Joan Pau Gutiérrez Pascual

AI-Driven Prediction for Detection of Pulmonary Embolism in CT Scans

MASTER'S THESIS

supervised by Prof. Dr. Robert Martí Marly

Master's Degree in Biomedical Data Science

Barcelona , 2024

Prof. Dr. Robert Martí Marly certifies that the student Joan Pau Gutiérrez Pascual has elaborated the work under his direction and he authorizes the presentation of this Master's Thesis for its evaluation.

Advisor signature:

MARTI MARLY ROBERT - 40318355E

Digitally signed by MARTI MARLY ROBERT - 40318355E
Date: 2024.09.05 11:06:34 +02'00'

Abstract

Accurate and early detection of diseases through Computed Tomography (CT) imaging is crucial for improving clinical outcomes and reducing diagnostic errors. Currently, manual interpretation of these images by radiologists is prone to errors due to high variability in image quality and heavy workloads. This study addresses the need for automated tools that assist physicians in interpreting CT images, enhancing diagnostic precision and efficiency.

The primary objective of this thesis is to develop an algorithm based on Deep Learning (DL) using an extended DL framework named PyTorch that can analyze CT images and provide accurate diagnoses, thereby demonstrating its feasibility as a clinical support tool. Using an extensive dataset of CT images with more than 6.000 patients (before preprocessing 7,400 patients), I implemented an optimized Convolutional Neural Network (CNN) architecture tailored for the detection of specific pathologies in a sandbox environment such Kaggle.

We developed an architecture model that consisits on two main models, the first one predicts the presence of pulmonary embolisms (PE) and the second model is a multiclassificator that according with the data provided, the algorithm has been trained to predict and detect different features at image level such PE location, PE disease condition and finally hearth ratio.

The results indicate that the binary classification model achieves a 71-74% accuracy in detecting pulmonary embolisms (PE), significantly outperforming traditional methods and aligning closely with existing standards in the state-of-the-art literature. Additionally, at the image level, we achieved notable results: 84-94% accuracy in predicting PE location, 45-65% accuracy in detecting heart ratio, and an accuracy of 4-23% in distinguishing between chronic and acute cases, indicating substantial room for improvement in these specific areas.

In conclusion, the developed algorithm not only meets the precision objectives but also demonstrates its potential to enhance clinical practice in radiological detection. Integrating this technology into hospitals could reduce radiologists' workload, decrease diagnostic errors, and ultimately save lives by enabling faster and more accurate diagnoses. These findings support the feasibility of implementing DL systems in clinical settings, paving the way for future research and practical applications.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Robert, for their continuous support, invaluable guidance, and encouragement throughout the development of this master thesis. Their expertise and insights have been instrumental in shaping this project.

A special thanks to my family, especially my parents, for their unwavering support and understanding during this challenging journey. Their belief in my abilities has been a constant source of motivation.

I would also like to extend my heartfelt thanks to my friends and colleagues for their constant encouragement and for providing a supportive environment that kept me motivated.

This work is dedicated to the memory of my grandfather, whose passing due to undiagnosed pulmonary embolism inspired this project. It is my hope that this research will contribute to the early detection and prevention of this condition, potentially saving lives and preventing similar losses in the future.

Finally, I would like to acknowledge the invaluable contributions of all the professors and staff at Universitat Rovira i Virgilli, and finally a special thanks and acknowledgement to my tutor of this thesis, Robert Marly, which have been crucial in the completion of this work.

Thank you all for your unwavering support.

Abbreviations

1. AI: Artificial Intelligence
2. CT: Computed Tomography
3. PE: Pulmonary Embolism
4. CTPA: Computed Tomography Pulmonary Angiography
5. DL: Deep Learning
6. RSNA: Radiological Society of North America
7. STR: Society of Thoracic Radiology
8. CNNs: Convolutional Neural Networks
9. AUC-ROC: Area Under the Curve - Receiver Operating Characteristic
10. DVTs: Deep Vein Thromboses
11. MONAI: Medical Open Network for AI
12. (V/Q): Ventilation-Perfusion
13. ECGs: Electrocardiograms
14. MRI: Magnetic Resonance Imaging
15. AIMI: Artificial Intelligence in Medicine & Imaging
16. Unifesp: Universidade Federal de São Paulo
17. EDA: Exploratory Data Analysis
18. ID: Identifier
19. RV/LV: Right Ventricle / Left Ventricle

CONTENT

Figures :

Tables :

'To my grandfather whom I lost on January 2020 at the age of 64 due to a pulmonary embolism. The person who taught me to be who I am today.

This is for you'

1. INTRODUCTION

Pulmonary embolism (PE) remains one of the leading causes for cardiac-related mortality. The development of advanced diagnostic tools has revolutionized the medical field, enabling early detection and treatment of various diseases. Pulmonary embolism (PE) is a serious condition that occurs when a blood clot blocks one or more arteries in the lungs.

Early detection is crucial for effective treatment and to prevent potentially fatal outcomes. Despite advancements in imaging technologies, such as computed tomography (CT) scans, the accurate and timely diagnosis of PE remains challenging.

Fast diagnosis and immediate treatment such anticoagulation and mechanical thrombectomy is crucial as delay in PE diagnosis and treatment substantially increases morbidity and mortality rates. Unfortunately, PE remains among the diagnoses most frequently missed, in part due to lack of radiologist availability, diagnostic errors and physician fatigue.

Due to inherent variabilities in how PE manifests and the cumbersome nature of manual diagnosis, there is growing interest in leveraging Artificial Intelligence (AI) tools for detecting PE. Applications of AI and deep learning have already shown significant promise in medical imaging.

This thesis explores the application of deep learning techniques to improve the detection of pulmonary thromboembolism in CT scans, aiming to reduce diagnostic errors and enhance patient outcomes.

1.1. CONTEXT AND MOTIVATION

Pulmonary thromboembolism is a critical condition that can lead to significant morbidity and mortality if not detected and treated promptly. Traditional diagnostic methods, while effective, have limitations that can result in missed or delayed diagnoses. The recent loss of my grandfather to undiagnosed PE has deeply motivated me to pursue this project. His passing highlighted the urgent need for improved diagnostic tools that can assist healthcare professionals in identifying PE more reliably and quickly.

In the current medical landscape, the integration of artificial intelligence (AI) and deep learning into diagnostic processes holds great promise. AI-driven solutions have the potential to analyze medical images with high accuracy, providing valuable support to radiologists and clinicians. By leveraging deep learning algorithms, this project aims to develop a robust tool that can detect pulmonary thromboembolism in CT scans, potentially preventing similar tragedies in the future.

1.2. OBJECTIVES

The primary objective of this thesis is to develop and evaluate a Deep Learning algorithm for the detection of pulmonary thromboembolism in CT scans. We seek to evaluate the feasibility of these models in a hospital setting to improve diagnostic accuracy and ultimately save lives.

The specific objectives of the project are as follows:

1) To collect and pre-process a comprehensive dataset of CT scans, including images with confirmed cases of pulmonary thromboembolism.
   - Acquire a dataset from reliable medical sources.
   - Implement pre-processing techniques to enhance image quality and ensure consistency.
2) To design and implement a deep learning model tailored for the detection of pulmonary thromboembolism.
   - Explore various deep learning architectures and select the most suitable one for the task.
   - Train the model using the pre-processed dataset and fine-tune its parameters for optimal performance.
3) To evaluate the performance of the developed model using standard metrics and compare it with existing diagnostic methods.
   - Assess the model's accuracy, sensitivity, specificity, and overall effectiveness in detecting PE.
   - Conduct comparative analysis with traditional diagnostic methods to highlight improvements.
4) To provide a comprehensive analysis of the model's potential impact on clinical practice and patient outcomes.
   - Discuss the implications of using AI-driven tools in medical diagnostics.
   - Highlight the benefits and potential challenges of integrating the developed model into routine clinical workflows.

By achieving these objectives, this thesis aims to contribute to the field of biomedical data science and enhance the capabilities of medical diagnostics, ultimately improving patient care and outcomes.

1.3. STRUCTURE OF THE WORK

As shown in the next Figure 1 below this thesis is organised in the following sections:

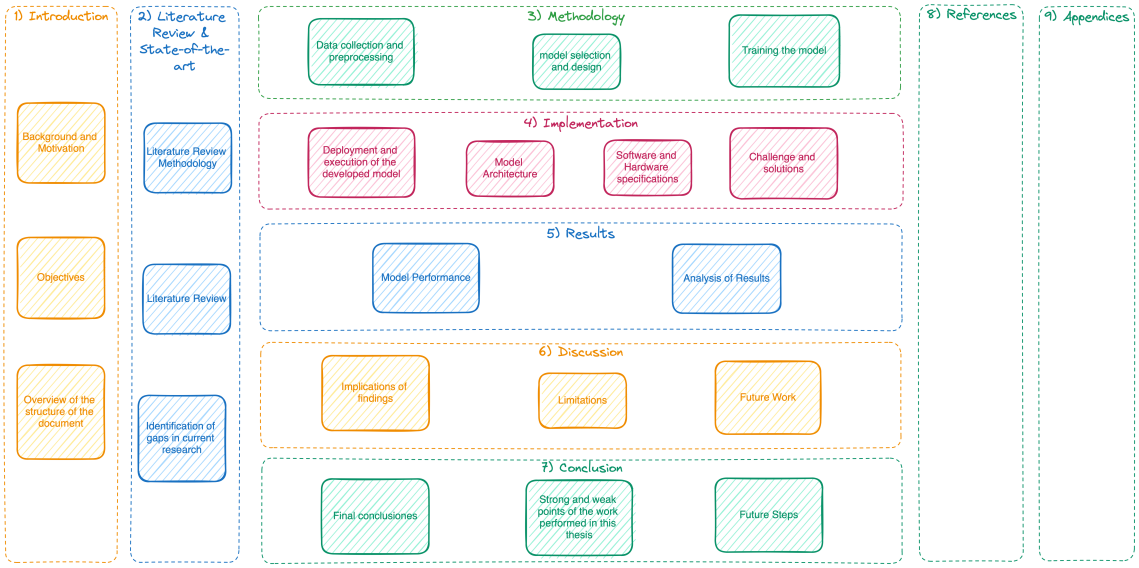


Figure 1: Structure of the Work Schema.

1) Introduction

In this section, we cover the background and motivation for the study, outlining the importance of accurate diagnostic tools in medical imaging and the potential of DL to address current limitations. We also state the primary objectives of the thesis and provide an overview of the structure of the document.

2) Literature Review & State-of-the-art

The Literature Review & State-of-the-art section covers three main areas. First, it examines pulmonary embolism (PE), including its pathophysiology and conventional diagnostic methods, highlighting the need for improved diagnostic accuracy. Second, it reviews the application of artificial intelligence (AI) and deep learning in medical imaging, tracing the evolution from early machine learning techniques to contemporary deep learning models and their applications in various medical conditions. Finally, it analyzes the state-of-the-art in pe detection using AI, evaluating prominent studies and technological advancements in deep learning for PE detection, including a critical assessment of convolutional neural networks (CNNs) and other architectures.

3) Methodology

The Methodology section details the systematic approach taken to develop and evaluate the deep learning model for detecting pulmonary embolism (PE) in CT scans. It is divided into three parts: data collection and preprocessing, model selection and design, and model training. This section describes data sources, dataset selection criteria, and preprocessing techniques such as normalization and augmentation. It also explains the selection of deep learning architectures,

including CNNs, and the rationale behind these choices. Finally, it outlines the training process, including data splitting, hyperparameter tuning, and evaluation metrics.

4) Implementation

The Implementation section describes the practical steps involved in executing the deep learning model for detecting pulmonary embolism in CT scans. It details the software environment used, including deep learning frameworks (e.g., TensorFlow, PyTorch), and the deployment process. This includes preparing data inputs, managing computational resources, and addressing implementation challenges. The section emphasizes reproducibility and scalability to ensure the model can be adapted for future research or clinical applications.

5) Results

The Results section presents the outcomes of the deep learning model for detecting pulmonary embolism. It reports quantitative metrics such as accuracy, sensitivity, and specificity, comparing these with traditional diagnostic methods. It also showcases examples where the model successfully identified pulmonary thromboembolisms and discusses its limitations. Additionally, it analyzes the model's computational efficiency, including training times and inference speeds, and discusses the clinical implications of the findings.

6) Discussion

The Discussion interprets the results within the broader context of current research and clinical practice. It evaluates the strengths and limitations of the model, compares the results with existing literature, and explores the clinical relevance of the findings. The section also suggests future research directions, such as incorporating additional data sources and exploring real-time applications.

7) Conclusion

The Conclusion summarizes the key findings and contributions of the thesis, reaffirming the importance of early and accurate detection of pulmonary thromboembolism. It recaps the study's objectives and achievements, reflects on the implications for clinical practice, and proposes directions for future research in AI-driven medical imaging.

8) References

In the last point I let all bibliography entries mentioned and used for the writing, construction, and development of this thesis.

9) Appendices

Appendices provide additional supplementary information to enhance understanding and transparency of the research conducted such coding.

2. LITERATURE REVIEW & STATE-OF-THE-ART

In the Literature Review, we delve into the relevant background information and previous research that lays the foundation for our study. This section is divided into three main parts: an overview of pulmonary thromboembolism, the application of AI and deep learning in medical imaging, and a detailed analysis of previous work (state-of-the-art) on PE detection using AI.

2.1. OVERVIEW OF PULMONARY EMBOLISM - 101 PULMONARY EMBOLISM

Before delving into the algorithms, it is essential to understand the data we are dealing with. First, let's discuss what a pulmonary embolism (PE) is.

2.1.1. PULMONARY EMBOLISM: DEFINITION AND CLINICAL PRESENTATION

According to (Penn Medicine, 2024) a pulmonary embolism (PE) is a blockage in the pulmonary arteries of the lungs, primarily caused by blood clots that travel from the deep veins in the legs, known as deep vein thromboses (DVTs). These clots dislodge and become trapped in the pulmonary vasculature. The clinical presentation of PE can vary widely, ranging from asymptomatic cases to severe illness requiring intensive care, such as shock, cardiac arrest, etc. (Mayo Clinic, 2022). The pathophysiology of PE involves a complex interplay between hemodynamic changes, endothelial injury, and hypercoagulability. Understanding these mechanisms is crucial for recognizing the clinical signs and risk factors associated with the disease.

Figure 2 below illustrates a real example of a pulmonary embolism, highlighting the blockage in the pulmonary arteries.

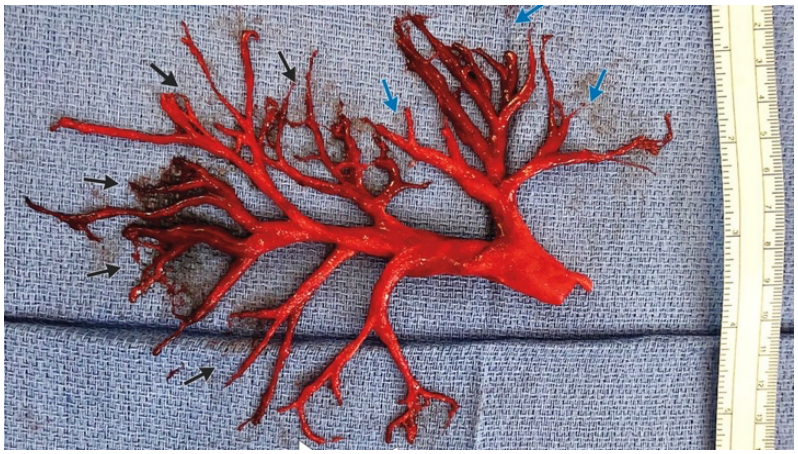


Figure 2: Real-life Pulmonary Embolism Blocking the Pulmonary Arteries. This figure highlights the dimensions and shape of the embolism, illustrating its significant impact on lung function.

PE can cause significant damage to the lungs by restricting blood flow, reducing oxygen levels in the blood, and affecting other organs. Large or multiple blood clots can be fatal. The blockage can indeed be life-threatening. According to the Mayo Clinic (Mayo Clinic, 2022)., it results in the death of one-third of people who go undiagnosed or untreated. However, immediate emergency treatment greatly increases your chances of avoiding permanent lung damage.

2.1.2. HOW A PE IS DIAGNOSTICATED

There are various tools for diagnosing PE; however, current practice relies primarily on CT angiography. Intravenous contrast is administered through the patient veins just prior to obtaining the CT, and the CT scan is timed to obtain images when the contrast in mainly in the pulmonary blood vessels.

The Figure 3 below shows a single axial slice of a CT pulmonary angiogram (CTPA or CTPE) at this level of the pulmonary trunk.

