problem domains. This capability empowers researchers and practitioners to adjust pre-trained models on custom datasets, making it adaptable to diverse object detection and classification tasks.

The YOLOv5 architecture, simplified in Figure 5.1, can be decomposed into three main components: the backbone, neck, and head. Each of these components plays a crucial role in the overall functioning of the object detection system. Components are described in Table 5.1.

Within the scope of this project, the YOLOv5 framework has been employed for object detection and multi-label classification on chest radiographs to identify various appearances potentially linked to Covid-19. Different YOLOv5 model architectures are explored, each exhibiting distinct complexities. The models undergo fine-tuning using the SIIM-FISABIO-RSNA COVID-19 Detection dataset, accompanied by variations in hyperparameters, all with the objective of achieving optimal results.

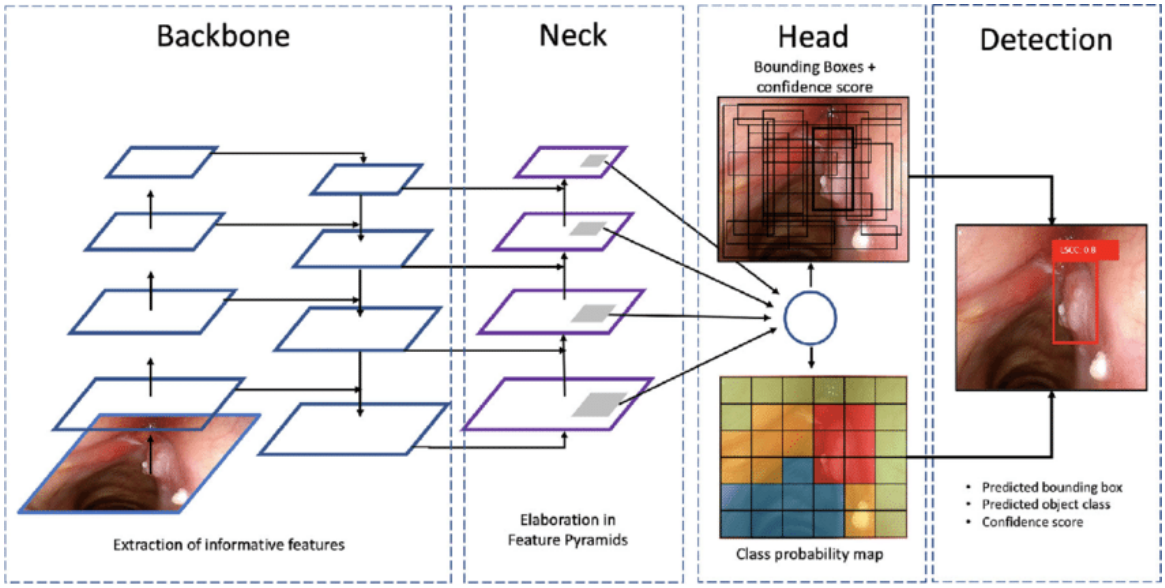

Figure 5.1: YOLOv5 simplified architecture[1].

5.3.1 YOLOv5 Loss Function

To further understand the training dynamics and guide the optimization process, it's essential to delve into the YOLOv5 loss function. The loss function serves as a critical component during the training phase, quantifying the disparity between predicted and actual values. Moreover, it has played a crucial role in guiding decisions for planning and conducting new experiments during the optimization process; a decreasing training loss indicates that the model is learning from the training data, while validation loss measures how well the model generalizes to new, unseen data. If the validation loss starts to increase while the training loss is still decreasing, it suggests that the model is overfitting and may not generalize well. YOLOv5's loss function [2] is determined by equation 5.1, where:

$$\text{Loss} = \lambda_1 \cdot \text{Object Loss} + \lambda_2 \cdot \text{Class Loss} + \lambda_3 \cdot \text{Box Loss} \tag{5.1}$$

- **Object Loss:** The object loss is calculated based on the binary cross-entropy between the predicted objectness score (indicating the presence of an object in a particular grid cell) and the ground truth objectness label. The loss penalizes the model for inaccuracies in determining whether an object exists in a given grid cell.

- **Class Loss:** Class loss is calculated using binary cross-entropy between the predicted class probabilities for each detected object and the actual class labels (ground truth) associated with those objects. This loss function penalizes deviations in the predicted class probabilities from the true class distribution. It guides the model to assign higher

Component	Function	Architecture	Details
Backbone	Serves as the foundation for feature extraction, capturing hierarchical features from the input image.	CSPDarknet53 backbone, an enhanced version of Darknet with a modified CSPNet module. Darknet consist on multiple convolutional layers with different filter sizes to capture features at different scales. CSPNet (Cross Stage Partial Network) is a neural network architecture designed to improve the efficiency and performance of convolutional neural networks.	Significant contribution to the model's parameters, influencing its ability to capture intricate patterns.
Neck	Enhances the model's ability to recognize objects and their contextual relationships.	PANet (Path Aggregation Network) facilitates information flow across different scales, aiding in the detection of objects of varying sizes.	Features from different backbone levels are integrated to create a feature pyramid.
Head	Responsible for making predictions based on extracted features. Predicts bounding boxes, class probabilities, and confidence scores.	Multiple convolutional layers followed by fully connected layers.	Produces predictions with confidence scores for determining the reliability of detected objects.

Table 5.1: Summary of YOLOv5 architecture components[2].

probabilities to the correct classes and lower probabilities to incorrect ones.

- **Box Loss:** The box loss is calculated using the Intersection over Union (IoU), depicted in equation 5.2, which quantifies the disparity between the predicted bounding box co-

ordinates and the ground truth coordinates. Minimizing the box loss during training is essential for the model to learn precise localization of objects in images.

$$IoU = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{5.2}$$

5.4 Model Selection

The exploration of YOLOv5 framework includes consideration of various model architectures[2], each designed to meet specific requirements in terms of computational efficiency and precision. Each YOLOv5 variant shares a common architectural structure, featuring multiple convolutional layers for effective feature extraction and fully connected layers for accurate bounding box and class predictions. The mean average precision (mAP), an evaluation metric, evaluated using the well-known COCO dataset[35] serves as a standardized benchmark, offering a consistent measure to assess model performance across different architectures. A breakdown of the YOLOv5 model variants tested in the study is provided in Table 5.2.

Model Variant	**Parameters**	**Model Size**	**Computational Efficiency**	**mAP**
YOLOv5s	7 million	Small	High	37.2%
YOLOv5m	21 million	Moderate	Balanced	45.2%
YOLOv5l	47 million	Large	Moderate	48.8%
YOLOv5x	90 million	Extra Large	Moderate to High	50.7%

Table 5.2: Summary of YOLOv5 Model Variants[2].

The decision on model selection revolves around thorough evaluation using metrics. The aim is to identify the most suitable architecture and hyper-parameters that achieve a balance between capturing intricate patterns in the data and avoiding overfitting. While more complex models with a greater number of parameters may excel at capturing nuanced patterns, the risk of overfitting must be carefully considered. Simpler models, less prone to overfitting, may outperform highly complex models in terms of generalization to new, unseen data. The iterative experimentation with different model architectures allows for the informed selection of a model that performs well across both training validation datasets. The ultimate assessment of the model's performance will be conducted using a test dataset comprising data that has not been previously encountered by the model.

5.5 Evaluation Metrics

In the evaluation of model performance, a set of key metrics is employed to provide nuanced insights into predictive capabilities[2]. Each metric serves a specific purpose, collectively offering a comprehensive assessment of the model's strengths and areas for improvement. Let's delve into these metrics:

5.5.1 Confusion Matrix

The confusion matrix offers a tabular representation of model predictions, detailing the distribution of true positive, true negative, false positive, and false negative instances. It provides a detailed examination of model performance in various prediction scenarios.

- **True Positives (TP):** Instances where the model correctly detects and classifies objects.
- **True Negatives (TN):** Instances where the model correctly identifies the absence of objects.
- **False Positives (FP):** Instances where the model incorrectly detects objects that are not present in the ground truth.
- **False Negatives (FN):** Instances where the model fails to detect objects that are present in the ground truth.

In the confusion matrix, the predicted classes are listed on the left, while the true classes are listed at the bottom. The values on the diagonal of the matrix represent True Positives (TP) and True Negatives (TN), making higher values on this diagonal indicative of the model's proficiency in making accurate predictions.

A critical consideration when computing confusion matrices for object detection lies in the treatment of the background class. The background class is a default class which is assigned to grid cells without detected objects in it. The confusion matrix primarily aims to assess the model's ability to detect objects across various classes. As a result, it typically calculates true positives (TP) and false negatives (FN) for the background class in relation to object-related classes, therefore an empty background-background cell in the confusion matrix is expected[36].

5.5.2 Precision (P)

Precision is a fundamental metric that measures the accuracy of positive predictions made by the model. Calculated as shown in equation 5.3, it quantifies the proportion of correctly

predicted positive instances (TP) relative to all positive predictions (TP + FP). High precision indicates proficiency in correctly identifying positive instances, albeit without considering missed positives.

$$P = \frac{\text{True Positives}}{\text{True Positives + False Positives}} \tag{5.3}$$

5.5.3 Recall (R)

Recall, also known as Sensitivity or True Positive Rate, evaluates the model's ability to capture all positive instances. It is calculated as shown in equation 5.4, and it quantifies the ratio of correctly predicted positive instances (TP) relative to all actual positive instances (TP + FN). A high recall score suggests that the model effectively identifies most positive instances, although it may also lead to more false positives.

$$R = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \tag{5.4}$$

5.5.4 Recall-Confidence Curve

The Recall-Confidence curve illustrates how recall varies at different confidence thresholds. It shows how the recall of the model changes as you adjust the confidence threshold for positive predictions. It helps in understanding the trade-off between recall and confidence. For instance, you can see how recall increases or decreases as you set higher or lower confidence thresholds for positive predictions.

5.5.5 Precision-Recall (PR) Curve

The Precision-Recall curve illustrates how changes in the decision threshold, which determines the classification boundary for predicted probabilities, impact precision and recall. A curve closer to the upper-right corner indicates superior model performance. The area under the PR curve (AUC-PR) offers a summarized evaluation of the model's precision-recall trade-off.

5.5.6 Mean Average Precision (mAP)

Mean Average Precision (mAP) serves as a comprehensive metric in assessing the performance of object detection models. It quantifies the area under the precision-recall curve, offering a holistic measure of the model's precision-recall balance. This evaluation is conducted based on the IoU (Intersection over Union) threshold, where 0.5 is a commonly used threshold denoted as mAP50.

For a more nuanced assessment, the notation mAP0.5:0.95 is employed. This indicates that the mean Average Precision is calculated by averaging precision values across a range of IoU thresholds, specifically from 0.5 to 0.95. This broader range provides a detailed evaluation of the model's detection capabilities, considering varying degrees of overlap between predicted and ground truth bounding boxes.

5.5.7 F1 Score Curve

The F1 Score, a harmonic mean of precision and recall, calculated with equation 5.5, which provides a balanced perspective on model performance. A high F1 score indicates that the model has a good balance between precision and recall. It means that the model is effective at both correctly identifying positive instances (precision) and capturing all relevant positive instances (recall). The F1 Score curve showcases how this metric varies across different decision thresholds, aiding in the identification of the threshold that optimally balances precision and recall[37].

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.5}$$

Chapter 6

Experiments

6.1 Experiment Overview

In this chapter, a thorough exploration of various YOLOv5 model architectures for object detection in chest radiographs is presented. The objective is to conduct experiments aimed at comprehending factors that could significantly impact the final performance of the models. The development of the experiments can be consulted in the Kaggle notebook titled "Detection on SIIM-COVID-19-DETECTION [YOLOv5]"[38]. Additionally, the "kaggle-siim-covid19" project[9] on the Weights & Biases platform serves as a repository for comprehensive experiment tracking, encompassing experiment configurations, training dynamics and evaluation on the validation set.

6.1.1 Experimental Setup

The experiments were conducted on the Kaggle platform, utilizing a GPU P100 as an accelerator. It's noteworthy that the free usage of this accelerator is constrained to 30 hours per week, imposing limitations on the number of experiments that could be performed. The development process was initially planned to leverage the Kaggle SIIM COVID-19 Detection competition submission platform for the evaluation of model results, providing accurate assessments on unseen data, but due to unknown errors it was not possible (the platform does not provide information on where the error is when a submission fails). For this reason, the original training dataset has been was splited in three subsets: training (80%), validation (10%), and test (10%). It's important to acknowledge that original dataset underwent data augmentation, potentially introducing some similarities in validation and test sets to the training data and consequently affecting the evaluation process. In order to save resources, x-rays from the dataset were resized to 256x256.

Furthermore, evaluations in both the test and validation sets were conducted at an Intersection over Union (IoU) threshold of 0.1. The IoU threshold defines the degree of overlap required for a predicted bounding box to be considered a true positive detection.

6.1.2 Key Experiments

6.1.2.1 No Data Augmentation Experiment

This experiment aimed to assess the necessity of employing additional data augmentation techniques, considering that the YOLO framework inherently applies some augmentations as needed. The objective was to evaluate the model's proficiency in predicting minority classes. YOLOv5m was chosen for this experiment due to its moderate complexity, and the experiment was conducted with a training duration of 100 epochs.

6.1.2.2 Background Class Experiment

A special experiment was designed to investigate the role of the YOLO's default background class. The goal was to help the model understand that **Negative for pneumonia** class can be considered 100% background. This experiment compared two scenarios: one including all **Negative for pneumonia** images to the training process and another, adhering to YOLOv5 training guidelines[2], where only 10% of the total images should correspond to full background images. The hypothesis posits that incorporating the negative class for pneumonia as background images would enable the model to recognize this class as devoid of objects, or background, potentially lowering the false positive (FP) rate for other classes. However, limiting these background images to 10% of the total number of images aims to prevent the model from not properly learning to detect objects, consequently reducing the false negative (FN) rate for classes associated with objects. Each resulting model's performance is later evaluated using evaluation metrics.

6.1.2.3 Individual Exploration of YOLOv5 Model Architectures

Each YOLOv5 model architecture—YOLOv5s (Small), YOLOv5m (Medium), YOLOv5l (Large), and YOLOv5x (Extra Large)—has been examined independently. The goal was to understand how the varying complexities of these architectures impact their ability to detect abnormalities in chest radiographs related to COVID-19. The experimentation phase involved the fine-tuning of each YOLOv5 variant using the SIIM-FISABIO-RSNA COVID-19 Detection dataset. This process included an automatic search for optimal hyperparameters, adjustments to the number of epochs based on observed losses during training and validation, and an evaluation of each resulting model's performance on both the validation and test sets using evaluation metrics.

6.1.2.4 Ensemble Learning

An ensemble learning approach was employed by combining the outputs of individual YOLOv5 model architectures. This strategy aims to enhance overall predictive performance and robustness by leveraging the unique strengths of each model. The ensemble method involves concatenating predictions from multiple models and utilizing Non-Maximum Suppression (NMS) as a post-processing technique. NMS helps refine the final predictions by eliminating redundant and overlapping bounding boxes, ensuring the selection of the most confident and accurate detections. This approach aims to improving the overall reliability and performance of the object detection system.

6.2 Results

6.2.1 No Data Augmentation Experiment

The results of the experiment, which excluded the application of data augmentation techniques to the dataset, can be readily seen by examining the resulting confusion matrix, shown in Figure 6.1.

The confusion matrix shows that the model struggles to predict **indeterminate** and **atypical** classes, often misclassifying them as background. It can be concluded that the application of data augmentation is imperative to improve the model's ability to effectively predict these classes. The evaluation metrics averaged over all classes from the experiment's model are shown in Table 6.1. These results will serve as a baseline model to be improved. It is important to note that data augmentation techniques described in Chapter 4 are applied when conducting the following experiments.

Experiment	P	R	mAP50	mAP0.5:0.95
Baseline	0.241	0.283	0.22	0.071

Table 6.1: Baseline model resulting from the experiment where data augmentation techniques are not applied in the dataset. The Experiment has been conducted using YOLOv5m with 100 epochs, and has been evaluated on Validation set.

6.2.2 Background Class Experiment

Examining Table 6.2, it becomes apparent that the evaluation metrics slightly improve when adhering to the recommended amount of **Negative for pneumonia** images as background images, as suggested by YOLOv5 guidelines. Although the observed difference may not be

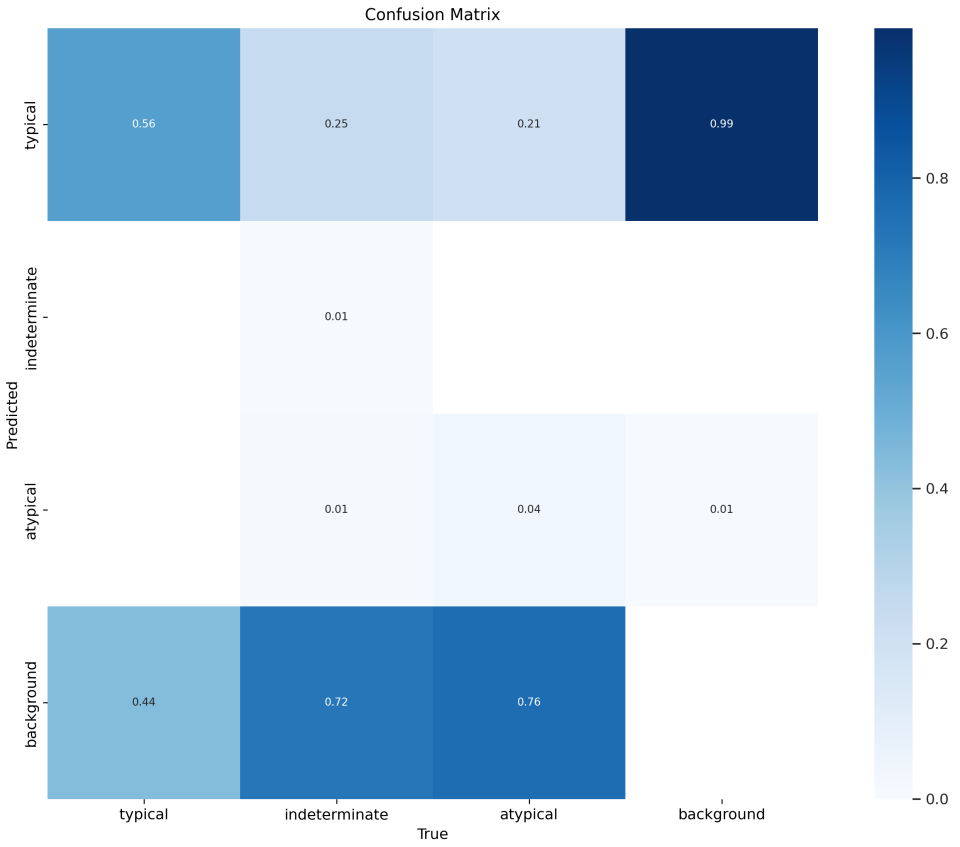

Figure 6.1: YOLOv5m's confusion matrix resulting from the experiment where data augmentation techniques are not applied in the dataset. The Experiment has been conducted using YOLOv5m with 100 epochs, and has been evaluated on Validation set.