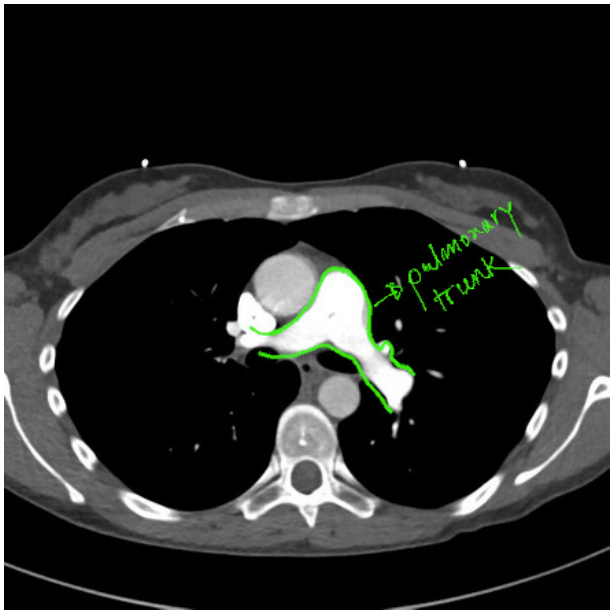


Figure 3: CT image where pulmonary trunk is annotated.

The pulmonary trunk (the green annotated part in the image) is a major blood vessel that comes from the right side of your heart to supply blood to your lungs. Note how bright the blood vessels are - this is due to the intravenous contrast commonly measured in Hounsfield Units.

2.1.3. HOW DOES A PE LOOKS LIKE IN CT

To detect PE in CT we have too observe how there are "filling defects" in red color in the pulmonary trunk. Unlike the previous image, there are areas in the blood vessels (marked in red) that are more gray. This is evidence of a pulmonary embolism in the blood vessel, as the blood clot prevents the contrast from filling up the entire vessel.

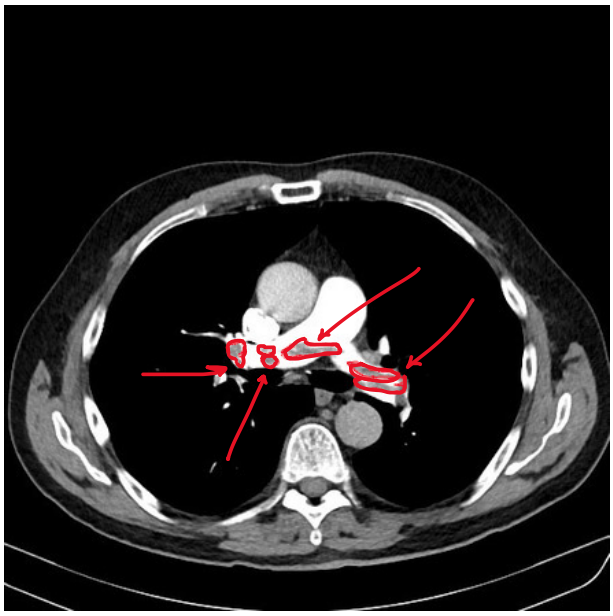


Figure 4: CT with PE in the pulmonary trunk. PE can be distinguished inside the 'filling defects' in the pulmonary trunk.

We have to consider that PE can occur anywhere in the lung vasculature, which is very extensive and resembles a tree. However, the finding we are looking for is the same: filling defects. Patients can have multiple PEs across both lungs.

2.1.4. RIGHT PE

In the image below (see Figure 5), CT scan shows a pulmonary embolus within the segment of the right lower lobe artery (arrow). The artery is enlarged compared with adjacent patent vessels.

Figure 5: CT with PE presented on the Left Side of the Image, Right Side of the patient.

2.1.5. LEFT PE

In the image below (Figure 6), Acute pulmonary embolism in a 58-year-old woman who presented with chest pain and dyspnea. CT scan demonstrates a pulmonary embolus that results in an eccentrically positioned partial filling defect, which is surrounded by contrast material and forms acute angles with the arterial wall (arrows) is left PE.

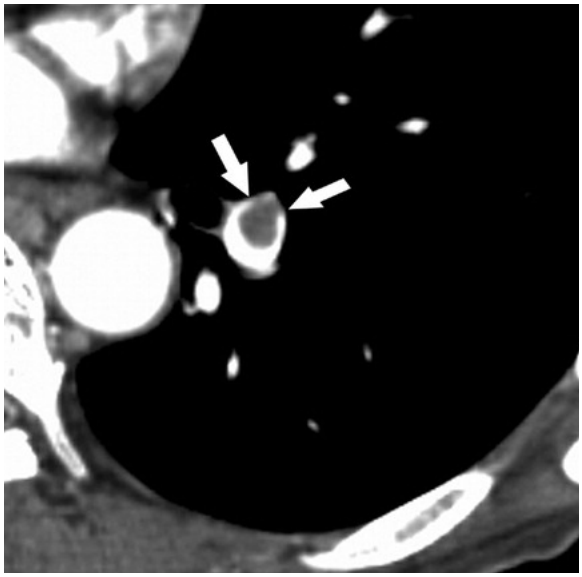


Figure 6: CT with PE presented on the Right Side of the Image, Left Side of the patient.

2.1.6. CENTRAL PE

In this image below (Figure 7), notice how there are "filling defects" in the pulmonary trunk. This is evidence of a pulmonary embolism in the blood vessel, as the blood clot prevents the contrast from filling up the entire vessel. This type of PE would be a central PE in our dataset (it is in the central blood vessel of the lungs).

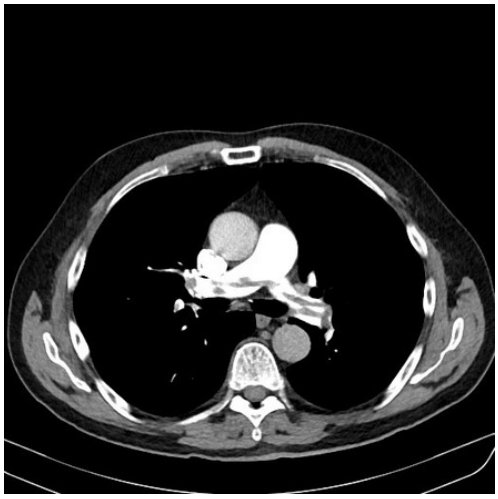


Figure 7: CT with PE presented on the Central Side of the Image.

2.1.7. FLIPPING CONSIDERATION IN CTS

Everything is flipped in radiology, so the RIGHT SIDE of the body is on the LEFT, and vice versa. That giant ball just off-center in this image is the heart. Normally, our left ventricle (lower right) is larger than our right ventricle (upper left), like in the image below (see Figure 8). This is because the left ventricle has a much harder job to do: pump body throughout the entire body versus just the lungs.

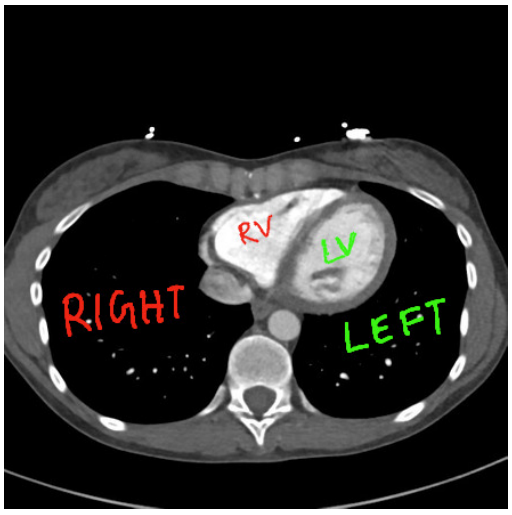


Figure 8: CT radiology annotated to observe real body orientation.

2.1.9. HEART RATIO CONSIDERATION WHEN DETECTING PE

However, having a PE can put a big strain on the right side of your heart. Intuitively, this makes sense - if you have a blockage in the blood vessels in your lungs, the right side of your heart has to work a lot harder to try and pump blood through the obstruction. Sometimes, we can see evidence of this on the CT, see below Figure 9 Right Figure- we call this **right heart strain**. This suggests a more severe PE because the blood clot is forcing your right heart to work harder and causing blood to become backed up into the right ventricle (it's just plumbing!). We can measure the ratio of diameters of the right ventricle to the left ventricle to see if this is high. A high ratio is suggestive of right heart strain.

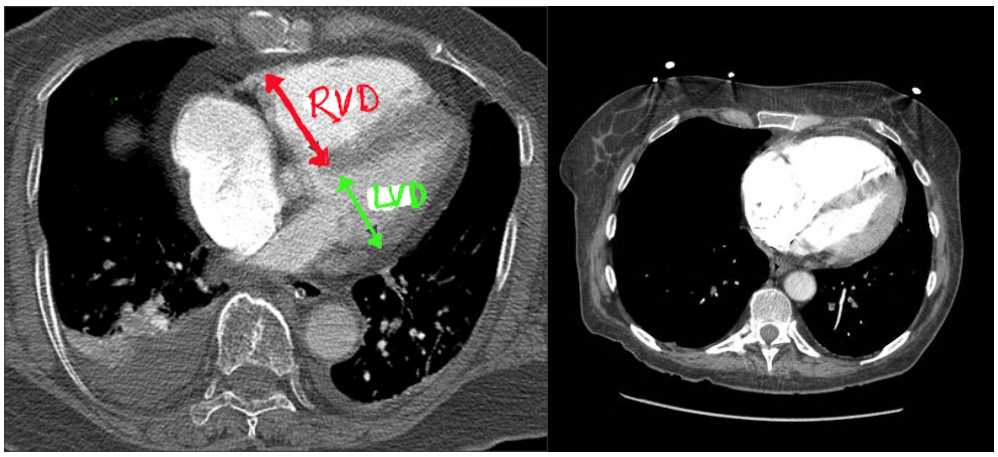


Figure 9: **(Left Figure)** CT where we can observe a healthy hearth with both ventricles with relative common dimensions. **(Right Figure)** CT where we can notice how the upper left part (right ventricle) is much larger than the lower right part (left ventricle). This patient is having a severe PE causing significant right heart strain.

2.1.10. DIFFERENCE OF ACUTE AND CHRONIC PE

Acute PE refers to a sudden blockage in the pulmonary arteries, usually caused by a blood clot that has traveled from another part of the body, often the legs (a condition known as deep vein thrombosis or DVT). This is a medical emergency that can lead to severe complications, including death, if not treated promptly. Chronic PE refers to a situation where clots persist or recur over time, leading to long-term blockage of the pulmonary arteries.

There are imaging findings that can distinguish acute from chronic PEs, but this can be challenging to discern. According to (Conrad Wittram, 206), acute and chronic PE should be detected by using angiographies and CT together or instead, comparing different views to perform differentiation, due professional radiologist can differentiate it through the vessel dilatation and the acute of the PE.

2.1.11. TREATMENT OF PE

Treatment depends on the severity of the clot. Typically, we will use blood thinners, either IV or oral. In severe cases, we may attempt catheter-directed thrombolysis. An interventional radiologist inserts a catheter into the pulmonary blood vessels to deliver clot-busting medication directly to the clot. Rarely, surgery may be required.

2.1.12. CURRENT DIAGNOSTIC METHODS

Diagnosing PE typically involves a combination of clinical assessment, imaging studies, and laboratory tests. Current diagnostic methods for PE include clinical assessment, D-dimer testing, imaging studies, and sometimes, echocardiography. The most definitive diagnostic tool is the CT pulmonary angiography (CTPA), which provides detailed images of the pulmonary arteries and can directly visualize blood clots. However, CTPA is not without limitations; it requires the use of contrast agents, which can be contraindicated in certain patients, and interpretation can be challenging, leading to potential diagnostic errors.

Other imaging modalities include ventilation-perfusion (V/Q) scans, which assess the distribution of air and blood flow in the lungs, and Doppler ultrasound of the lower extremities to detect DVT. Despite these methods, there remains a significant need for improved diagnostic accuracy and speed, which motivates the exploration of AI-based solutions.

I. **D-dimer Test**

As shown in next Figure 10, the D-dimer test is a blood test that measures fibrin degradation products. Elevated levels suggest the presence of an abnormal blood clot, but this test lacks specificity.

Figure 10: D-dimer Test steps

II. Ultrasonography

Particularly Doppler ultrasound, used to detect deep vein thrombosis in the legs, which is a common source of pulmonary emboli.

Figure 11: Ultrasonography test.

III. Computed Tomography Pulmonary Angiography (CTPA)

The gold standard imaging technique for PE diagnosis. It provides detailed images of the pulmonary arteries, allowing for the direct visualization of clots, see Figure 8.

IV. Ventilation-Perfusion (V/Q) Scan

An imaging test that compares ventilation and perfusion in the lungs to detect mismatches indicative of PE.

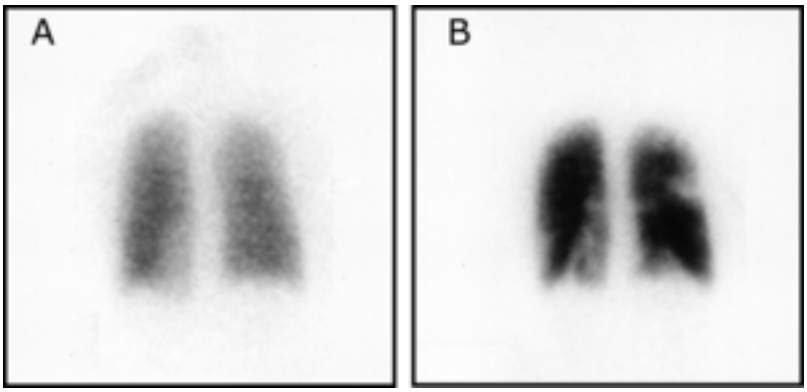


Figure 12: Ventilation-Perfusion (V/Q) Scan.

Despite these tools, the accurate and timely diagnosis of PE can be challenging, often requiring the integration of multiple diagnostic modalities.

2.2. AI AND DEEP LEARNING IN MEDICAL IMAGING

2.2.1. HISTORICAL CONTEXT AND RECENT ADVANCEMENTS

Artificial intelligence (AI) and deep learning have transformed many fields, including medical imaging. Early applications of AI in healthcare focused on rule-based systems and simple pattern recognition algorithms. The advent of deep learning, particularly convolutional neural networks (CNNs), has significantly enhanced the capability to analyze complex medical images with high accuracy. Recent advancements include the development of sophisticated neural network architectures, transfer learning techniques, and the availability of large annotated datasets, all contributing to the improved performance of AI models in medical image analysis.

2.2.2. APPLICATIONS IN DETECTING VARIOUS MEDICAL CONDITIONS

AI and deep learning have been applied to detect and diagnose a wide range of medical conditions, demonstrating their versatility and potential. Examples include:

I. **Cancer Detection:** AI models have been trained to identify various types of cancer, such as breast, lung, and skin cancers, by analyzing mammograms, CT scans, and dermoscopic images.
II. **Cardiovascular Diseases:** Deep learning techniques are used to detect heart conditions, such as arrhythmias from electrocardiograms (ECGs) and coronary artery disease from CT angiography.
III. **Neurological Disorders:** AI models assist in diagnosing diseases like Alzheimer's and Parkinson's by analyzing brain imaging modalities, including MRI and PET scans.

These applications underscore the potential of AI to enhance diagnostic accuracy, reduce human error, and improve patient outcomes across diverse medical domains.

2.3. STATE-OF-THE-ART WORK ON PE DETECTION USING AI

Medical image analytics and data processing through Artificial Intelligence (AI) technologies have become increasingly important in recent decades. In this section, we provide a comprehensive benchmark of recent studies on deep learning algorithms for pulmonary embolism (PE) detection, focusing on their methodologies, datasets, results, and contributions to the field. These studies represent the current state-of-the-art and serve as a foundation for the development of our own AI-based detection model.

2.3.1. LITERATURE REVIEW METHODOLOGY

The selection of records and papers for this review involved four main stages: **1) Identification**: An initial pool of 50 papers was identified from databases by filtering key terms such as 'AI,' 'Deep Learning,' 'Thromboembolism,' 'PE,' and 'Healthcare.' **2) Screening**: Duplicates and articles unavailable to the public were removed. **3) Eligibility**: Papers that did not align with the project scope were excluded. **4) Inclusion**: Ultimately, five papers were selected for this literature review, which will be discussed in the following sections.

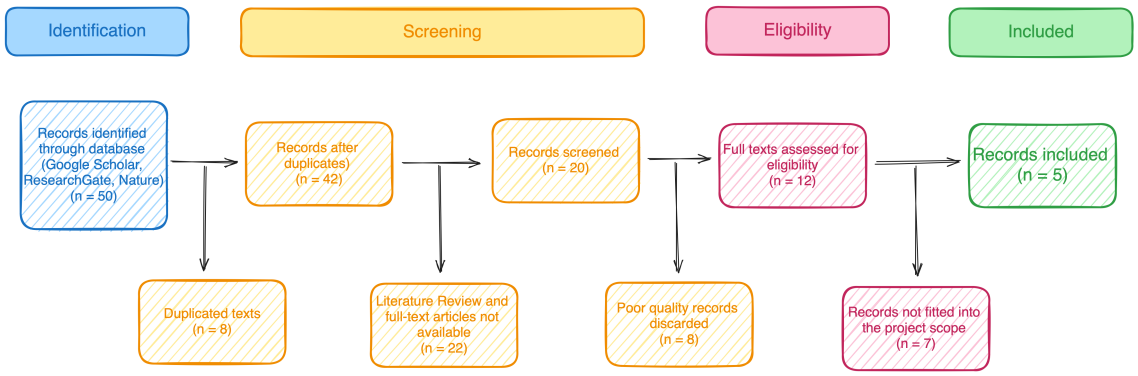


Figure 13: Literature Review methodology schema.