Experiment	**P**	**R**	**mAP50**	**mAP0.5:0.95**
All Backgrounds	0.845	0.687	0.744	0.523
10% Backgrounds	0.849	0.703	0.751	0.547

Table 6.2: Background Class Experiment using YOLOv5m with 300 epochs, and different amount of background images, evaluated on Validation set.

substantial, it validates the hypothesis put forth. The overall reduction in false positive and false negative rates is marginally better when limiting the number of background images to 10% as observable in resulting confusion matrices from both models showcased in Figures 6.2 and 6.3. Moving forward, the experiments will align with the training guidelines recommended in YOLO's framework manual [2], reducing the amount of negative for pneumonia images introduced in training as all background images to 10%.

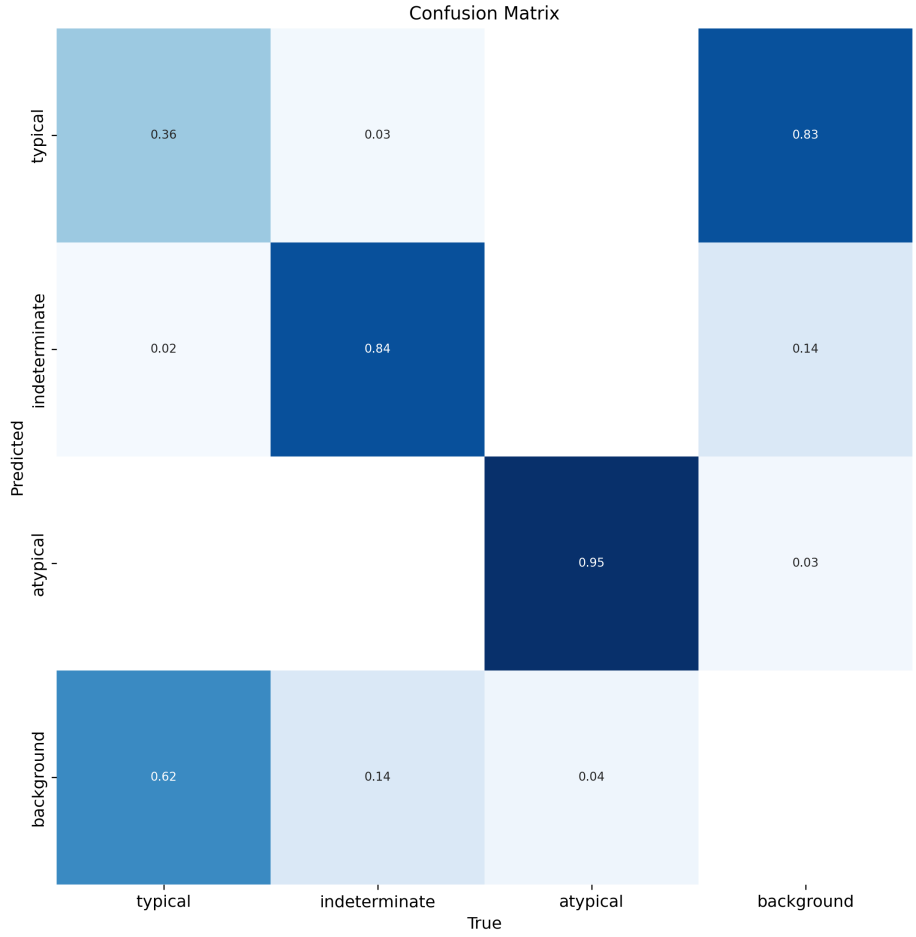

Figure 6.2: YOLOv5m's confusion matrix evaluated on Validation set for the experiment where all Negative for Pneumonia images where included in the training of the model as background images.

6.2.3 Individual Exploration of YOLOv5 Model Architectures

6.2.3.1 Validation set

Table 6.3 contains the performance evaluation of experiments performed on individual YOLO models for the validation set.

Analyzing the performance metrics across various experiments reveals nuanced insights into the effectiveness of different YOLOv5 model configurations. YOLOv5s, YOLOv5m and YOLOv5l architectures that have been trained for 500 epochs, exhibits favorable precision, recall, and mAP50 scores across all classes. However, a closer examination of the training dynamics reveals that YOLOv5s and YOLOv5m with 500 epochs experienced overfitting, evident in a significant increase in validation object loss from epoch 300 onwards as it can be seen in Figures 6.4 and 6.5. The same case can be observed for YOLOv5l with 500 epochs, where

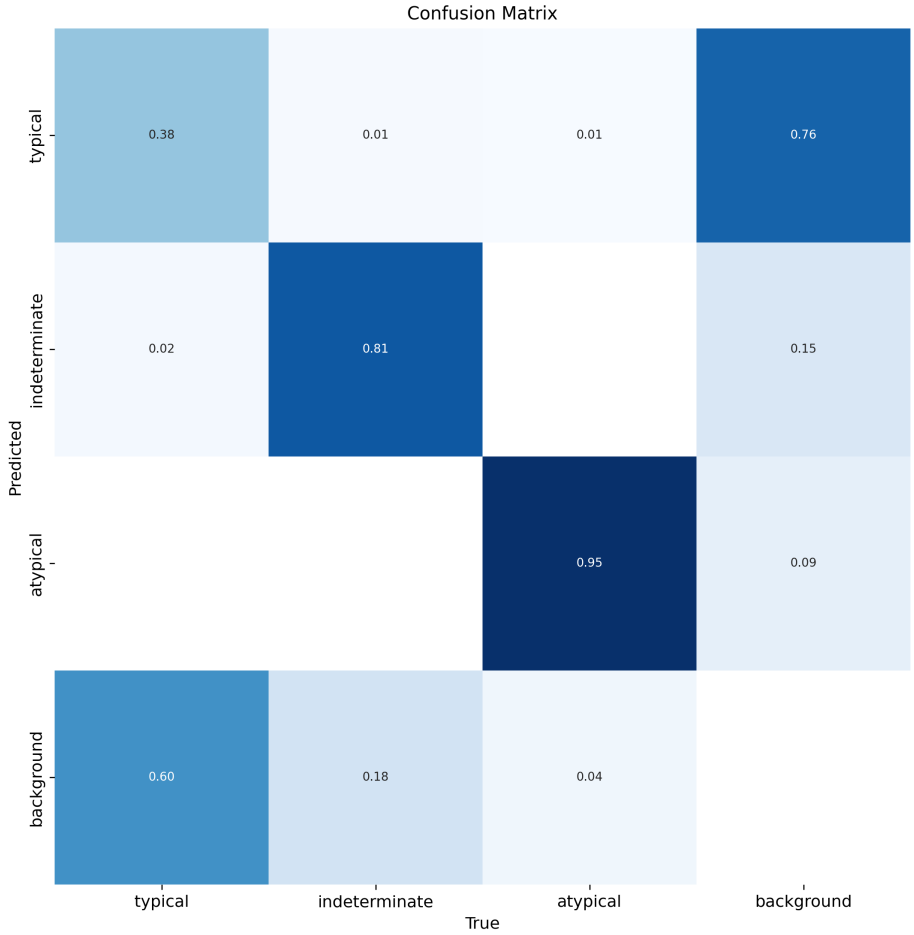

Figure 6.3: YOLOv5m's confusion matrix evaluated on Validation set for the experiment where 10% of Negative for Pneumonia images where included in the training of the model as background images.

while training object loss decreases, validation object loss increases significantly from epoch 100 onwards, observed in Figure 6.6. This overfitting raises concerns about the model's ability to generalize effectively to real-life instances. For these reasons, these models have been discarted from further explorations.

The introduction of data augmentation for the **Indeterminate** and **Atypical** classes positively contributes to the model's capability to predict these classes, demonstrating the model's ability to learn from augmented instances. However, the risk of overfitting for these specific classes is inherent in this process, with the model potentially learning the augmented examples too well, which may hinder its generalization to real-life scenarios of these categories.

Additionally, certain architectural choices, such as YOLOv5m, trained at 300 epochs, YOLOv5l and YOLOv5x, trained at 100 epoch, demonstrate promising results, highlighting the possible benefit of ensemble learning to improve overall predictions.

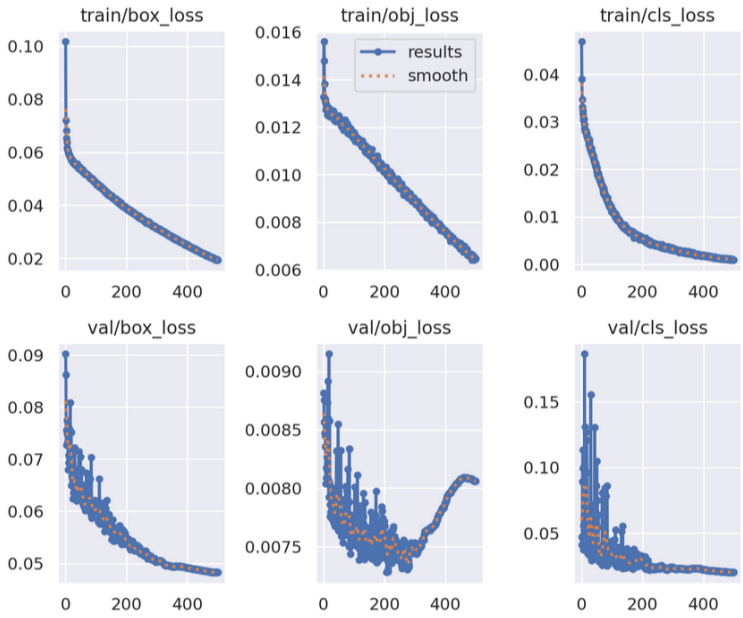

Figure 6.4: Training dynamics of YOLOv5s trained with 500 epoch.

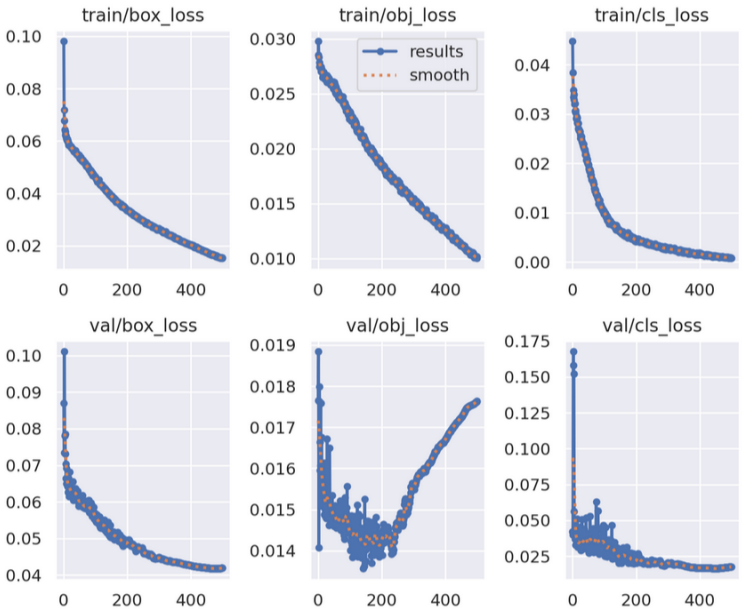

Figure 6.5: Training dynamics of YOLOv5m trained with 500 epoch.

6.2.3.2 Test set

The remaining effective models were assessed on the test set, and the evaluation metrics are presented in Table 6.4. Overall, all studied models demonstrated satisfactory performance in both validation and test sets. Notably, YOLOv5m, trained for 300 epochs, outperformed the other models. The confusion matrix (Figure 6.7) reveals YOLOv5m's capability to accurately

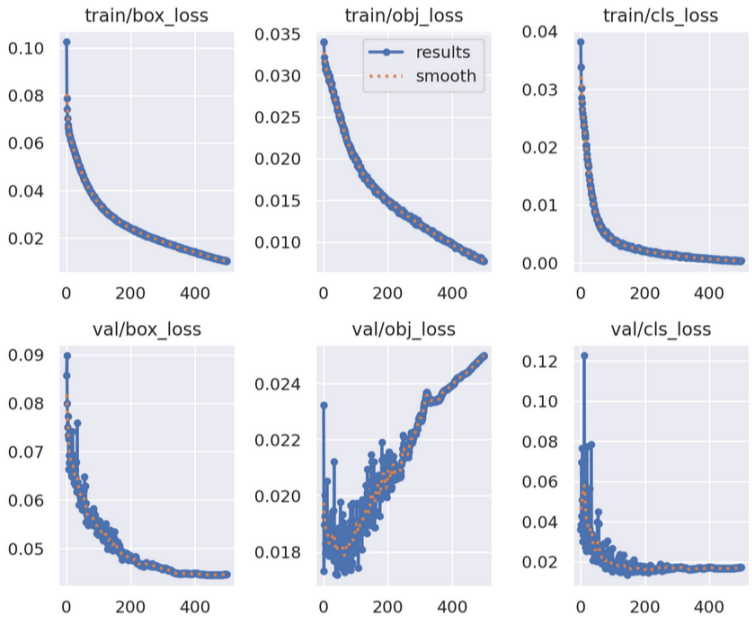

Figure 6.6: Training dynamics of YOLOv5l trained with 500 epoch.

identify *Atypical* and *Indeterminate* appearances, with some confusion in differentiating *Typical for pneumonia* objects. The F1-Confidence curve (Figure 6.8) highlights a balanced precision and recall for *Atypical* and *Indeterminate* classes, but a notable imbalance for the *Typical* class. Similarly, the Precision-Recall curve (Figure 6.10) indicates strong model performance for *Atypical* and *Indeterminate* classes but a lower curve for the *Typical* class. Lastly, the Recall-Confidence curve illustrates how recall varies with increasing confidence thresholds, with an expected reduction as the threshold increases. For an example of model predictions compared to ground truth, refer to Figures 6.11 and 6.12.

6.2.4 Ensemble Learning

The ensemble learning technique was applied by combining YOLOv5s (trained for 300 epochs), YOLOv5m (trained for 300 epochs), YOLOv5l (trained for 100 epochs), and YOLOv5x (trained for 100 epochs). The results of this ensemble approach are presented in Table 6.5 for both the Validation and Test sets.

In contrast to the YOLOv5m model, which was identified as the top performer in the *Individual Exploration of YOLOv5 Model Architectures* experiment, the ensemble of YOLOv5s (300 epochs), YOLOv5m (300 epochs), YOLOv5l (100 epochs), and YOLOv5x (100 epochs) demonstrates comparable results, albeit with a slight superiority retained by YOLOv5m. The confusion matrix depicted in Figure 6.13 illustrates a notable enhancement in minimizing false positives and false negatives, particularly for the *Typical* class. This improvement is evident in

Experiment	Epochs	Class	P	R	mAP50	mAP0.5:0.95
YOLOv5s	300	all	0.773	0.562	0.644	0.391
		typical	0.62	0.325	0.385	0.115
		indeterminate	0.831	0.5	0.633	0.385
		atypical	0.868	0.862	0.913	0.673
YOLOv5s	500	all	0.767	0.642	0.698	0.47
		typical	0.557	0.332	0.388	0.119
		indeterminate	0.838	0.69	0.766	0.553
		atypical	0.906	0.903	0.941	0.737
YOLOv5m	300	all	0.846	0.692	0.744	0.535
		typical	0.656	0.329	0.416	0.131
		indeterminate	0.922	0.789	0.842	0.661
		atypical	0.961	0.959	0.975	0.811
YOLOv5m	500	all	0.826	0.71	0.751	0.568
		typical	0.645	0.365	0.419	0.14
		indeterminate	0.88	0.805	0.857	0.715
		atypical	0.953	0.959	0.977	0.849
YOLOv5l	500	all	0.865	0.71	0.743	0.58
		typical	0.687	0.332	0.383	0.127
		indeterminate	0.949	0.83	0.863	0.571
		atypical	0.96	0.969	0.984	0.86
YOLOv5l	100	all	0.822	0.675	0.718	0.469
		typical	0.628	0.374	0.389	0.116
		indeterminate	0.903	0.727	0.806	0.537
		atypical	0.934	0.923	0.959	0.754
YOLOv5x	100	all	0.855	0.674	0.731	0.494
		typical	0.667	0.358	0.415	0.12
		indeterminate	0.925	0.74	0.808	0.575
		atypical	0.973	0.926	0.97	0.786

Table 6.3: Performance metrics for experiments with YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x evaluated on the Validation set.

the Recall-Confidence curve (Figure 6.15), surpassing the performance of the YOLOv5m Recall-Confidence curve (Figure 6.9). However, both the F1-Confidence curve (Figures 6.14 and 6.8) and Precision-Recall curve (Figures 6.16 and 6.10) indicate that the YOLOv5m model consistently outperforms the Ensemble learning approach. For an illustration of model predictions in comparison to the ground truth, consult Figures 6.17 and 6.18.

Experiment	Epochs	Class	P	R	mAP50	mAP0.5:0.95
YOLOv5s	300	all	0.72	0.573	0.657	0.429
		typical	0.473	0.295	0.359	0.13
		indeterminate	0.787	0.57	0.711	0.498
		atypical	0.899	0.854	0.901	0.659
YOLOv5m	300	all	0.832	0.687	0.765	0.566
		typical	0.609	0.33	0.447	0.167
		indeterminate	0.915	0.767	0.869	0.71
		atypical	0.972	0.963	0.98	0.82
YOLOv5l	100	all	0.767	0.643	0.717	0.492
		typical	0.519	0.325	0.401	0.148
		indeterminate	0.887	0.703	0.811	0.597
		atypical	0.895	0.9	0.939	0.731
YOLOv5x	100	all	0.821	0.652	0.742	0.521
		typical	0.573	0.299	0.413	0.153
		indeterminate	0.919	0.717	0.841	0.622
		atypical	0.971	0.941	0.971	0.789

Table 6.4: Performance metrics for experiments with YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x evaluated on the Test set.