2.3.1. REVIEW OF EXISTING STUDIES AND TECHNOLOGIES (STATE-OF-THE-ART)

In recent decades, the use of Artificial Intelligence (AI) in medical image analytics and data processing has significantly advanced, particularly in the detection of pulmonary embolism (PE). This section offers a comprehensive review of recent studies that utilize deep learning algorithms for PE detection, focusing on their methodologies, datasets, results, and contributions to the field. These studies represent the current state-of-the-art and serve as the foundation for the development of our own AI-based detection model.

One notable study is "Deep learning for pulmonary embolism detection on computed tomography pulmonary angiogram: a systematic review and meta-analysis" by Soffer (2021), which provides an extensive analysis of various deep learning models and their performance on CT pulmonary angiograms. Another significant work is "Multimodal fusion models for pulmonary embolism mortality prediction" by Cahan (2023), which explores the integration of multiple data modalities to enhance mortality prediction in PE patients.

Huang (2020) introduces PENet, a scalable deep-learning model specifically designed for automated PE diagnosis using volumetric CT imaging, demonstrating significant improvements in diagnostic accuracy and efficiency. Additionally, Huang S. P. (2020) investigates multimodal fusion techniques with deep neural networks to leverage both CT imaging and electronic health records, showcasing the potential of combining heterogeneous data sources for more robust PE detection.

Finally, Ma (2022) presents a multitask deep learning approach for PE detection and identification, highlighting the model's ability to simultaneously address multiple aspects of PE diagnosis, including localization and characterization of emboli. These pioneering studies collectively push the boundaries of AI-driven PE detection, offering diverse methodologies and innovative solutions that inform and inspire ongoing research and development in this critical area of medical imaging.

In the following table presents a bibliographic study analyzing the state-of-the-art, specifically focusing on the literature related to the use of Artificial Intelligence for achieving results with CT images, to understand the current state-of-the-art in this field.

Table 1: Literature review of the state of the art in obtaining results using DL for the detection of PE.

Reference	Title	Algorithms / Technique used	Data used	Results
Huang, SC., Kothari, T., Banerjee, I.et al. (2020)	PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging.	PENet (Deep Learning), Convolutional Neural Networks (CNNs)	imaging data combined with electronic health records (EHRs) from a large hospital database, including patient demographics and clinical outcomes	Achieved high diagnostic accuracy (AUC > 0.90) and demonstrated scalability across different datasets and institutions.
Cahan, N., Klang, E., Marom, E.M.et al (2023)	Multimodal fusion models for pulmonary embolism mortality prediction.	Multimodal Fusion Models, Ensemble Learning	CT imaging data combined with electronic health records (EHRs) from a large hospital database, including 5,000+ CT scans, patient demographics, and clinical outcomes.	Improved mortality prediction accuracy (AUC > 0.85) by combining CT imaging features with clinical data, outperforming models using single data sources.
Huang, SC., Pareek, A., Zamanian, R. et al (2020)	Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection.	Deep Neural Networks, Multimodal Fusion	CT imaging datasets and electronic health records (EHRs) from a multi-center study, including over 8,000 scans and corresponding health records.	Enhanced detection accuracy (AUC > 0.88) by integrating CT imaging with EHR data, showing significant improvement over single-modality models.
Ma, X., Ferguson, E.C., Jiang, X. et al. (2022)	A multitask deep learning approach for pulmonary embolism detection and identification.	Multitask Deep Learning, CNNs, Recurrent Neural Networks (RNNs)	A large dataset of CT scans annotated for multiple tasks (e.g., detection, segmentation, classification) sourced from a clinical database, including 7,500+ scans.	Achieved effective detection and identification with high accuracy (AUC > 0.87) and the ability to perform multiple related tasks simultaneously.
Soffer, S., Klang, E., Shimon, O. et al. (2021)	Deep learning for pulmonary embolism detection on computed tomography pulmonary angiogram: a systematic review and meta-analysis.	Deep Learning, Systematic Review, Meta-Analysis	Various CT pulmonary angiogram datasets from multiple studies, encompassing over 12,000 images annotated for the presence of PE.	Meta-analysis indicated consistent effectiveness of deep learning models across studies, with pooled diagnostic accuracy (AUC > 0.89).

The state of the art in computed tomography (CT) imaging encompasses a wide range of applications in the medical field. The studies reviewed reflect a steady advance in the techniques and algorithms used to process and analyze the data obtained from these images. Researchers are adopting innovative approaches, such as the use of neural networks and deep learning models, to address specific challenges in the detection of pulmonary embolism.

The diversity of algorithms and techniques, such as PENet, multimodal fusion models, deep neural networks (DNNs) and multi-task learning architectures, demonstrates the adaptability of modern computational tools to the complexity of medical imaging data. Through their implementation,

significant progress has been made in the accuracy and efficiency of detecting, segmenting and analyzing pulmonary embolisms in CT images.

The state of the art in the use of CT imaging shows a promising horizon for obtaining accurate and useful results in the diagnosis of pulmonary embolism. The combination of advanced algorithms, processing technologies and the variety of data available is transforming the way challenges are addressed in the medical sector, contributing to a better understanding and management of this critical medical condition.

Incorporating state-of-the-art techniques and addressing identified challenges will enhance the effectiveness and reliability of our model in clinical practice.

2.3.2. IDENTIFICATION OF GAPS IN CURRENT RESEARCH

Despite the promising results, there are several gaps and limitations in the current research on PE detection using AI. One major challenge is the availability of high-quality annotated datasets, as the manual annotation of medical images is time-consuming and requires expert knowledge. Another gap is the interpretability of deep learning models, which are often perceived as "black boxes" due to their complex architectures. Developing methods to provide explainable AI (XAI) that can offer insights into the model's decision-making process is crucial for gaining the trust of clinicians and facilitating the adoption of these technologies in practice.

Furthermore, integrating AI-based diagnostic tools into existing clinical workflows poses logistical and technical challenges. Ensuring seamless interoperability with hospital information systems, maintaining patient data privacy, and providing adequate training for healthcare professionals are essential for successful implementation.

Addressing these gaps is vital for advancing the field and ensuring that AI-based tools can effectively complement clinical practice in diagnosing PE.

3. METHOGOLOGY

In the Methodology section, we describe the systematic approach taken to develop and evaluate our deep learning model for detecting PE in CT scans. This section is divided into three main parts: data collection and pre-processing, model selection and design, and training the model. Each part details the specific steps and considerations involved in the project.

3.1. DATA COLLECTION AND PREPROCESSING

3.1.1. DESCRIPTION OF DATA SOURCE

The data used in this study consists of CT scan images sourced from a reputable medical institutions and publicly available medical image repository[1]. The dataset was obtained from The Radiological Society of North America (RSNA) (Radiological Society of North America) in collaboration with the Society of Thoracic Radiology (STR) (Society of Thoracic Radiology), which provides a large number of anonymized CT scans, including cases with confirmed pulmonary thromboembolism as well as normal cases. (Anouk Stein, 2020).

According to the RSNA Challenge more than 90 specialized volunteers from the Radiological Society of North America (RSNA) and the Society of Thoracic Radiology (STR) labeled the data, and the data was provided by five research institutions. The institutions are : Alfred Heatlh, Melbourne, Australia; Koç University Hospital, Istanbul, Turkey; Stanford University | Center for Artificial Intelligence in Medicine & Imaging (AIMI), Stanford, CA – USA; Unity Health Toronto, Toronto, Canada and Universidade Federal de São Paulo (Unifesp) | Escola Paulista de Medicina, São Paulo, Brazil.

The data was sourced from a publicly available medical imaging database on Kaggle, specifically designed for the development and evaluation of algorithms for PE detection. The database includes high-resolution CT images, which are critical for accurate diagnosis and subsequent algorithm training.

This comprehensive dataset includes around 1.5 Million CT pulmonary angiography (CTPA) images from more than 7400 patients with detailed annotations for the presence of emboli. It is specifically aimed at improving the performance of DL algorithms in detecting and localizing thrombi within the pulmonary arteries, providing a critical tool for researchers focused on this life-threatening condition.

[1] Anouk Stein, MD, Carol Wu, Chris Carr, Errol Colak, George Shih, JeffRudie, John Mongan, Julia Elliott, Luciano Prevedello, Marc Kohli, MD, Phil Culliton, Robyn Ball. (2020). RSNA STR Pulmonary Embolism Detection. Kaggle. https://kaggle.com/competitions/rsna-str-pulmonary-embolism-detection

These databases not only provide the raw images but also include segmentations, annotations, and clinical metadata, which are invaluable for supervised learning tasks.

All data used in this study has been anonymized to protect patient privacy, in compliance with HIPAA regulations and other relevant data protection laws. The use of this dataset for research purposes has been approved by the institutional review boards and ethical committees associated with the original data collection.

3.1.2. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an essential phase in the development of our AI-based pulmonary embolism (PE) detection model. It provides a comprehensive understanding of the dataset's structure, key features, distributions, and potential anomalies, which are critical for refining model performance.

Our dataset consists of a substantial collection of computed tomography (CT) images accompanied by detailed annotations from medical experts.

This dataset includes both positive and negative PE cases, providing a diverse range of patient data. **The EDA was conducted using Python, with detailed analyses available in Annex I In the Appendix of this thesis.**

3.1.2.1 DESCRIPTION OF THE DATASET

The dataset, referred to as the "RSNA and STA Dataset," originates from the RSNA Pulmonary Embolism Detection Challenge on Kaggle. It was created to facilitate the development of artificial intelligence models aimed at detecting pulmonary embolism (PE) from computed tomography (CT) images. The primary goal is to enhance early diagnosis of PE, a potentially life-threatening condition requiring prompt treatment.

The dataset is organized into structured directories containing both patient images and metadata. The training portion includes 7,279 unique labeled patients, comprising a total of 1,790,594 CT images. The testing portion consists of 650 unique labeled patients and 146,853 CT images.

The original train dataset is annotated with 16 distinct labels, which include various factors such as patient ID, image-level data, and exam-level information (excluding the ID labels, which are unique identifiers). These labels are crucial for categorizing the data for different types of analysis and model training.

Below as shown in the next Table 2 a description of the different labels used is provided.

Table 2: Label description

Label	Type	Description
StudyInstanceUID	ID	Unique ID for each study (exam) in the data.
SeriesInstanceUID	ID	Unique ID for each series within the study.
SOPInstanceUID	ID	Unique ID for each image within the study (and data).
pe_present_on_image	image-level	Notes whether any form of PE is present on the image
negative_exam_for_pe	exam-level	Whether there are any images in the study that have PE present
qa_motion	informational	Indicates whether radiologists noted an issue with motion in the study.
qa_contrast	informational	Indicates whether radiologists noted an issue with contrast in the study.
flow_artifact	informational	Not described (meant to be used as helpers)
rv_lv_ratio_gte_1	exam-level	Indicates whether the RV/LV ratio present in the study is >= 1
rv_lv_ratio_lt_1	exam-level	Indicates whether the RV/LV ratio present in the study is < 1
leftsided_pe	exam-level	Indicates that there is PE present on the left side of the images in the study
chronic_pe	exam-level	Indicates that the PE in the study is chronic
true_filling_defect_not_pe	exam-level	Indicates a defect that is NOT PE
rightsided_pe	exam-level	indicates that there is PE present on the right side of the images in the study
acute_and_chronic_pe	exam-level	Indicates that the PE present in the study is both acute AND chronic
central_pe	exam-level	Indicates that there is PE present in the center of the images in the study
indeterminate	exam-level	Indicates that while the study is not negative for PE, an ultimate set of exam-level labels could not be created, due to QA issues

The next figure (Figure 14) illustrates the structure of the training dataframe, emphasizing the binarization of all labels except for the IDs.

Train DataFrame

	StudyInstanceUID	SeriesInstanceUID	SOPInstanceUID	pe_present_on_image	negative_exam_for_pe	qa_motion	qa_contrast	flow_artifact	rv_lv_ratio_gte_1
0	6897fa9de148	2bfbb7fd2e8b	c0f3cb036d06	0	0	0	0	0	0
1	6897fa9de148	2bfbb7fd2e8b	f57ffd3883b6	0	0	0	0	0	0
2	6897fa9de148	2bfbb7fd2e8b	41220fda34a3	0	0	0	0	0	0
3	6897fa9de148	2bfbb7fd2e8b	13b685b4b14f	0	0	0	0	0	0
4	6897fa9de148	2bfbb7fd2e8b	be0b7524ffb4	0	0	0	0	0	0

Figure 14: Train Dataframe dataset structure.

The following image (Figure 15) presents a flowchart outlining how each feature is labelled and the relationships between the labels. It is important to note that four labels in the training set—*QA Contrast, QA Motion, True Filling Defect Not PE,* and *Flow Artifact*—are purely informational and do not require predictions.

These labels are not scored but serve as auxiliary information. Additionally, the label '*Acute PE*' is not explicitly defined; it is implied by the absence of '*Chronic PE*' or '*Acute and Chronic PE*' labels.

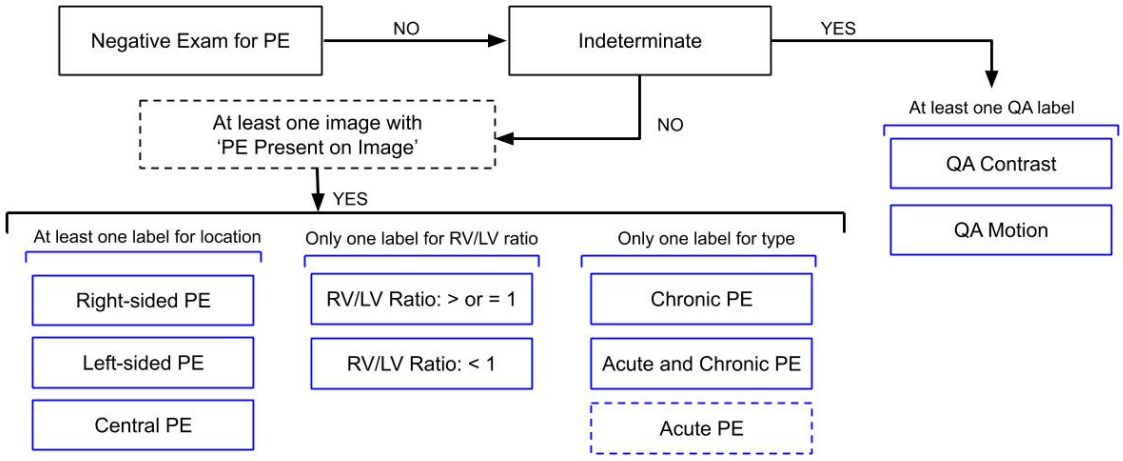


Figure 15: Labelling flowchart performed by physicians.

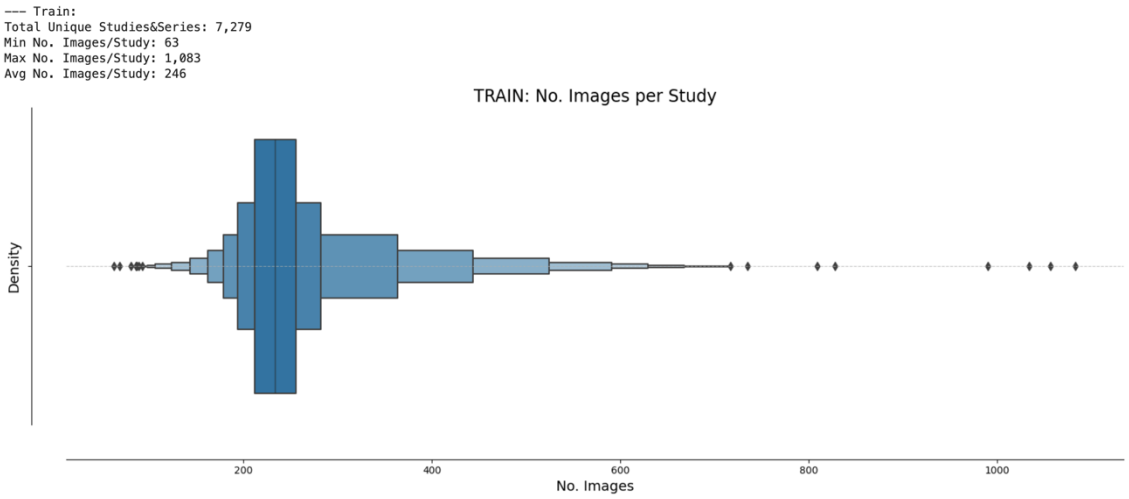
3.1.2.2 DATA DISTRIBUTION

INITIAL DATA QUALITY ASSESSMENT

Initial exploration of the data revealed that there are no missing values, highlighting its high quality and reliability. This thoroughness ensures the data is accurate and comprehensive, providing a solid foundation for developing a robust AI model.

IMAGES PER STUDY ANALYSIS

An in-depth analysis was conducted to calculate the ratio of images per study, which is presented in the following figures (Figure 16). The analysis revealed variations in the number of images per study due to the diverse sources of patient data. On average, there are 246 images per study in the training set and 226 images per study in the testing set.