Set	Class	P	R	mAP50	mAP0.5:0.95
Validation	All	0.83	0.695	0.776	0.558
	Typical	0.635	0.35	0.477	0.159
	Indeterminate	0.909	0.79	0.872	0.69
	Atypical	0.945	0.944	0.979	0.827
Test	All	0.797	0.688	0.756	0.548
	Typical	0.563	0.356	0.433	0.157
	Indeterminate	0.896	0.753	0.848	0.675
	Atypical	0.933	0.955	0.987	0.812

Table 6.5: Ensemble Learning performance metrics on Test and Validation Sets

6.3 Conclusions

In summary, the experiments underscore the vital role of balancing the dataset through the implementation of data augmentation (DA) techniques, crucial for effective prediction of minority classes. The application of DA, while enhancing model performance overall, introduces the potential for overfitting in the **Indeterminate Appearance** and **Atypical Appearance** classes. This possibility is suggested by the notable disparity in results for these classes compared to the **Typical Appearance** class, particularly in the augmented test and validation sets.

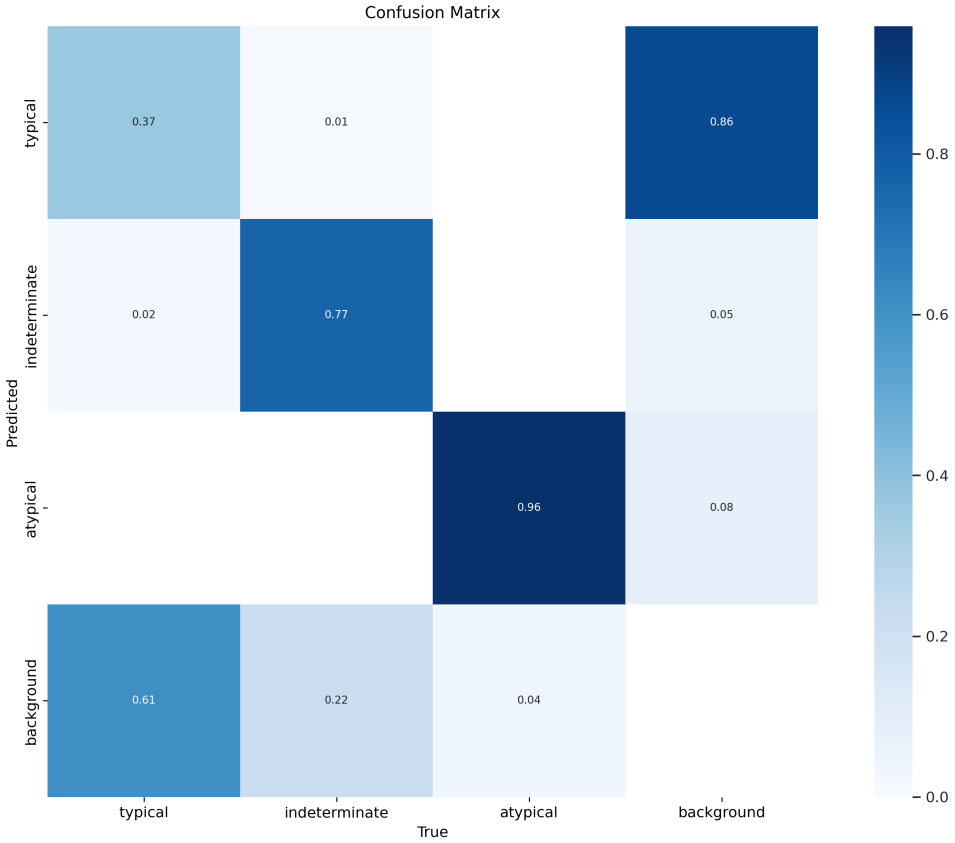

Figure 6.7: YOLOv5m's confusion matrix for the Test set.

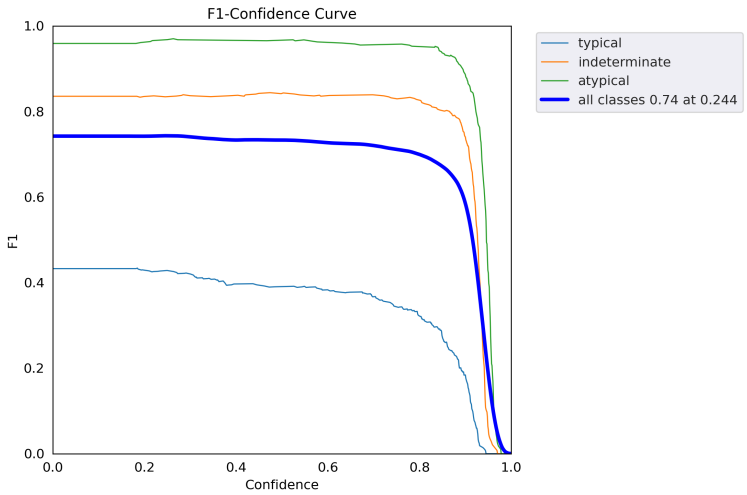

Figure 6.8: YOLOv5m's F1-Confidence curve for the Test set.

Moreover, reducing the number of negative for pneumonia images used as background during training to 10% of the total dataset leads to a slight improvement in overall model performance.

YOLOv5m, trained for 300 epochs, emerges as the standout performer, consistently ex-

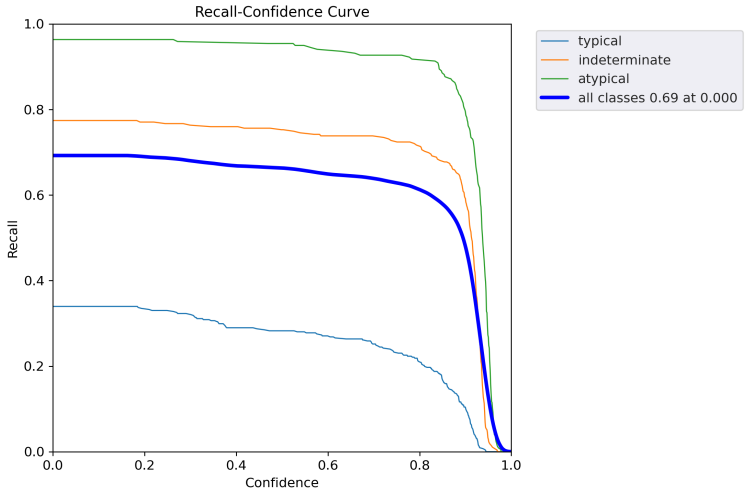

Figure 6.9: YOLOv5m's Recall-Confidence for the Test set.

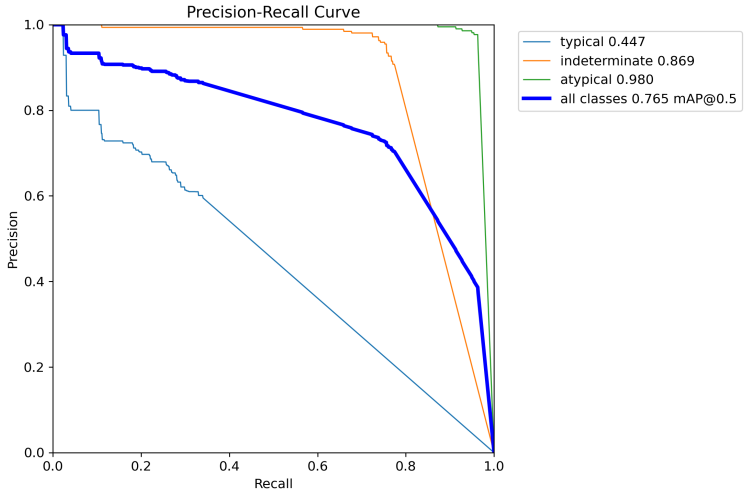

Figure 6.10: YOLOv5m's Precision-Recall curve for the Test set.

celling in predicting **Indeterminate Appearance** and **Atypical Appearance** classes while displaying relatively modest performance for the **Typical Appearance** class.

Although ensemble learning demonstrates enhanced results for the **Typical Appearance** class, the overall model performance remains superior in the YOLOv5m model. Lastly, evaluating accuracy for the **Negative for pneumonia** class has not been possible given the inherent nature of object prediction models.

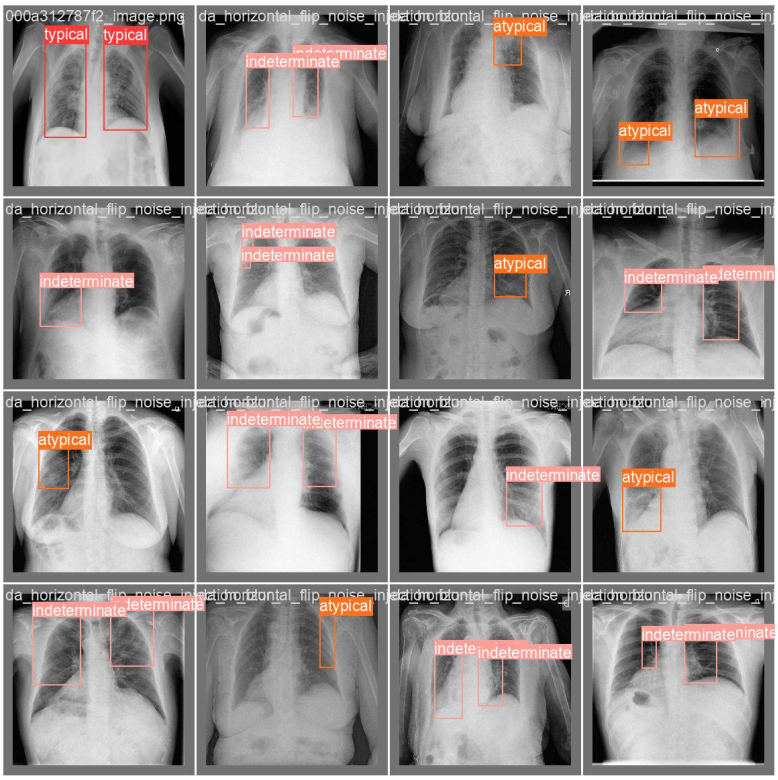

Figure 6.11: YOLOv5m's ground truth for batch 0 of the Test set.

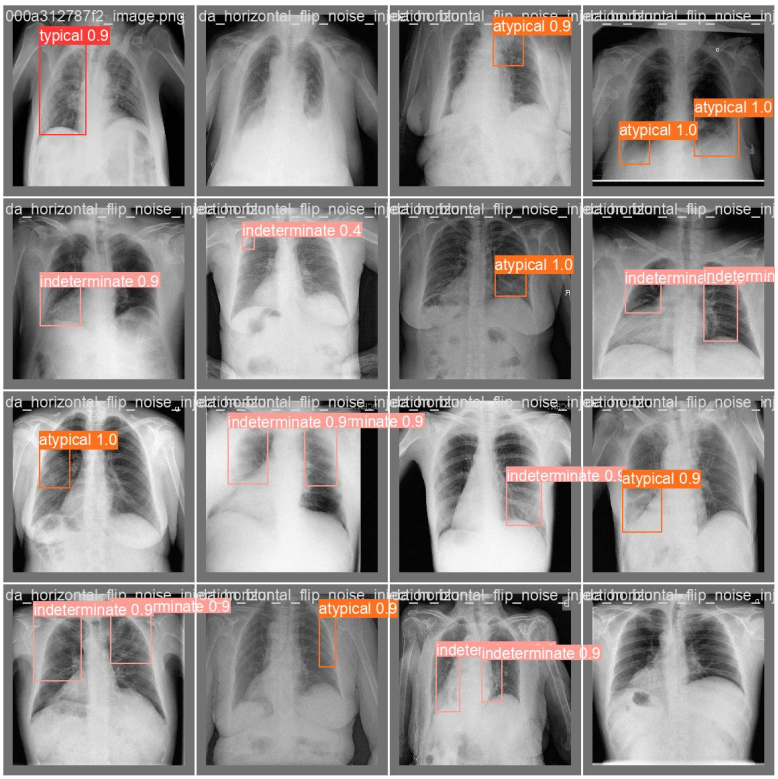

Figure 6.12: YOLOv5m's predictions for batch 0 of the Test set.

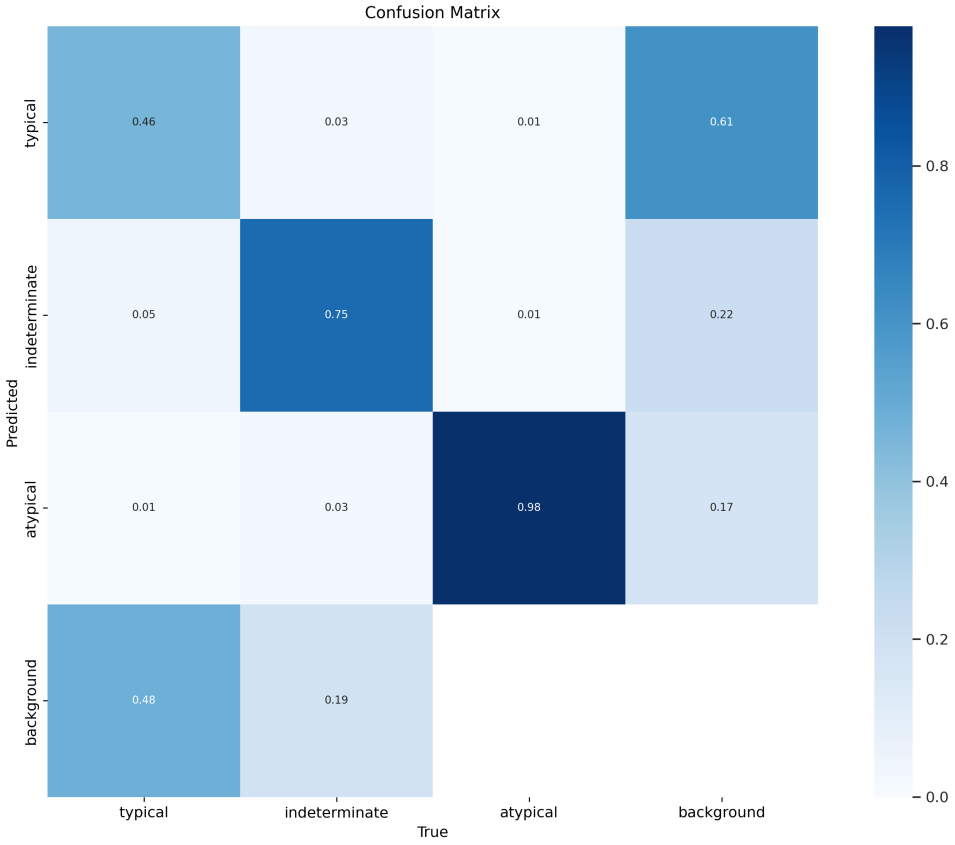

Figure 6.13: Ensemble learning confusion matrix for the Test set.

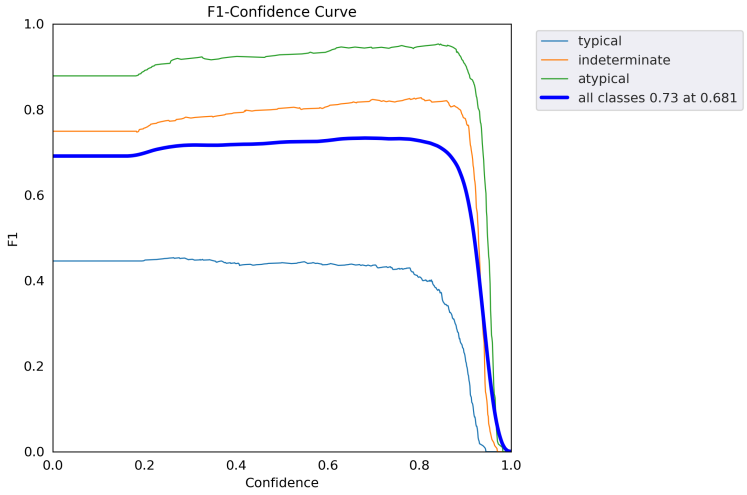

Figure 6.14: Ensemble learning F1-Confidence curve for the Test set.

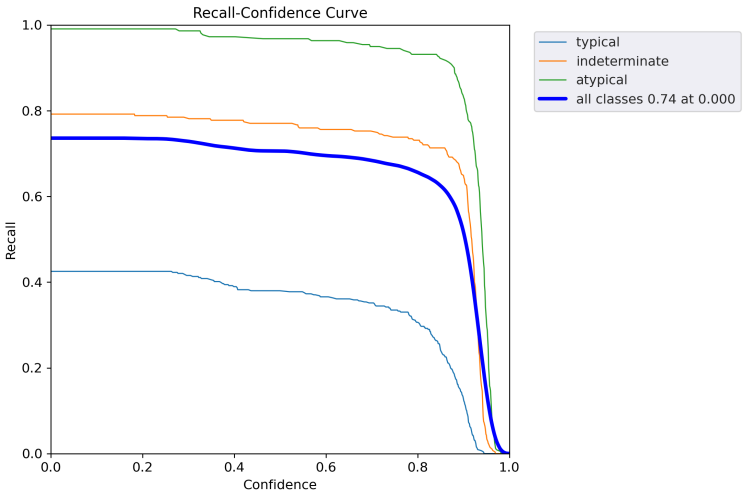

Figure 6.15: Ensemble learning Recall-Confidence for the Test set.

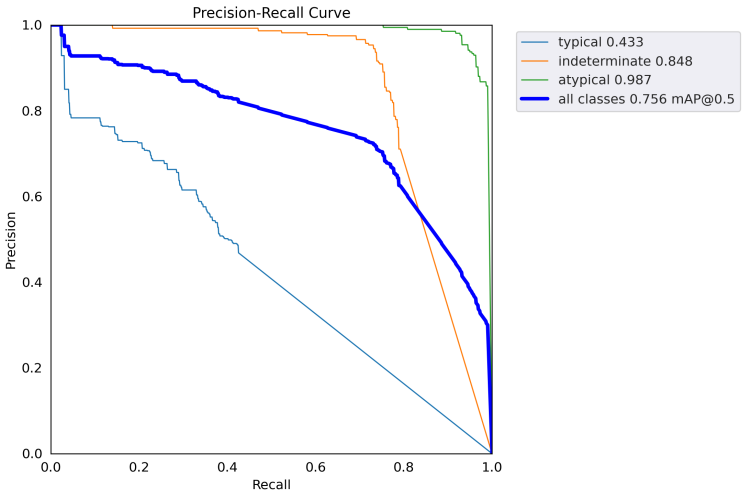

Figure 6.16: Ensemble learning Precision-Recall curve for the Test set.