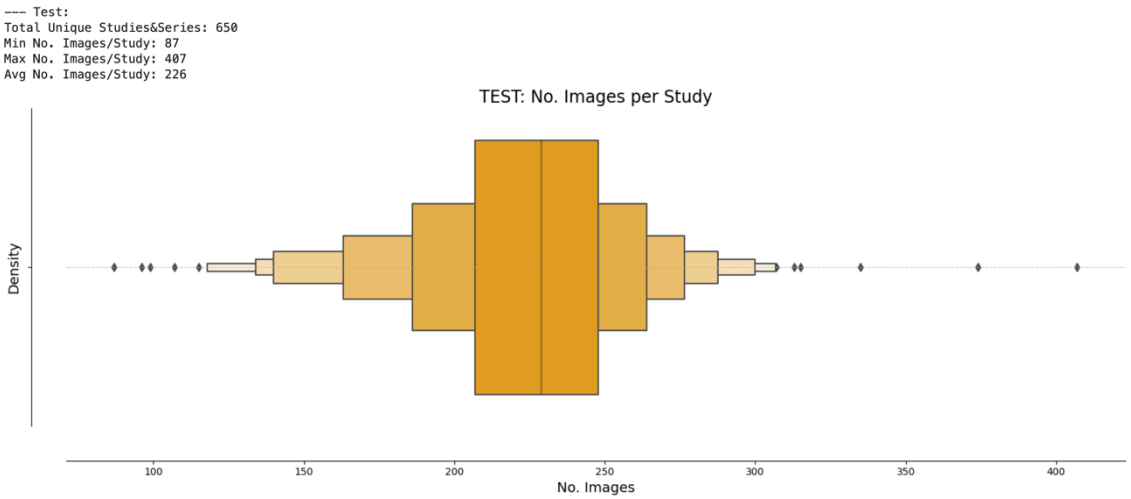


Figure 16: Train and Test Distribution of Images per Study.

The subsequent figure (Figure 17) summarizes these findings, illustrating the variability in the dataset.

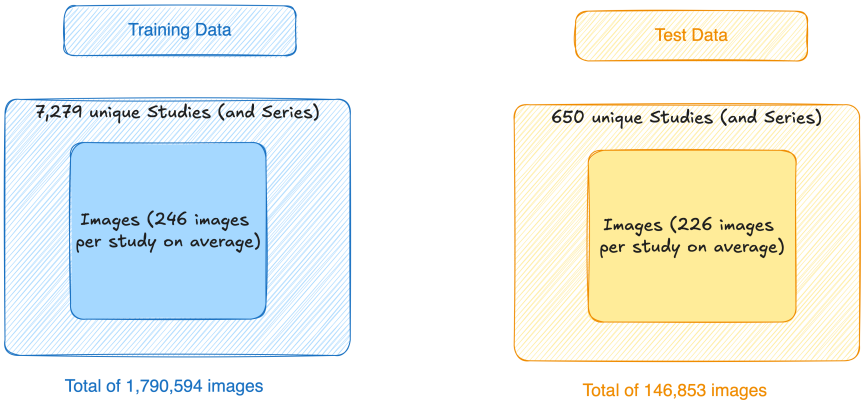


Figure 17: Summary of Image Distribution Analysis

CLASS DISTRIBUTION ANALYSIS

We will present an analysis of the class distribution within our dataset, focusing on the distinction between positive and negative cases of Pulmonary Embolism (PE). Understanding this distribution is crucial for evaluating the dataset's balance, which can significantly impact the performance and interpretation of machine learning models.

To provide a comprehensive view of the class distribution, we present the number of positive and negative PE cases at the image level:

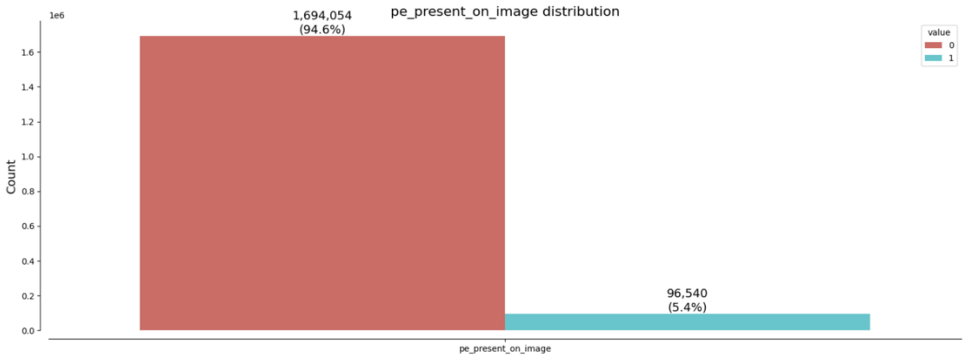


Figure 18: Class Distribution of PE Cases.

This chart demonstrates that the dataset contains a clear significant imbalance favouring the negative cases, which may influence the model's ability to detect PE accurately. The percentage of positive cases are 5,4% of PE presence and 94,6% of negative cases among the images (not studies (patients).

EXAM-LEVEL ANALYSIS

Now we will go in-deep between the rest of exam-levels features, which will give us the notion about the real sample of our data.

First of all, in the next figure (Figure 19) we present the distribution among the negative exams for PE, meaning that this label indicates cases where tests performed to diagnose pulmonary embolism (among all the images, whether there are any images in the study that have PE present) did not detect the presence of the condition, i.e. the results were negative.

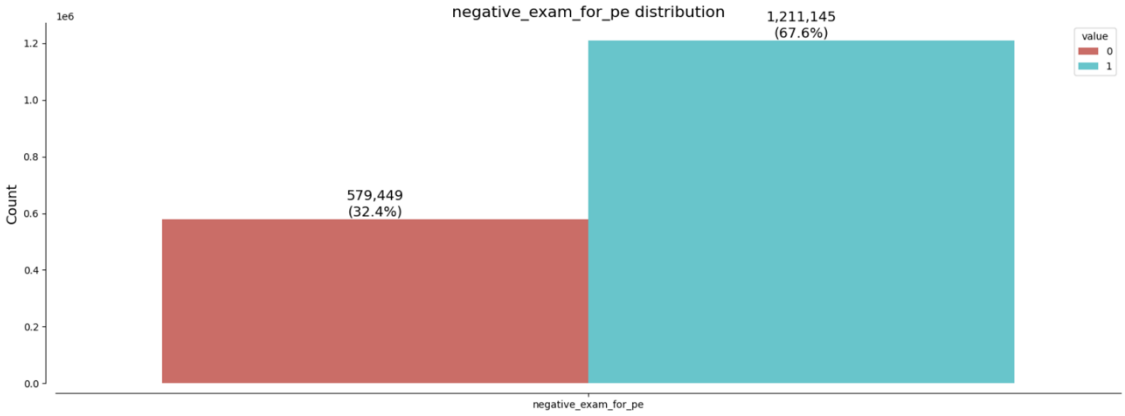


Figure 19: Example of a visualization of the distribution of negative and positive exams for PE.

In this dataset, 67.6% of the exams are classified as negative, while 32.4% are positive for PE. This distribution provides insight into the burden of the disease within the study population and highlights the importance of accurate diagnostic procedures.

This type of analysis is important to understand the burden of disease in the study population and to assess whether diagnostic procedures are being applied effectively.

THROMBOUS QUANTITY ANALYSIS

Observing the previous analysis, we can conclude that the ratio of PE per patient is not consistency, and the trombo may appear more than one time per patient. This means that patients with the disease usually have more than one PE in their lungs, which is crucial to understand. In the next figure (Figure 20) I illustrate representation of number of PE per samples.

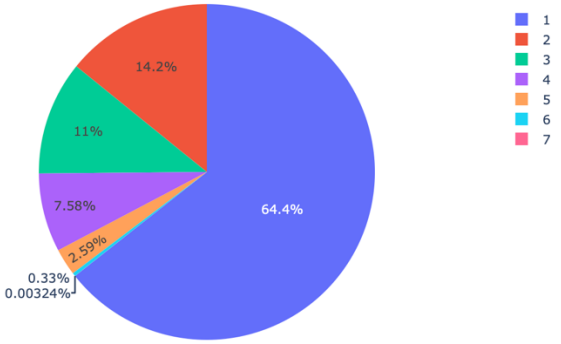


Figure 20: Distribution of Thrombus Quantity per Patient.

In this analysis, we observe that among patients diagnosed with PE, the majority—64.4%—present with a single thrombus. However, a noteworthy observation is that 35.6% of patients have multiple thrombi. This diversity in thrombus quantity significantly enhances the learning phase of our AI algorithm, improving its predictive accuracy.

CHRONIC AND ACUTE PE ANALYSIS

The dataset also distinguishes between chronic and acute PE cases, which is crucial for understanding the nature and severity of the disease. Chronic PE, present in 4.0% of cases, involves long-standing thrombi that have led to the organization and scarring of emboli within the pulmonary arteries. This condition can result in chronic thromboembolic pulmonary hypertension (CTEPH). Acute PE, characterized by recent thrombi formation, is less frequent. Cases featuring both acute and chronic PE account for 1.9% of the dataset.

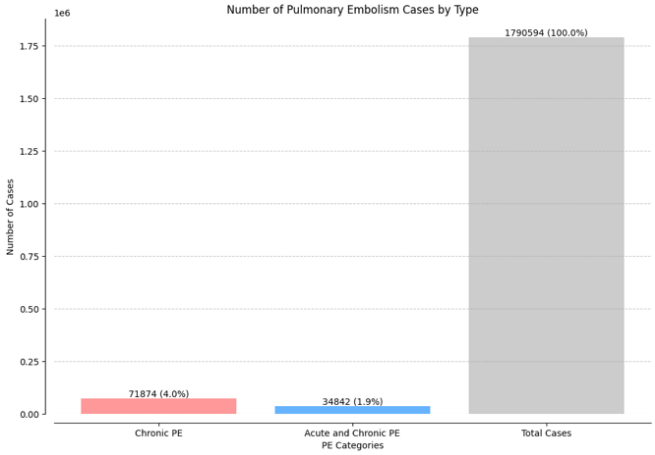


Figure 21: Distribution of Chronic and Acute PE Cases respect to the total cases.

The noted imbalance in chronic PE cases (4.0%) compared to acute and chronic PE cases (1.9%) is significant for model training, as it suggests a need for the model to be particularly sensitive to these less common types to avoid bias.

The mention of these percentages highlights the imbalance within the dataset. Specifically, the chronic PE cases constitute a larger portion of the dataset (4.0%) compared to the acute and chronic PE cases (1.9%). This imbalance is important for understanding the distribution of different types of PE in the dataset and may have implications for the training and evaluation of AI models, as models may need to be particularly sensitive to detecting less common types of PE to avoid bias.

RIGHT VENTRICLE / LEFT VENTRICLE (RV/LV) RATIO ANALYSIS

In patients with PE, the RV/LV ratio is an important metric, often indicating right ventricular enlargement—a critical factor as it can lead to right ventricular dysfunction and, potentially, cardiogenic shock.

As shown in the next Figure 22, RV/LV ratio analysis shown that in our sample the 17.4% of patients suffer from rv/lv ratio lower than 1. An in the other side, the 12.9% of patients suffer from RV/LV ratio greater than 1.

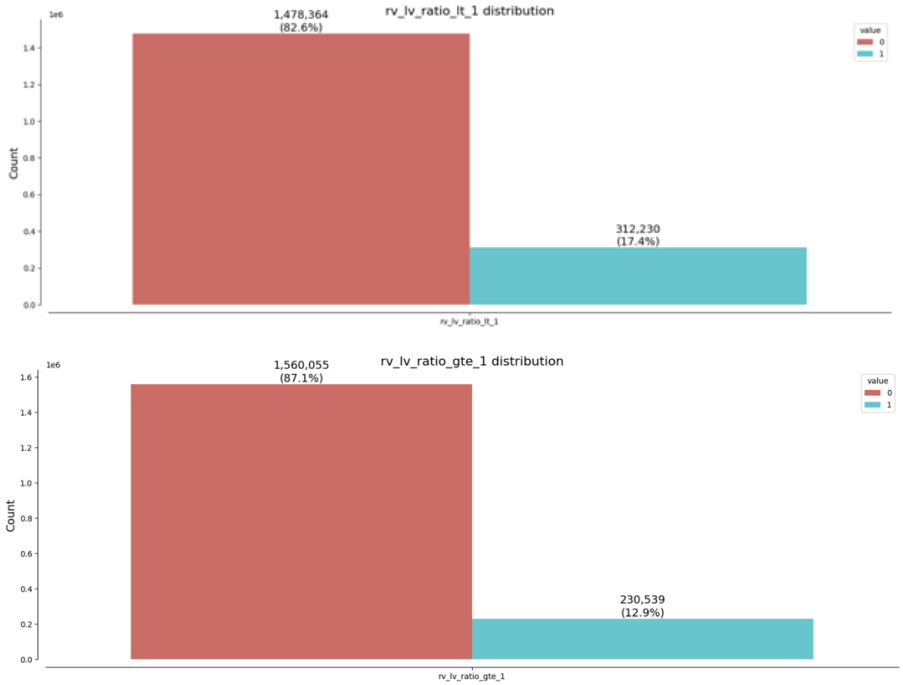


Figure 22: RV/LV Ratio Analysis and Its Implications

In the next figure we can appreciate how people who has a ratio of ventricle greater than 1, has more predominance to suffer chronic and acute PE. This is something to highlight and that has discovered by making this analysis.

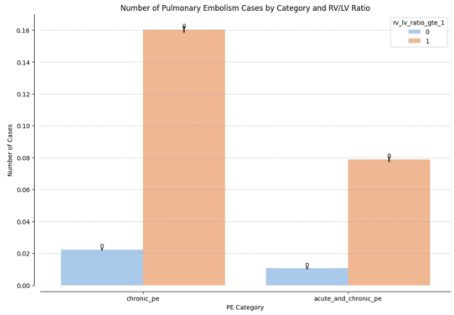


Figure 23: Distribution of PE cases by rv/lv ratio greater than 1 versus chronic and acute and chronic PE.

LOCATION RATIO ANALYSIS VS RATIOS

In addition to analysing the general distribution of pulmonary embolism (PE) cases, it is crucial to examine the location of the emboli within the lungs, specifically differentiating between right-sided and left-sided PE, as well as central PE cases.

The analysis revealed that the majority of PE cases are located in the right lung, with less cases in the left lung, and a certain proportion involving central PE. This distribution is consistent with clinical observations where the right lung, being larger, is more frequently affected by emboli. Furthermore, the analysis of the RV/LV ratio, which indicates the size disparity between the right and left ventricles, is critical in assessing the severity of PE. Elevated RV/LV ratios, particularly those greater than 1, often indicate significant right ventricular strain and are associated with worse clinical outcomes.

This detailed location ratio analysis not only aids in understanding the anatomical distribution of PE but also helps in tailoring the AI model to be sensitive to variations in clinical presentations, ultimately improving its diagnostic accuracy and clinical applicability.

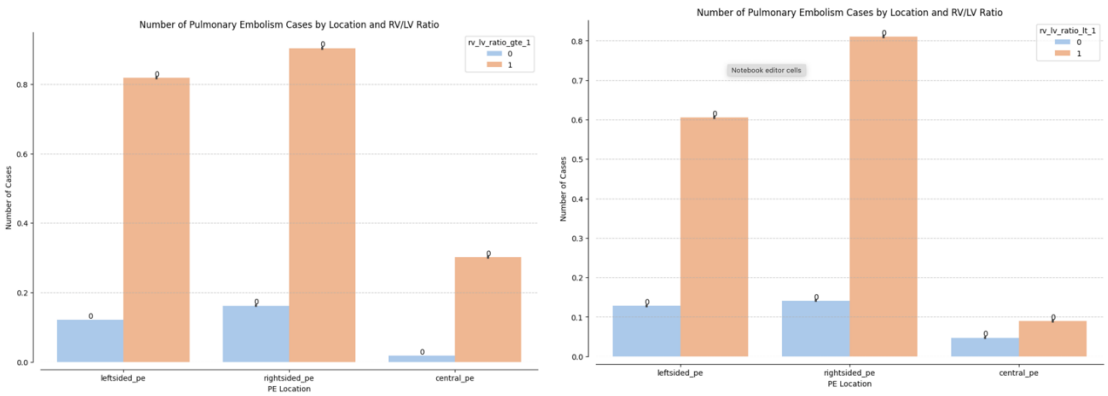


Figure 24: Number of PE cases by location and RV/LV Ratio

LABEL DISTRIBUTION AND IMBALANCE VISUALIZATION

The final figure presents the distribution of all the predicted labels, highlighting the dataset's imbalance. It also includes a visualization of labels that are informational or indicate noise, which are outside the scope of this project but still essential for understanding the dataset's composition.

Finally in the next Figure 25 we can observe the overall aimed to predict labels and the magnitude of the unbalance in the dataset.