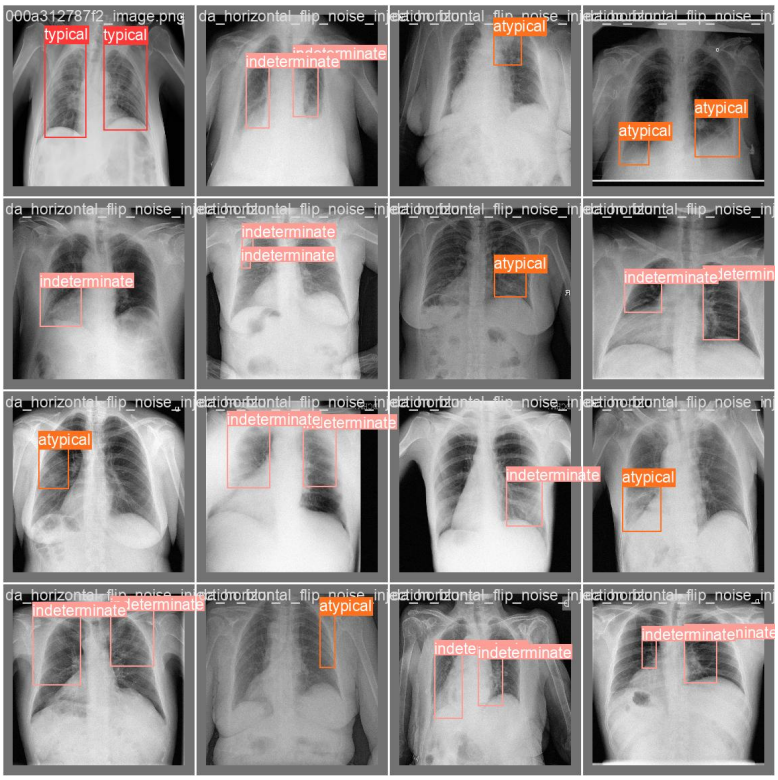

Figure 6.17: Ensemble learning ground truth for batch 0 of the Test set.

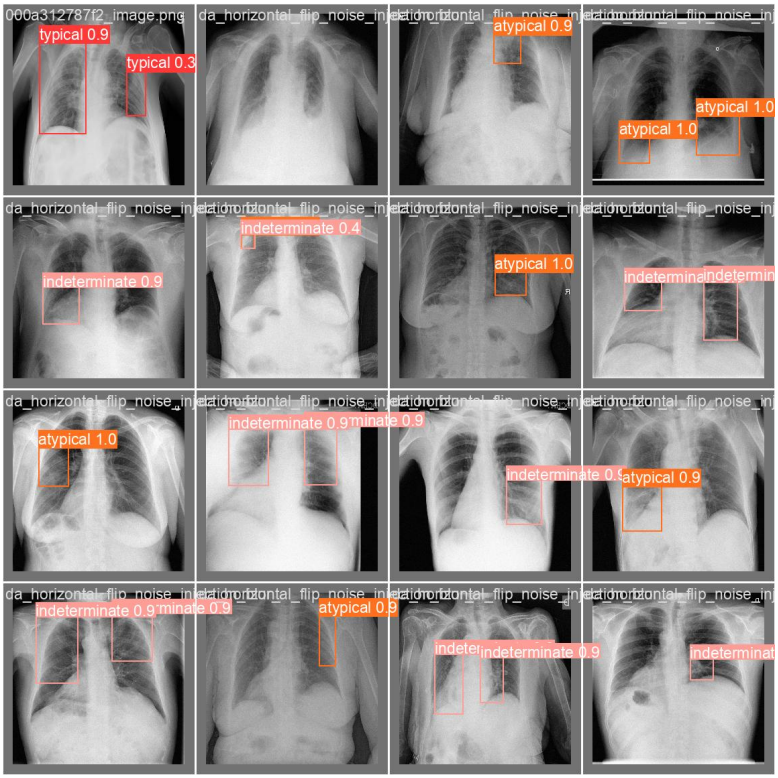

Figure 6.18: Ensemble learning predictions for batch 0 of the Test set.

Chapter 7

Future Work Areas

The following improvement areas have been identified:

- **Test Different Image Sizes:** Investigate the impact of varying image sizes on model performance. Experimenting with different input dimensions can provide valuable insights into the trade-offs between computational efficiency and detection accuracy.
- **Explore Other Frameworks:** Extend the investigation beyond YOLOv5 and explore the performance of other object detection frameworks. Comparing results across different architectures can offer a broader perspective on the suitability of various models for chest radiograph analysis.
- **Evaluate on New Unseen Data:** To obtain a more robust evaluation, experiments should be evaluated on entirely new and unseen datasets. This will help assess the generalization capability of the models and ensure their effectiveness in detecting objects in diverse and real-world scenarios.
- **Develop Evaluation Metrics for Negative Class:** Address the challenge of evaluating the *Negative for pneumonia* class by devising specialized metrics or methodologies. This will enable a more comprehensive assessment of the model's accuracy in predicting the absence of pneumonia, contributing to a more nuanced understanding of its performance.
- **Advanced Data Augmentation Techniques:** Explore more sophisticated data augmentation techniques to mitigate the risk of overfitting, particularly for augmented classes. Introducing new augmentation strategies can enhance the model's ability to generalize to unseen variations in the data while avoiding memorization of specific augmented instances.

Bibliography

[1] Muhammad Adeel Azam, Claudio Sampieri, Alessandro Ioppi, Stefano Africano, Alberto Vallin, Davide Mocellin, Marco Fragale, Luca Guastini, Sara Moccia, Cesare Piazza, Leonardo S. Mattos, and Giorgio Peretti. Deep learning applied to white light and narrow band imaging videolaryngoscopy: Toward real-time laryngeal cancer detection. *Scientific Figure on ResearchGate*, 2021. URL https://www.researchgate.net/figure/YOLOv5-architecture-representation-Color-figure-can-be-viewed-in-the-online-issue fig2_356562369. [Access date: 29/12/2023] Licensed under CC BY-NC-ND 4.0 Deed: https://creativecommons.org/licenses/by-nc-nd/4.0/.

[2] Ultralytics. Yolov5 manual. 2023. URL https://docs.ultralytics.com/yolov5/. [Access date: 29/12/2023].

[3] Andrew Kemp, Anna Zawacki, Chris Carr, George Shih, John Mongan, Julia Elliott, Kaiwen, Paras Lakhani, and Phil Culliton. Siim-fisabio-rsna covid-19 detection. 2021. URL https://kaggle.com/competitions/siim-covid19-detection. [Access date: 10/10/2023].

[4] S. Mirsadraee, M. Pourabrollah Toutkaboni, M. Bakhshayeshkaram, M. Rezaei, E. Askari, S. Haseli, and N. Sadraee. Radiological and laboratory findings of patients with covid-19 infection at the time of admission. 2020. URL https://ijp.iranpath.org/article_240036.html. [Access date: 10/10/2023].

[5] Gopal Singh Panwar. Top 8 image-processing python libraries used in machine learning. 2023. URL https://neptune.ai/blog/image-processing-python-libraries-for-machine-learning. [Access date: 10/10/2023].

[6] Deep learning frameworks. URL https://developer.nvidia.com/deep-learning-frameworks. [Access date: 10/10/2023].

[7] Mohamed Elgendy. Deep learning for vision systems. URL https:

//livebook.manning.com/book/deep-learning-for-vision-systems/deep-learning-for-vision-systems/10. [Access date: 10/10/2023].

[8] Mounika Narang. Top 10 python libraries for data visualization. 2023. URL https://www.knowledgehut.com/blog/business-intelligence-and-visualization/python-data-visualization-libraries#frequently-asked-questions. [Access date: 10/10/2023].

[9] Weights & Biases. Siimcovid19detection project by marta coll pol tracked with weights & biases. URL https://wandb.ai/siimcovid19detection/kaggle-siim-covid19?workspace=user-mcollpol. [Access date: 08/01/2024].

[10] ClearML. Clearml. build better ai at any scale, faster. URL https://clear.ml/. [Access date: 10/10/2023].

[11] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, Peihua Niu, Faxian Zhan, Xuejun Ma, Dayan Wang, Wenbo Xu, Guizhen Wu, George F. Gao, and Wenjie Tan. A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*, 382(8):727–733, 2020. doi: 10.1056/NEJMoa2001017. URL https://doi.org/10.1056/NEJMoa2001017. PMID: 31978945.

[12] Nanshan Chen, Min Zhou, Xuan Dong, and et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The Lancet*, 395(10223):507–513, 2020.

[13] Di Dong, Zhenchao Tang, Shuo Wang, Hui Hui, Lixin Gong, Yao Lu, Zhong Xue, Hongen Liao, Fang Chen, Fan Yang, Ronghua Jin, Kun Wang, Zhenyu Liu, Jingwei Wei, Wei Mu, Hui Zhang, Jingying Jiang, Jie Tian, and Hongjun Li. The role of imaging in the detection and management of covid-19: A review. *IEEE Reviews in Biomedical Engineering*, 14: 16–29, 2021. doi: 10.1109/RBME.2020.2990959.

[14] Diletta Cozzi, Marco Albanesi, Edoardo Cavigli, Chiara Moroni, Alessandra Bindi, Silvia Luvarà, Silvia Lucarini, Simone Busoni, Lorenzo Nicola Mazzoni, and Vittorio Miele. Chest x-ray in new coronavirus disease 2019 (covid-19) infection: findings and correlation with clinical outcome. *La Radiologia Medica*, 125(8):730–737, 2020. ISSN 1826-6983. doi: 10.1007/s11547-020-01232-9. URL https://doi.org/10.1007/s11547-020-01232-9.

[15] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of chest ct and rt-pcr testing for coronavirus

disease 2019 (covid-19) in china: A report of 1014 cases. *Radiology*, 296(2):E32–E40, 2020. doi: 10.1148/radiol.2020200642. URL https://doi.org/10.1148/radiol.2020200642. PMID: 32101510.

[16] G. D. Rubin and et al. The role of chest imaging in patient management during the covid-19 pandemic: A multinational consensus statement from the fleischner society. *Radiology*, page 201365, 2020. URL https://pubs.rsna.org/doi/full/10.1148/radiol.2020201365.

[17] Liqa A. Rousan, Eyhab Elobeid, Musaab Karrar, and Yousef Khader. Chest x-ray findings and temporal lung changes in patients with covid-19 pneumonia. *BMC Pulmonary Medicine*, 20(1):245, 2020. ISSN 1471-2466. doi: 10.1186/s12890-2020-01286-5. URL https://doi.org/10.1186/s12890-020-01286-5.

[18] Giovanni Volpicelli, Luna Gargani, Stefano Perlini, Stefano Spinelli, Greta Barbieri, Antonella Lanotte, Gonzalo García Casasola, Ramon Nogué-Bou, Alessandro Lamorte, Eustachio Agricola, Tomas Villén, Paramjeet Singh Deol, Peiman Nazerian, Francesco Corradi, Valerio Stefanone, Denise Nicole Fraga, Paolo Navalesi, Robinson Ferre, Enrico Boero, Giampaolo Martinelli, Lorenzo Cristoni, Cristiano Perani, Luigi Vetrugno, Cian McDermott, Francisco Miralles-Aguiar, Gianmarco Secco, Caterina Zattera, Francesco Salinaro, Alice Grignaschi, Andrea Boccatonda, Fabrizio Giostra, Marta Nogué Infante, Michele Covella, Giacomo Ingallina, Julia Burkert, Paolo Frumento, Francesco Forfori, Lorenzo Ghiadoni, Thomas Fraccalini, Alessandro Vendrame, Vittoria Basile, Alessandro Cipriano, Francesca Frassi, Massimo Santini, Marco Falcone, Francesco Menichetti, Bruno Barcella, Marzia Delorenzo, Flavia Resta, Giulia Vezzoni, Marco Bonzano, Domenica Federica Briganti, Giovanni Cappa, Ilaria Zunino, Lorenzo Demitry, Damiano Vignaroli, Lorenzo Scattaglia, Santi Di Pietro, Marco Bazzini, Vincenzo Capozza, María Mateos González, Rosa Vilella Gibal, Ramon Piñol Ibarz, Luis Martin Alfaro, Carlos Martin Alfaro, Maria Galindo Alins, Alice Brown, Hannah Dunlop, Maria Luisa Ralli, Paolo Persona, Frances M. Russel, Peter S. Pang, Serena Rovida, Cristian Deana, and Diego Franchini. Lung ultrasound for the early diagnosis of covid-19 pneumonia: an international multicenter study. *Intensive Care Medicine*, 47(4):444–454, 2021. ISSN 1432-1238. doi: 10.1007/s00134-2021-06373-7. URL https://doi.org/10.1007/s00134-021-06373-7.

[19] M. J. Smith, S. A. Hayward, S. M. Innes, and A. S. C. Miller. Point-of-care lung ultrasound in patients with covid-19 – a narrative review. *Anaesthesia*, 75:1096–1104, 2020. doi: 10.1111/anae.15082. URL https://doi.org/10.1111/anae.15082.

[20] Stanford University. Convolutional neural networks for visual recognition. 2023. URL http://cs231n.github.io/convolutional-networks/.

[21] S. Wang, Y. Zha, W. Li, and et al. A fully automatic deep learning system for covid-19 diagnostic and prognostic analysis. *Eur Respir J*, 56:2000775, 2020. doi: 10.1183/13993003.00775-2020. URL https://doi.org/10.1183/13993003.00775-2020.

[22] Dandi Yang, Cristhian Martinez, Lara Visuña, Hardev Khandhar, Chintan Bhatt, and Jesus Carretero. Detection and analysis of covid-19 in medical images using deep learning techniques. *Scientific Reports*, 11(1):19638, 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-99015-3. URL https://doi.org/10.1038/s41598-021-99015-3.

[23] Loveleen Gaur, Ujwal Bhatia, N. Z. Jhanjhi, Ghulam Muhammad, and Mehedi Masud. Medical image-based detection of covid-19 using deep convolution neural networks. *Multimedia Systems*, 29(3):1729–1738, 2023. ISSN 1432-1882. doi: 10.1007/s00530-021-00794-6. URL https://doi.org/10.1007/s00530-021-00794-6.

[24] Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, Marisa Caparrós, Germán González, and Jose María Salinas. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. 2020.

[25] Emily B. Tsai, Scott Simpson, Matthew P. Lungren, Michelle Hershman, Leonid Roshkovan, Errol Colak, Bradley J. Erickson, George Shih, Anouk Stein, Jayashree Kalpathy-Cramer, Jody Shen, Mona Hafez, Susan John, Prabhakar Rajiah, Brian P. Pogatchnik, John Mongan, Emre Altinmakas, Erik R. Ranschaert, Felipe C. Kitamura, Laurens Topff, Linda Moy, Jeffrey P. Kanne, and Carol C. Wu. The rsna international covid-19 open radiology database (ricord). *Radiology*, 2021. doi: 10.1148/radiol.2021203957.

[26] BIMCV-COVID19 Project. Bimcv-covid19 dataset research use agreement. URL https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/bimcv-covid19-dataset-research-use-agreement-2/. [Access date: 18/12/2023].

[27] The Cancer Imaging Archive. Tcia data usage policies and restrictions. URL https://www.cancerimagingarchive.net/data-usage-policies-and-restrictions/. [Access date: 18/12/2023].

[28] Creative Commons. Creative commons attribution-noncommercial 4.0 international license. URL https://creativecommons.org/licenses/by-nc/4.0/. [Access date: 18/12/2023].

[29] Paras Lakhani, John Mongan, Chirag Singhal, Qian Zhou, Kathy P Andriole, William F Auffermann, Prasanth Prasanna, Thanh Pham, Melissa Peterson, Peter J Bergquist, Tessa S Cook, Saulo F Ferraciolli, Guillermo C de Antonio Corradi, Marcelo Takahashi, Stephen S Workman, Meghal Parekh, Sherif Kamel, Jordan H Galant, Alberto Mas-Sanchez, Enrique Carrillo Benítez, Marta Sánchez-Valverde, Lineu Jaques, Maria Panadero, Montserrat Vidal, María Culiáñez-Casas, David M Angulo-Gonzalez, Steve G Langer, Maria de la Iglesia Vaya, and George Shih. The 2021 siim-fisabio-rsna machine learning covid-19 challenge: Annotation and standard exam classification of covid-19 chest radiographs. *OSF Preprints*, 2021. URL https://osf.io/532ek.

[30] Marta Coll. Eda - siim covid-19 detection. 2023. URL https://www.kaggle.com/code/martacoll/eda-siim-covid19-detection. [Access date: 29/12/2023].

[31] Innolitics LLC. Dicom standard - image pixel module. URL https://dicom.innolitics.com/ciods/rt-dose/image-pixel/00280004. [Access date: 18/12/2023].

[32] Marta Coll. Data preprocessing - siim covid-19 detection. 2023. URL https://www.kaggle.com/code/martacoll/data-preprocessing-siim-covid19-detection. [Access date: 29/12/2023].

[33] Manuel Cossio. Augmenting medical imaging: A comprehensive catalogue of 65 techniques for enhanced data analysis. March 2 2023. URL https://arxiv.org/pdf/2303.01178.pdf.

[34] Ultralytics. Yolov5: You only look once, version 5. 2023. URL https://github.com/ultralytics/yolov5. [Access date: 29/12/2023].

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

[36] Ultralytics. Ultralytics metrics documentation. 2023. URL https://docs.ultralytics.com/reference/utils/metrics/. [Access date: 29/12/2023].

[37] Ted Tigerschiold. What is accuracy, precision, recall, and f1 score? November 7 2022. URL https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score. [Access date: 29/12/2023].

[38] Marta Coll. Detection on siim covid-19 detection - yolov5. 2023. URL https://www.kaggle.com/code/martacoll/detection-on-siim-covid-19-detection-yolov5/notebook. [Access date: 29/12/2023].