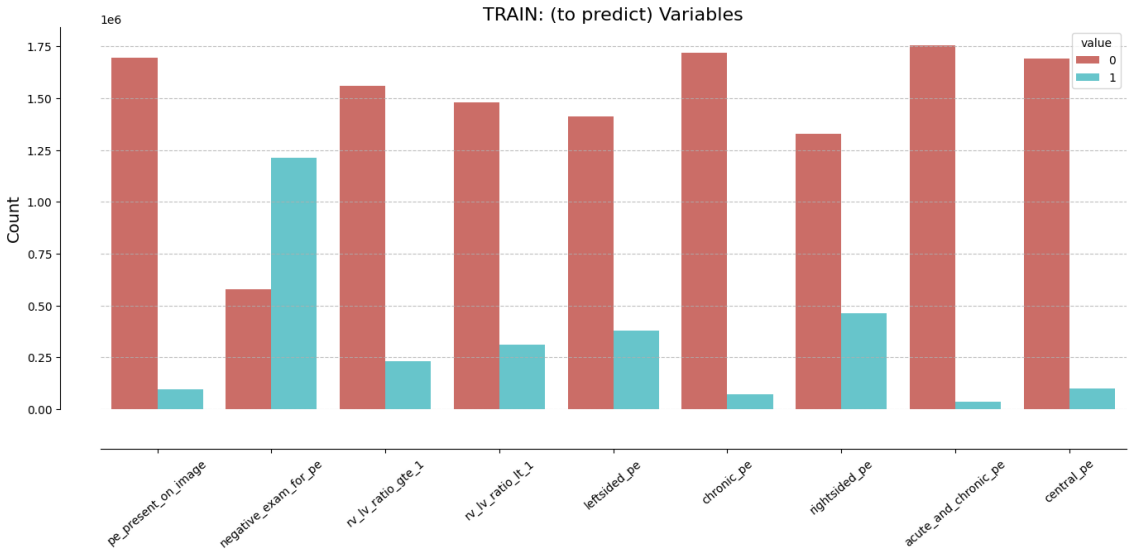


Figure 25: Overall Predicted Label Distribution and Data Imbalance.

Also, we ploted the labels that indicate noise, errors or helpers, as long as they will be out of the scope of this project. As shown in the next Figure 26 we can observe the noises count vs the good images.

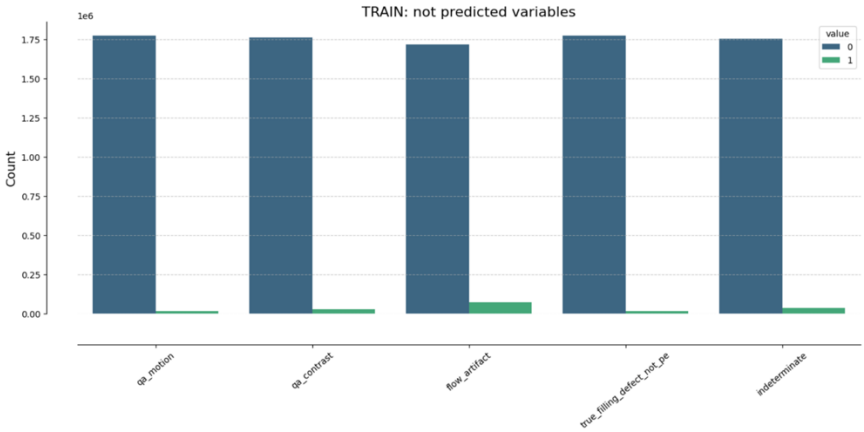


Figure 26: Overall non-Predicted Label Distribution and Data Imbalance.

In this next figure (Figure 27) we can observe all the labels with non-zero entries to have an holistic vision of all the labels.

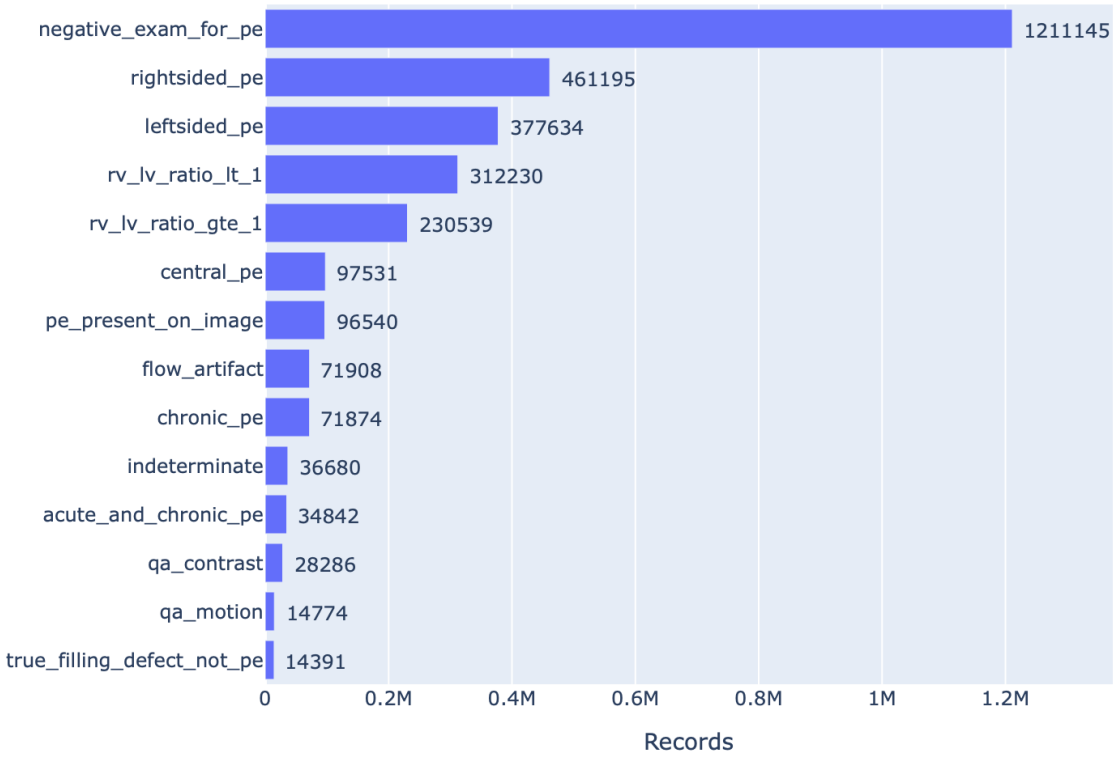


Figure 27: Positives Labels (Non-Zero Entries)

From the previous figure, we can observe that mostly patients with the disease is more possible to observe PE in the right-side of the lungs. And most patients present a RV/LV ratio less than 1 with 461195 followed by the patients who present RV/LV ratio greater or equal to 1.

This comprehensive analysis underscores the importance of considering dataset imbalance and the diversity of PE presentations in developing effective AI models. Addressing these imbalances and variations is crucial for improving the predictive accuracy and generalizability of the AI system.

3.1.2.2 VISUAL ANALYSIS AND DICOM METADATA

The RSNA STR Challenge dataset consists of CT axial thorax images in DICOM format. The DICOM format not only preserves high image quality but also embeds additional metadata, which is valuable for detailed analysis. The axial thorax images primarily capture the lungs and surrounding body structures as seen when the patient lies inside the CT machine.

Figure 28 illustrates a typical image from the dataset, where we can see that the lungs, which are the primary area of interest for detecting pulmonary embolism (PE), do not occupy the entire image. This observation is crucial for the preprocessing stage, where the focus will be on isolating the lungs, either through image cropping or other techniques. By concentrating only on the lung regions, we can enhance the accuracy of PE detection.

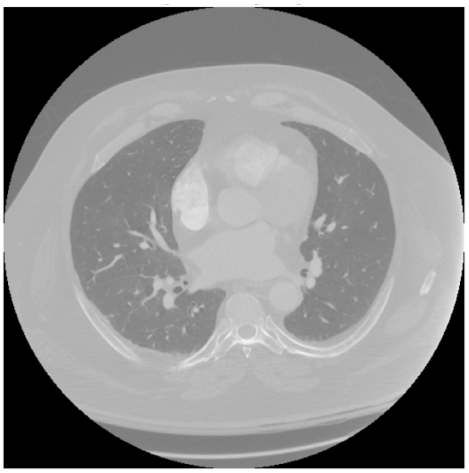


Figure 28: RSNA PE Challenge DICOM Image (Original Pixel Values).

Since the dataset includes both images with and without pulmonary embolism (PE), we can compare these two conditions in Figure 29 below.

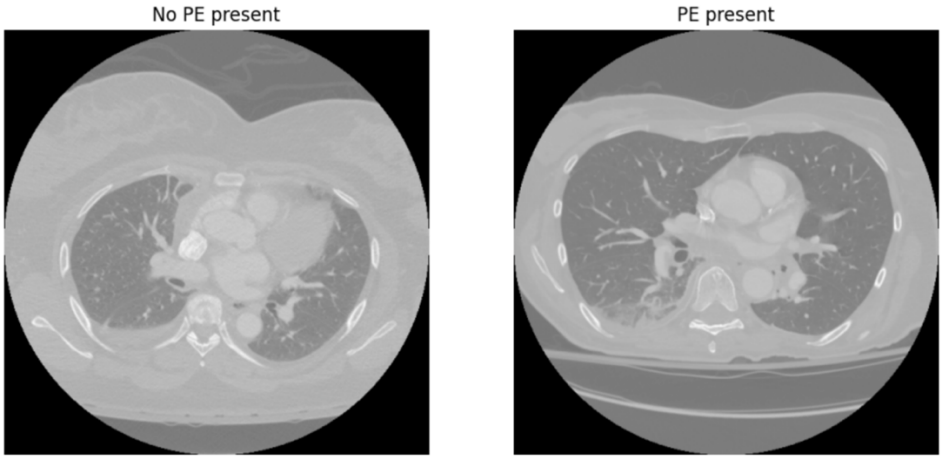


Figure 29: CT images from the dataset showing both a diseased patient (with PE) and a healthy patient (without PE).

In both cases, identifying the thrombus (blood clot) is a challenging task. Even for experienced physicians, detecting PE can be difficult under certain conditions due to subtle or ambiguous visual indicators.

Figure 30 shows a sequence of axial thoracic CT slice images for a single patient.

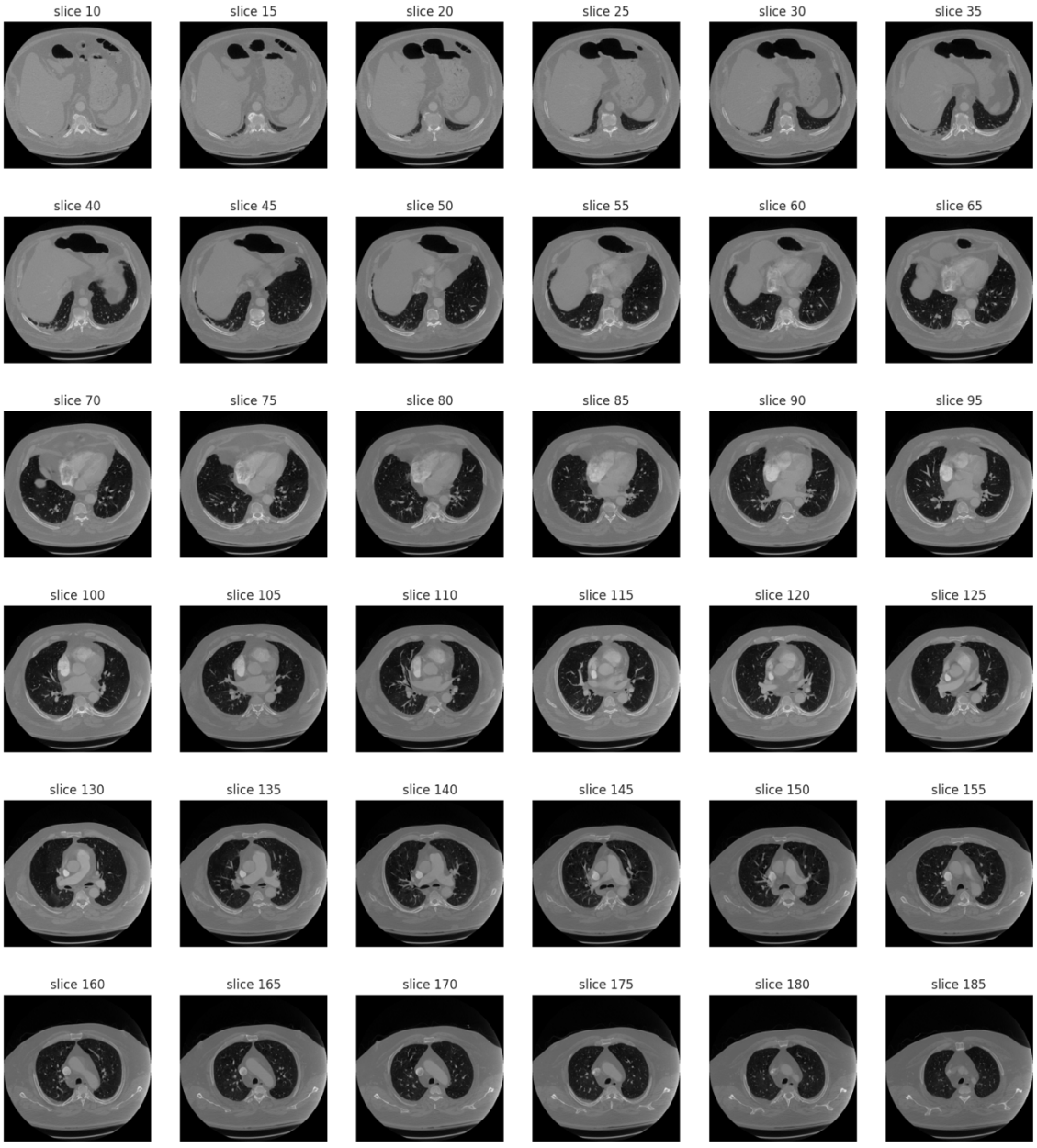


Figure 31: Sequence of axial thoracic CT slices for a single patient, displayed at intervals of 5 slices each.

Regarding the DICOM data of the overall dataset we can visualize 'hardware' aspects of the images such images sizes resolutions and slices volume spacing.

In the following Figure 31 can assess that all the images from the original RSNA STR Challenge Dataset is 512 x 512 pixels.

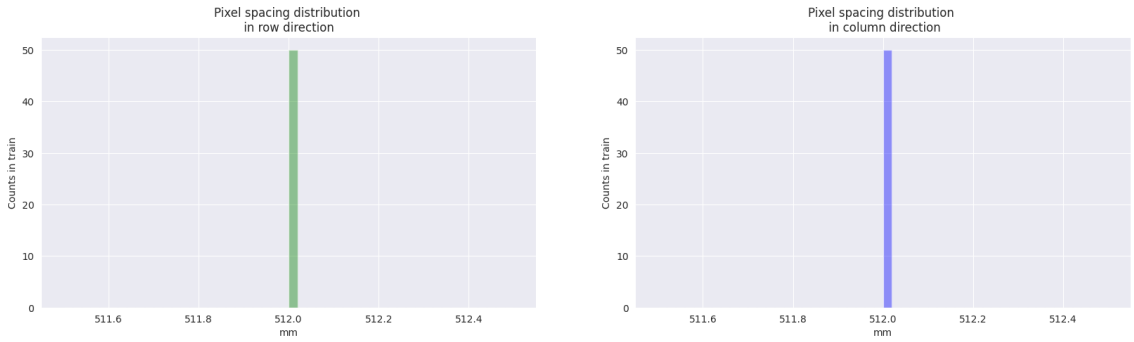


Figure 32: Pixel Resolution distribution for all dataset DICOM images. Here we can assess that all images are 512 x 512 pixels.

Another advantage of DICOM files is their ability to provide Hounsfield Unit (HU) values, which are essential for categorizing different types of tissues and substances in CT images. HU values help differentiate between soft tissues, bones, air, blood, and emboli (such as those found in pulmonary embolism). In the next Figure 32 , we present the distribution of HU values for a single CT image, ranging from -1000 to 3000 HU. This distribution illustrates the varying densities of tissues and substances, aiding in more accurate identification and analysis of abnormalities.

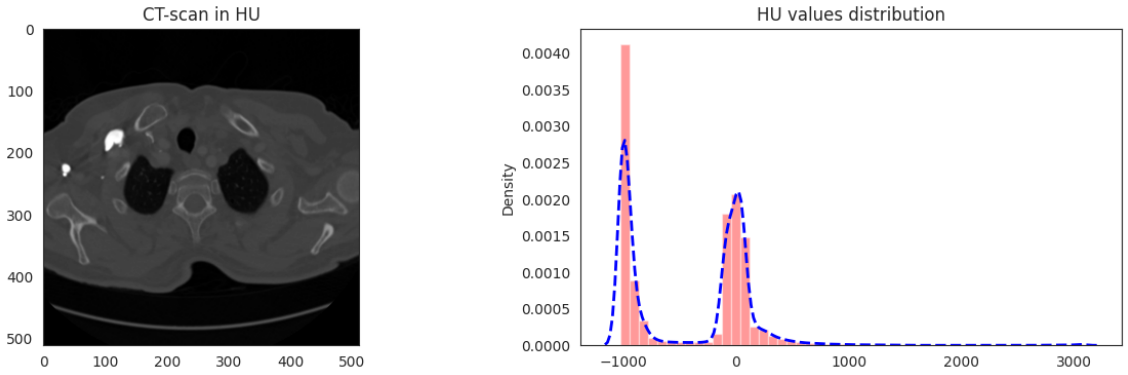


Figure 33: Distribution of Hounsfield Unit (HU) values for a single CT image, ranging from -1000 to 3000 HU. This distribution highlights the varying densities of tissues and substances within the image.

Another aspect of the data analysis involved evaluating slice volumes based on pixel spacing and area. Figure 33 illustrates this assessment by comparing three images. The analysis shows a consistent correlation across these images, indicating uniformity in slice volumes throughout the dataset.

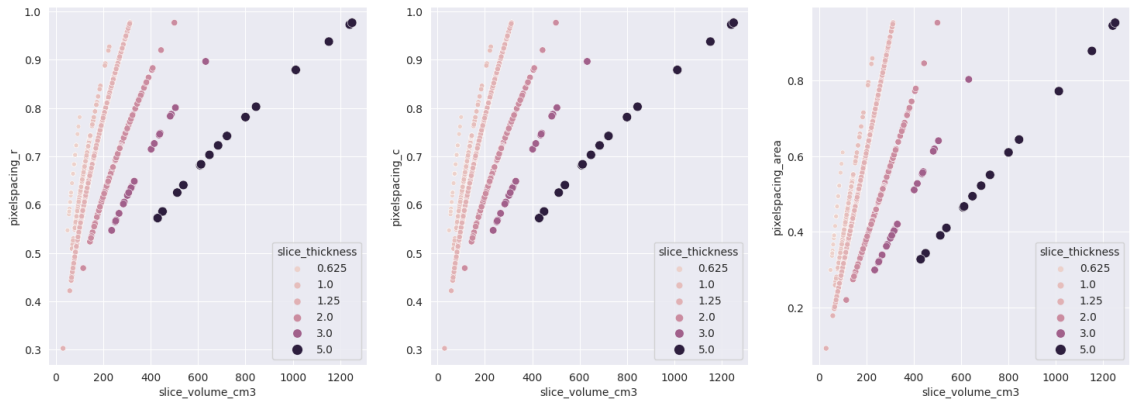


Figure 34: Correlation of slice volumes based on pixel spacing and area across three images. This assessment confirms the consistency of slice volumes throughout the dataset.

The next assessment focused on the pixel spacing in both the column and row directions across all images. As shown in Figure 34, the distribution of pixel spacing is consistent between columns and rows, indicating uniformity in the image layout. However, there is notable variability in pixel spacing values across different images, revealing inconsistencies in how pixel spacing is applied throughout the dataset.

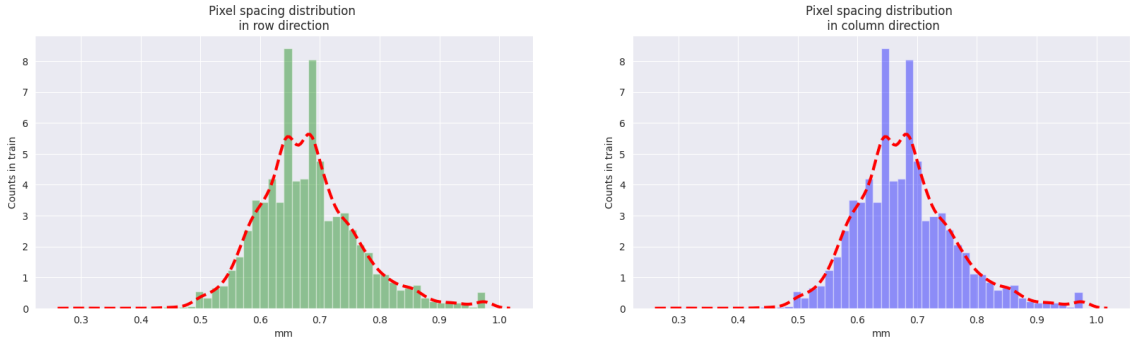


Figure 35: Distribution of pixel spacing in the column and row directions across all images.

The last assessment involved the consideration of the patient position at the time of taking CT exam. As shown in the next Figure 35, we can assess that 6.977 patientes 95.9% where taking in FFS position, 301 patients 4.1% where taking in HFS position and only 1 patient was in FFP position.

The final assessment analyzed patient positioning during the CT exams. As depicted in Figure X, the majority of patients—6,977 (95.9%)—were positioned in the feet-first supine (FFS) position, while 301 patients (4.1%) were in the head-first supine (HFS) position. Only one patient was positioned in the feet-first prone (FFP) position.

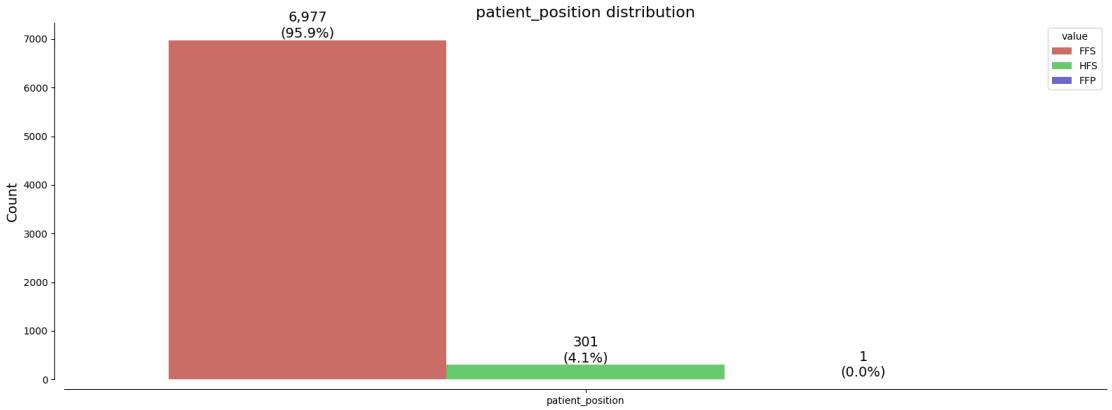


Figure 36: Distribution of patient positioning during CT exams.

3.1.2.3 STATISTICAL ANALYSIS

Starting with the Pearson correlation analysis for the RSNA dataset provided, we observed that the target label 'pe_present_on_image' is most strongly correlated with 'leftsided_pe', 'rightsided_pe', 'central_pe', and 'rv_lv_ratio_gte_1'. The high correlation with the first three labels is expected since they directly relate to the location of the pulmonary embolism (PE), naturally leading to a strong relationship.

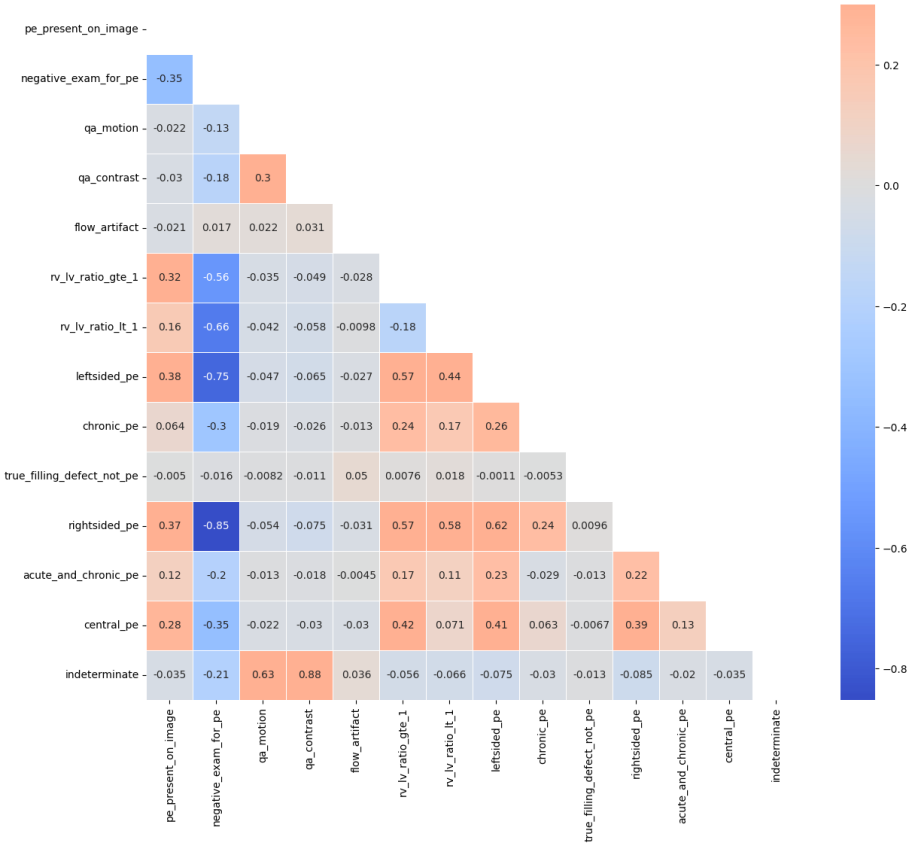


Figure 37: Dataset Pearson Correlation Analysis.

Preliminary insights suggest that the RV/LV ratio (the ratio of the right ventricular to left ventricular dimensions) may be associated with the presence and severity of PE. This finding aligns with clinical understanding of PE, where a clot in the lungs increases pressure in the pulmonary vessels, causing the heart to work harder. This additional strain can lead to an elevated right ventricle (RV) to left ventricle (LV) ratio, which reflects changes in heart structure due to increased pressure.

Specifically, patients with an RV/LV ratio greater than 1 were more likely to exhibit PEs in the central regions of the chest, suggesting a more severe presentation that could involve central pulmonary arteries.

Additionally, it was observed that PEs more frequently appeared in the left-sided parts of the lungs, particularly in the deeper pulmonary arteries. This pattern may indicate a possible anatomical predisposition or diagnostic artifact, but it warrants further investigation to determine if it correlates with patient outcomes or specific risk factors.

These initial patterns highlight the importance of considering anatomical and physiological indicators like the RV/LV ratio in the analysis. Understanding these patterns could lead to better predictive models and more targeted clinical interventions, as these features may indicate higher risk or severity in patients with PE. Further exploration of these trends, including a breakdown by demographic factors such as age and gender, could provide more nuanced insights into the factors influencing PE distribution and presentation.

In summary, the analysis highlights correlations related to PE location and image-related factors such as contrast quality, motion artifacts, and indeterminate labels due to poor image quality as identified by physicians. Based on these insights, we plan to filter out and remove images with indeterminate labels and significant noise from the dataset to improve overall analysis accuracy.

When conducting a correlation analysis on the DICOM metadata across all images, as shown in Figure 37 we observe strong consistency among technical factors related to the CT hardware settings, such as pixel spacing, window level, and exposure. This consistency suggests a standardized imaging process, which is beneficial for maintaining uniformity across the dataset and reducing variability due to imaging conditions.

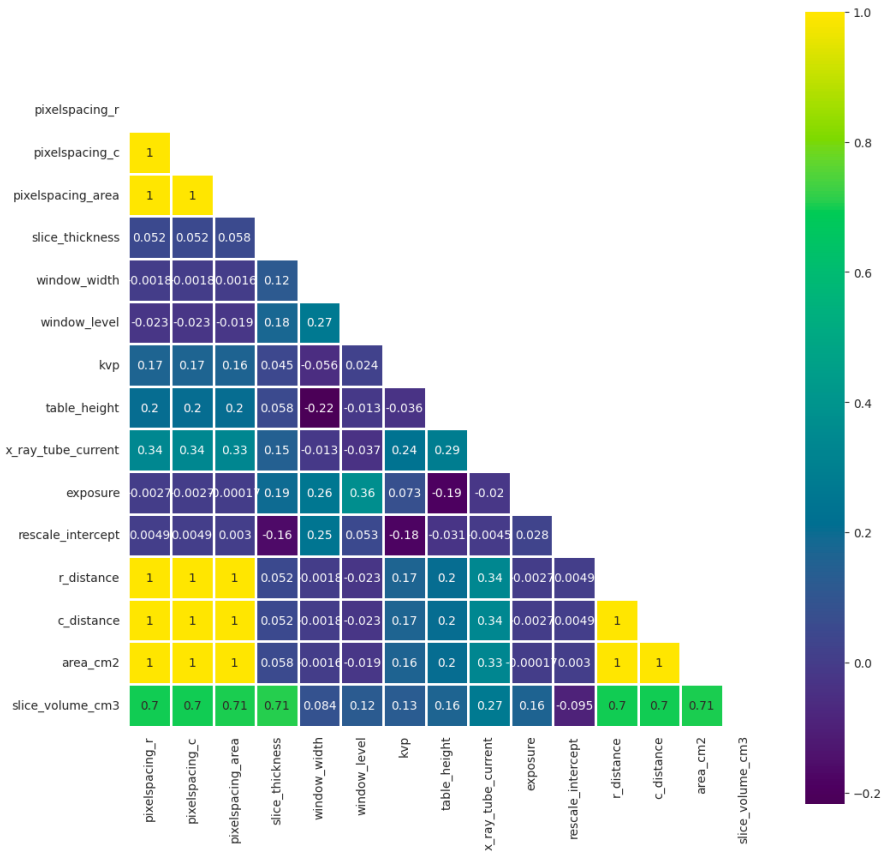


Figure 38: Correlation analysis of DICOM metadata across all images.

The uniformity in technical parameters across the dataset implies that the imaging protocol was consistent, reducing potential noise and ensuring that the variations in image quality are minimal.

In future projects, incorporating these technical factors into the model could improve performance. For instance, adjusting preprocessing steps or fine-tuning models based on these settings may lead to better generalization and prediction accuracy.

Understanding the impact of technical factors like window level and exposure could help in building more robust models that are less sensitive to hardware-related variations, making them applicable to datasets from different imaging sources.

3.1.2.3 DATA CLEANING AND PREPROCESSING

Due to the variability in image quality, such as noise, motion artifacts, and contrast variations, we decided to exclude these images from our sample to avoid compromising accuracy. Specifically, we removed data instances labelled with *qa_motion*, *qa_contrast*, *true_filling_defect_not_pe*, and *flow_artifact*. These labels are categorized as auxiliary or quality control markers in the RSNA dataset and are not utilized as predictors in our analysis. Their removal helps streamline the

dataset by excluding non-predictive information that does not contribute to the detection of pulmonary embolism.

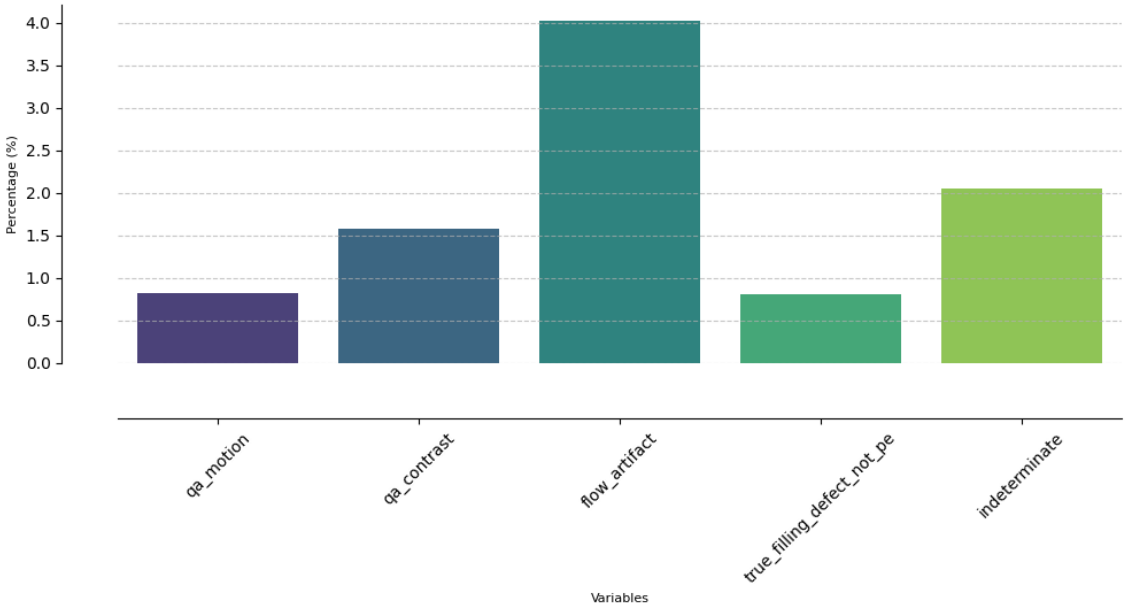


Figure 39: Percentage of occurrences of errors to be suppressed in each column with repect to the total number of variables in Train.

Additionally, we excluded patients with fewer than 100 or more than 500 CT scans to maintain consistency across the samples. This threshold was chosen to ensure a sufficient amount of data per patient for reliable analysis while avoiding potential biases from patients with an excessively high number of scans, which could indicate repeated or follow-up imaging rather than initial diagnostic scans. This filtering process helps to create a more homogeneous and manageable dataset, improving the robustness of our subsequent analysis and modelling efforts.

As shown in the next Figure 40 after applying this data cleaning the sample reduced by an average of images per study of 240 in order to homogenize the sample and prevent overfitting and help model to improve the accuracy and sensitivity.

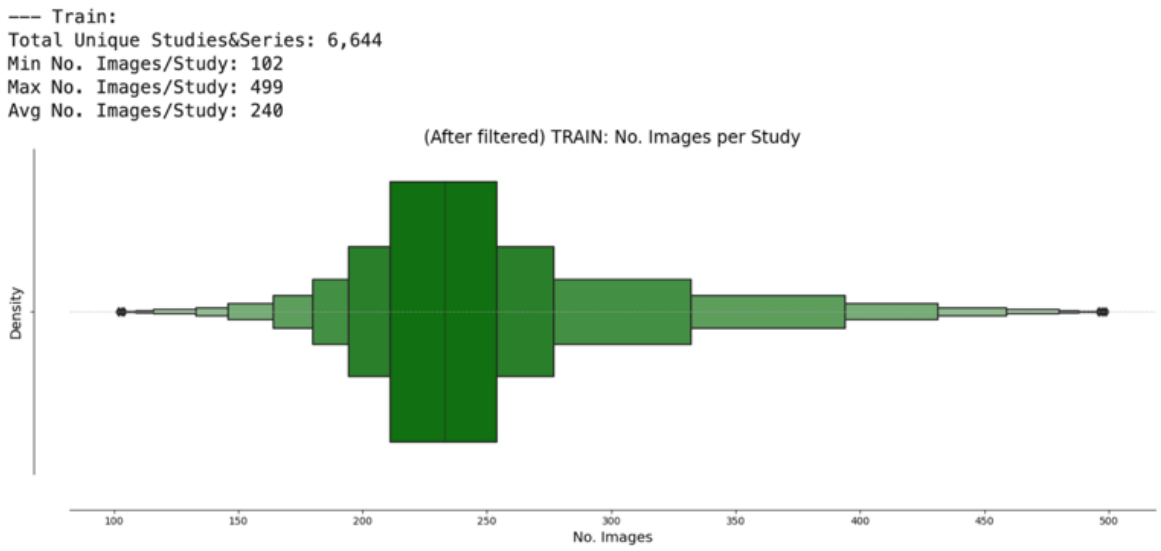


Figure 40: Train distribution of images per study after data cleaning.

After applying these cleaning procedures, the dataset was reduced from 1,790,594 to 1,592,613 samples, representing a 11.05% reduction. Additionally, the number of unique studies (patients) decreased from 7,279 to 6,644. In the following Figure 41 a final statistics schema in order to clarify steps as been made.

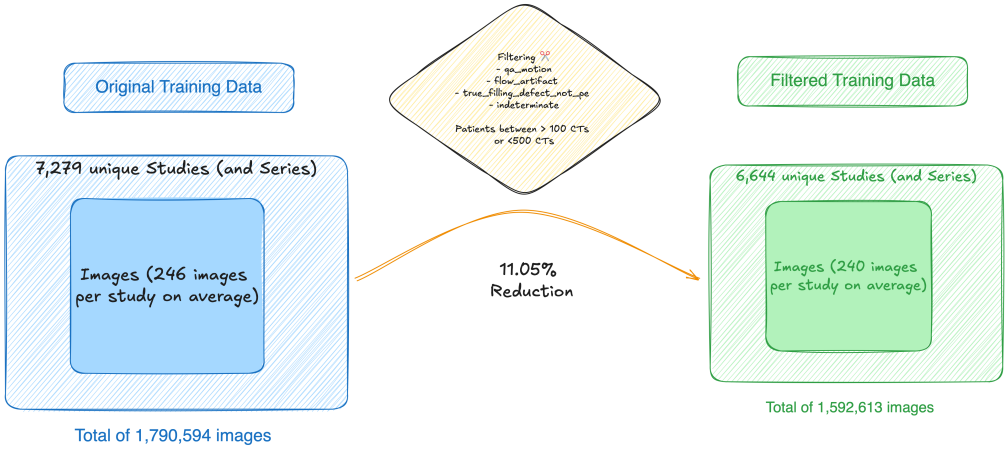


Figure 41: Schema with the numbers of images, average of images per study and unique studies (patients) diference and the cleanings made.

3.1.2.4 INSIGHTS AND PRELIMINARY FINDINGS AFTER DATA CLEANING AND PREPROCESSING

Finally, after making the data cleaning, to assess if there are changes in the distributions, a analysis as been provided in Figure 42 below.

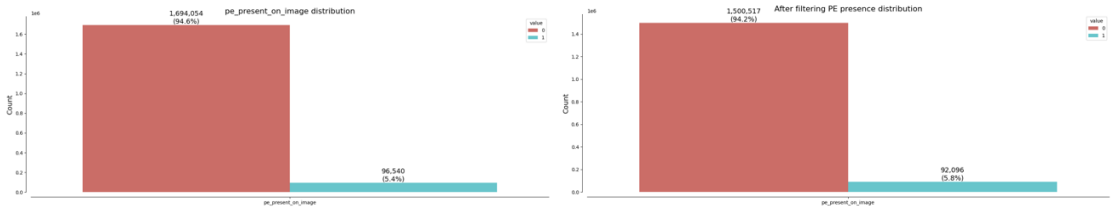


Figure 42: Assessment of distrubtion changes in the PE presence after and before data cleaning and preprocessing.

There is evidence that distributions have not been change and distribution has not moved with the reduction of the sample.

Finally, negative_exam_for_pe for Study after data cleaning stayed in 31.1% (520,600) patients with PE disease and 68.9% (1,152,417) patients without the PE disease.

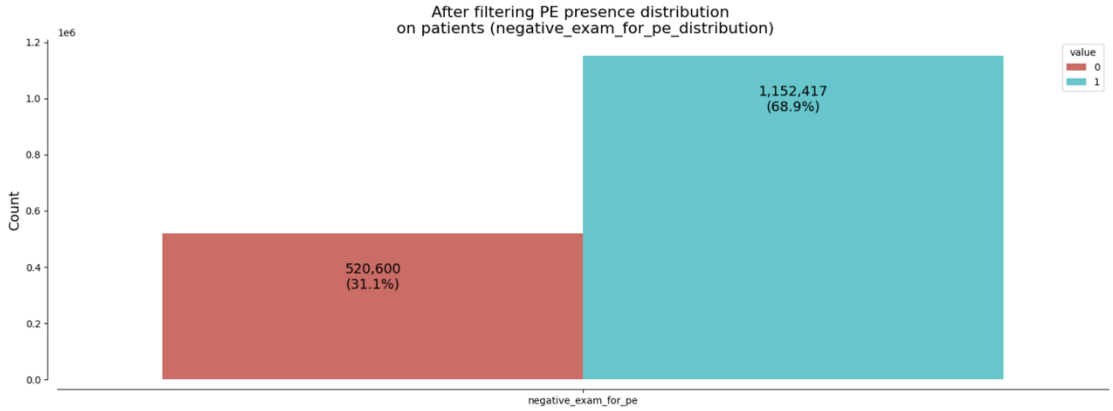


Figure 43: Negative_exam_for_pe distribution after data cleaning and preprocessing on patients with PE.

Finally a Person Correlation heatmap as been performed in order to compare correlation among labels. In the next Figure 44, we can observe how correlations among the target label 'pe_present_on_image' has not been moved.

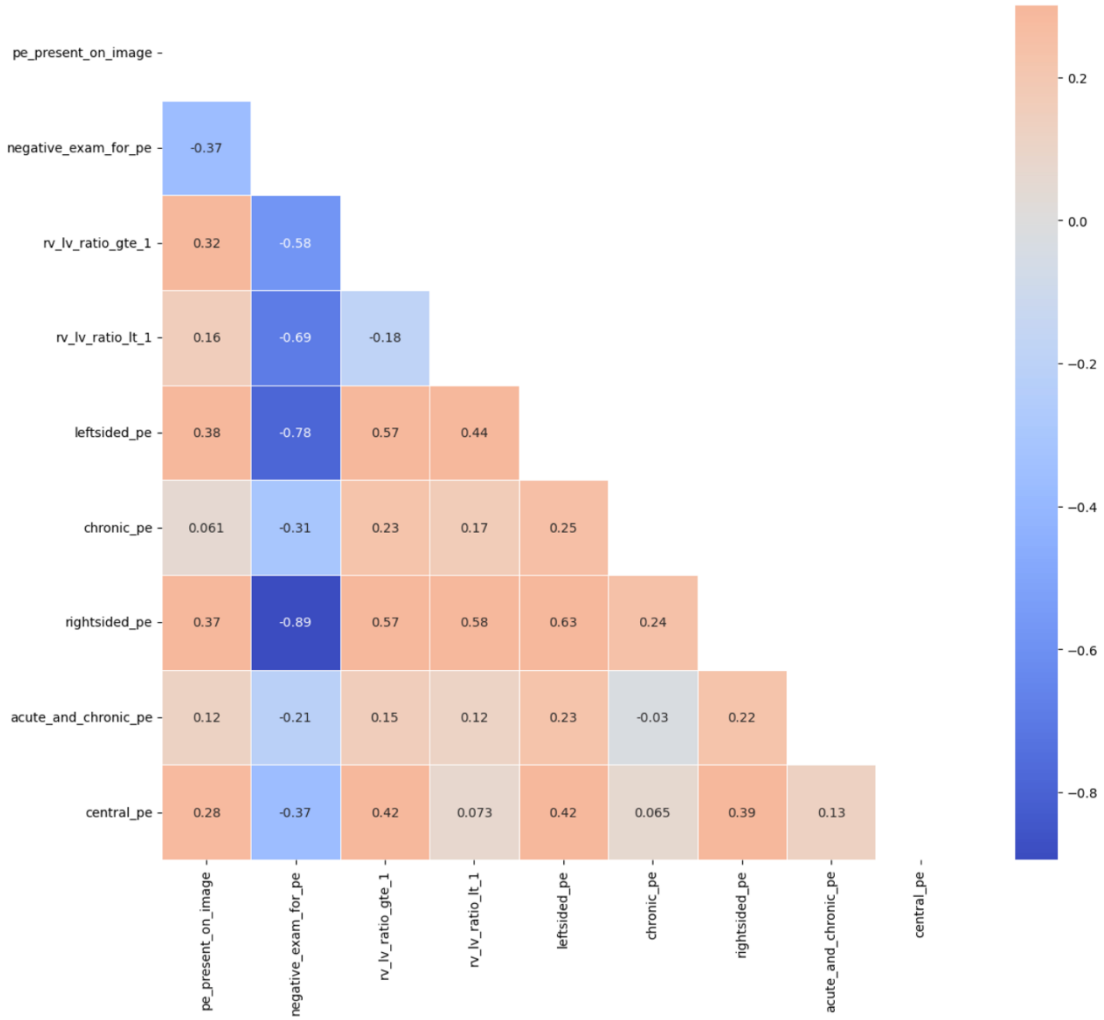


Figure 44: Pearson Correlation heatmap after the data cleaning and preprocessing.

3.1.1. EXPLORATORY DATA ANALYSIS CHALLENGES AND LIMITATIONS

During the Exploratory Data Analysis (EDA), several challenges were encountered, including data inconsistencies and inherent limitations within the dataset. For instance, there was a noted lack of data diversity in patient demographics, gender, life habits, which could impact the model's ability to generalize. This lack of diversity may lead to biases in the model, potentially affecting its performance when applied to populations not represented in the training data.

Another lack of information will be the precedence of the data, and the professionals who annotated the data, because sometimes this kind of data must be taken as an important feature to discard or maintain the data and ensure the consistency of the results.

Some state-of-the-art research has explored approaches that combine both ethnic data (ETH) and imaging data to provide improved results and reduce bias. Integrating ethnic characteristics and other demographic variables could result in a more robust and less biased model, with better generalization capabilities across diverse populations.

To mitigate the bias due to limited diversity, techniques such as data augmentation or class weighting could be employed. Additionally, regularization approaches that penalize bias towards specific groups could be considered.

Acknowledging and addressing these challenges is crucial for developing a reliable and generalizable model for PE detection.

3.1.2. PREPROCESSING STEPS WITH IMAGES USING DICOM FORMAT

Preprocessing the CT scan images is a critical step to ensure the quality and consistency of the data fed into the deep learning model.

In Annex II, a preprocessing steps has been performed in order to segmentate the lungs in each image.

In the Figure 45 shown below, we can observe the lung segmentation in one single image.

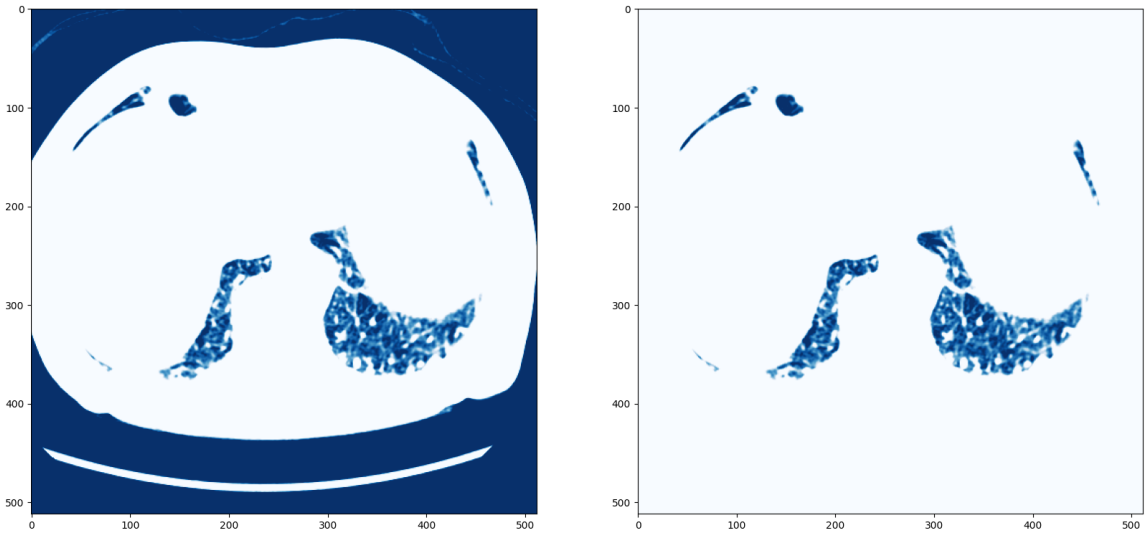


Figure 45: Image before and after lung segmentation.

Applying this lung segmentation preprocessing to a single patient we obtain the following result as we can see in the next Figure 46.

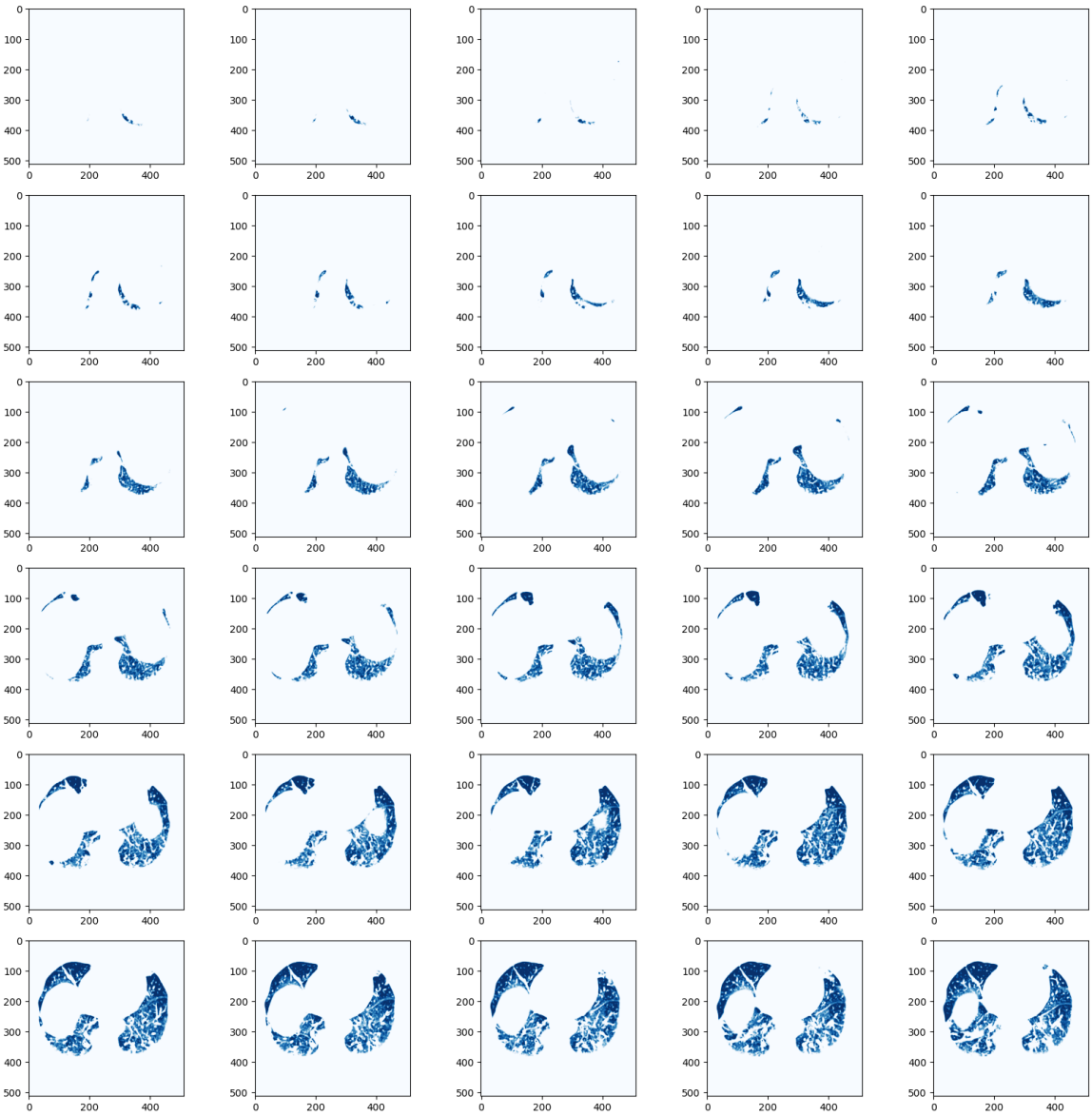


Figure 46: Lung segmentation slices for a single patient

Once the lungs have been segmented from the DICOM image and the entire sequence of slices for each patient is processed, we can create powerful 3D visualizations like the one shown in Figure 47. These visualizations can display both the external and internal structures, thanks to the importance of selecting the appropriate HU unit, meaning filtering by the HU value we want to observe.

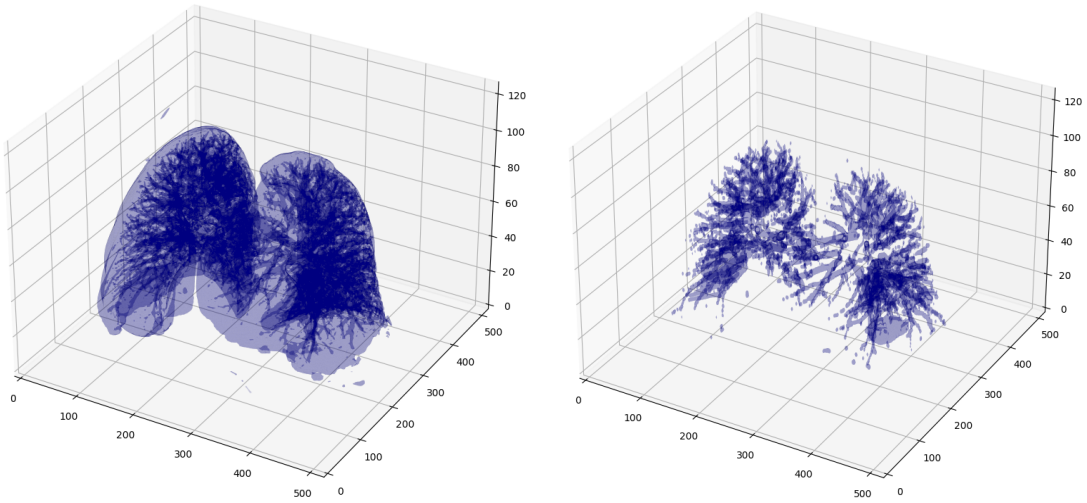


Figure 47: 3D Lung segmentation (left side) and lungs vessels segmentation (right side).

Finally, after considering the advantages of using the DICOM format, including the ability to leverage HU values for precise lung segmentation, one key challenge became evident: the processing time required when working with DICOM files. Given the weekly computational limitations, alternative techniques were employed. However, this project provided valuable insights into the significant benefits that could be realized by incorporating DICOM data more extensively. In future work, with increased computational resources and extended processing time, these approaches could lead to notable improvements in model performance and overall results.

3.2. MODEL SELECTION AND DESIGN

3.2.1. EXPLANATION OF CHOSEN DEEP LEARNING ARCHITECTURES

For the detection of pulmonary thromboembolism, convolutional neural networks (CNNs) were chosen due to their proven effectiveness in image analysis tasks. Specifically, the following architectures were considered:

I. **ResNet (Residual Networks):** Known for its deep architecture and ability to mitigate the vanishing gradient problem, ResNet allows for the training of very deep networks by using residual connections. This architecture is particularly effective in learning complex features from medical images.

II. **VGG (Visual Geometry Group):** With a simpler and more uniform architecture, VGG networks are effective in image classification tasks and serve as a good baseline model. Despite its simplicity, VGG has been shown to perform well in various image recognition challenges.

III. **U-Net:** Particularly suited for medical image segmentation, U-Net was considered for its ability to provide pixel-level predictions, which is useful for detecting and localizing PE. Its encoder-decoder structure allows for precise delineation of regions of interest within the images.

IV. **EfficientNet**: EfficientNet was included due to its state-of-the-art performance in image classification tasks. EfficientNet uses a compound scaling method that uniformly scales all dimensions of depth, width, and resolution using a simple yet highly effective principle. This architecture achieves high accuracy with fewer parameters, making it both efficient and powerful for medical imaging applications.

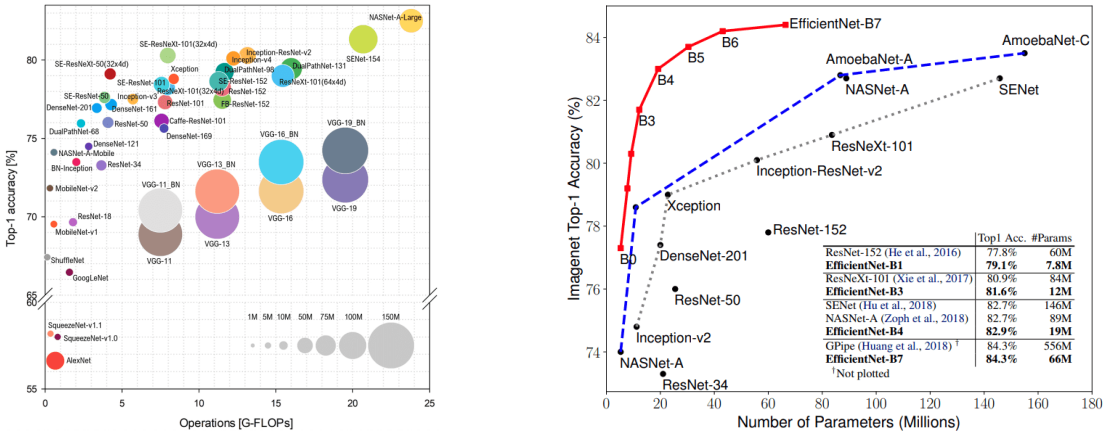


Figure 48: Different CNN Arquitectures Model Performance.

3.2.2. JUSTIFICATION FOR MODEL SELECTION

The choice of CNN architectures can be based on several factors:

I. **Performance**: ResNet, VGG, and EfficientNet have consistently shown high performance in image classification benchmarks, making them suitable for the detection of PE in CT scans. U-Net's success in medical image segmentation tasks makes it a strong candidate for localizing clots.

II. **Architecture Depth**: ResNet deep architecture allows it to learn complex features, which is crucial for accurately identifying PE. EfficientNet compound scaling approach provides a balance between depth, width, and resolution, leading to efficient and effective model training.

III. **Segmentation Capability**: U-Net's architecture is designed for segmentation tasks, making it ideal for detecting and localizing clots within the CT scans. Its ability to produce detailed segmentation maps is essential for identifying the precise location of PE.

IV. **Efficiency**: EfficientNet ability to achieve high accuracy with fewer parameters is particularly beneficial for medical imaging tasks, where computational resources and efficiency are critical. This makes it a suitable choice for large-scale deployment in clinical settings.

Preliminary experiments and literature reviews indicated that these models are well-suited for the task, providing a balance between accuracy, computational efficiency, and the ability to generalize across different datasets.

Finally, upon diverse test with Resnet18, Resnet50 and Efficientnet, Efficientnet-B0 has been selected due the computational time efficiency due the limitation of GPU time in the Kaggle platform and the model has been selected also for the accuracy compared to Resnet architectures, and the version B0 has been selected because the input of this version is the same of the input image size (256 pixels).

3.3. TRAINING THE MODEL

The training of the deep learning models for the detection of pulmonary embolism (PE) involves a systematic approach to optimize performance through appropriate data handling, training processes, and rigorous evaluation. This section outlines the detailed steps involved in training the model, including hyperparameter tuning and the metrics used for evaluation.

The training process is conducted in several stages, from data loading and preparation to model definition and training. Below is an explanation of the training process.

3.3.1. DATASET PREPARATION

Initially, the dataset was read and processed. The original dataset was in CSV format and contained multiple CT images with their respective labels, indicating the presence or absence of pulmonary embolism in each image.

```
train_csv_path = '/kaggle/input/df-filtered-after-
eda/final_df_Pulmonary_Embolism_CT_scans_data.csv'
jpeg_dir = '/kaggle/input/rsna-str-pe-detection-jpeg-256/train-jpegs'

train_df = pd.read_csv(train_csv_path)
train_df.head()
```

Filtering and Subsampling Images

Since a series of images can contain multiple CT slices for a single patient, a filtering step was implemented to reduce the number of images per patient. This filtering retained only 80% of the central images of each series, discarding 10% from both extremes, to focus on the most representative images of the series.

```
def filter_slices(group):
    n = len(group)
    start = int(0.1 * n)
    end = int(0.9 * n)
    return group.iloc[start:end]

train_df_80 =
train_df.groupby('SeriesInstanceUID').apply(filter_slices).reset_index(drop=True)
```

Subsequently, the dataset was balanced by randomly subsampling the classes to have an equal number of images with and without pulmonary embolism. This step is crucial to avoid bias in the model training due to a dominant class.

```
df_positive_downsampled = resample(df_positive, replace=False, n_samples=72500,
random_state=42)
df_negative_downsampled = resample(df_negative, replace=False, n_samples=72500,
random_state=42)

df_balanced = pd.concat([df_positive_downsampled, df_negative_downsampled])
df_balanced = df_balanced.sample(frac=1, random_state=42).reset_index(drop=True)
```

3.3.2. IMAGE AUGMENTATION AND PREPROCESSING

Data augmentation is a fundamental technique in deep learning, especially when working with images. It is used to increase the diversity of the dataset and improve the model's generalization by simulating possible variations in the test images. For this purpose, various augmentation techniques were implemented using the albumentations library:

Horizontal and Vertical Flips: These random flips allow the model to learn the features of pulmonary embolism without depending on the specific orientation of the image.

Random Rotation: A random rotation of up to 20 degrees was applied to simulate different possible orientations of the images.

Random Resized Crop: The RandomResizedCrop method was used to perform random cropping followed by resizing to the original dimensions, helping the model focus on different parts of the image.

Elastic Transformation and Grid Distortion: These techniques deform the images non-linearly, helping the model become robust to structural variations.

Grayscale Conversion: Finally, the images were converted to grayscale, as CT images are traditionally monochromatic.

```
def get_training_augmentation(y=256, x=256):
    train_transform = [
        albu.VerticalFlip(p=0.5),
        albu.HorizontalFlip(p=0.5),
        albu.Rotate(limit=20, p=0.5),
        albu.RandomResizedCrop(height=y, width=x, scale=(0.8, 1.0), ratio=(0.75,
1.33), p=0.5),
        albu.ElasticTransform(alpha=1, sigma=50, alpha_affine=50, p=0.5),
        albu.GridDistortion(p=0.3),
        albu.Downscale(p=0.5, scale_min=0.35, scale_max=0.75),
        albu.ToGray(p=1.0),
        albu.Resize(y, x)
    ]
    return albu.Compose(train_transform)
```

For validation, a simpler augmentation strategy was used, consisting of a central crop followed by resizing. This ensures that the validated images are centered and have the same scale, allowing for a more consistent evaluation of the model.

```
def get_validation_augmentation(y=256, x=256):
    test_transform = [
        albu.CenterCrop(height=int(y * 0.8), width=int(x * 0.8), p=1.0),
        albu.Resize(y, x)
    ]
    return albu.Compose(test_transform)
```

3.3.3. MODEL DEFINITION AND TRAINING CONFIGURATION

For this project, EfficientNet-B0 was selected, a convolutional neural network (CNN) architecture that has shown excellent performance in various image classification tasks with high computational efficiency. The model was initialized with pretrained weights on ImageNet to leverage prior knowledge in generic visual classification tasks.

```
from efficientnet_pytorch import EfficientNet
model = EfficientNet.from_pretrained('efficientnet-b0')
num_ftrs = model._fc.in_features
model._fc = nn.Linear(num_ftrs, 1)
model = model.cuda()
```

The last fully connected layer of the model was replaced to adapt to the binary classification task (presence or absence of PE). The BCEWithLogitsLoss function was used as the loss function, which is suitable for binary classification problems.

```
loss_fn = torch.nn.BCEWithLogitsLoss()
```

For the optimizer, Adam was used with an initial learning rate of 0.001, combined with a learning rate scheduling policy using CosineAnnealingLR, which gradually reduces the learning rate as training progresses, promoting model convergence.

```
optimizer = torch.optim.Adam(model.parameters(), lr=0.001, weight_decay=0.00001)
scheduler = torch.optim.lr_scheduler.CosineAnnealingLR(optimizer, T_max=300,
eta_min=0.00001)
```

The decision made to select this parameters (learning rate, weight decay, T_max) where by testing and analyzzing results and timing also observing different projects in Kaggle Community.

3.3.4. TRAINING AND VALIDATION

To efficiently manage the training and validation processes of the deep learning model, a custom class named trainer was implemented. This class encapsulates all the logic necessary for training the model on the training data, evaluating it on the validation set, and updating the model's parameters accordingly.

Class trainer Overview

The trainer class is designed to handle the end-to-end workflow of training a neural network. This includes processing batches of images, calculating loss, performing backpropagation, updating model parameters, and adjusting the learning rate. By organizing the training logic into a class, the code becomes more modular, reusable, and easier to maintain or extend.

Here is an in-depth breakdown of the trainer class components:

```
class trainer:
    def __init__(self, loss_fn, model, optimizer, scheduler):
        self.loss_fn = loss_fn
        self.model = model
        self.optimizer = optimizer
        self.scheduler = scheduler
```

Initialization (__init__ method)

The __init__ method is the constructor of the class and is responsible for initializing the core components required for training. This method takes four key arguments:

loss_fn: This is the loss function used to calculate the discrepancy between the model's predictions and the actual labels. In this project, the BCEWithLogitsLoss function was used, which combines a sigmoid layer and binary cross-entropy loss in one function, making it ideal for binary classification tasks.

model: The deep learning model itself, which in this case is an instance of EfficientNet-B0, initialized with pre-trained weights and modified for binary classification.

optimizer: The optimizer is responsible for updating the model parameters based on the gradients computed during backpropagation. The Adam optimizer was chosen due to its efficiency and adaptive learning rate capabilities.

scheduler: A learning rate scheduler, specifically CosineAnnealingLR, is used to adjust the learning rate during training. This scheduler reduces the learning rate according to a cosine function, which can help in fine-tuning the model as training progresses.

Training Epoch (train_epoch method)

The train_epoch method handles the training process for one complete pass (epoch) through the training dataset. This method iterates over the dataset in batches, updating the model's weights using backpropagation. Here's a step-by-step explanation:

```
def train_epoch(self, loader):
    self.model.train()
    tqdm_loader = tqdm(loader)
    current_loss_mean = 0
    for batch_idx, (imgs, labels) in enumerate(tqdm_loader):
        loss, predicted = self.batch_train(imgs, labels, batch_idx)
        current_loss_mean = (current_loss_mean * batch_idx + loss) / (batch_idx + 1)
        tqdm_loader.set_description('loss: {:.4} lr:{:.6}'.format(
                current_loss_mean, self.optimizer.param_groups[0]['lr']))
        self.scheduler.step(batch_idx)
    return current_loss_mean
```

Set Model to Training Mode: self.model.train() sets the model to training mode, enabling features like dropout and batch normalization, which behave differently during training compared to evaluation.

Batch Iteration: The method uses a for loop to iterate through the data loader (loader), which provides the training data in batches. Each batch contains a set of images (imgs) and their corresponding labels (labels).

Batch Training: For each batch, the batch_train method is called to perform the forward pass, calculate the loss, and update the model's weights.

Loss Calculation: The average loss for the epoch is calculated dynamically. current_loss_mean tracks the mean loss across all batches, giving a running average as training progresses.

Progress Tracking: The tqdm library is used to provide a progress bar for the training loop, which displays the current loss and learning rate, helping to monitor the training process in real-time.

Learning Rate Adjustment: self.scheduler.step(batch_idx) adjusts the learning rate at each batch iteration, following the cosine annealing schedule. This helps in reducing the learning rate gradually, which can improve convergence.

Return: At the end of the epoch, the method returns the average loss for the epoch, which is used to monitor training progress and to compare with validation performance.

Batch Training (batch_train method)

Although not explicitly shown in the previous code snippet, the batch_train method is a crucial part of the training loop. This method performs the forward and backward passes for a single batch of data.

```
def batch_train(self, imgs, labels, batch_idx):
    self.optimizer.zero_grad()  # Clear previous gradients
    batch_imgs = imgs.cuda().float()  # Transfer images to GPU
    batch_labels = labels.cuda()  # Transfer labels to GPU
    predicted = self.model(batch_imgs)  # Forward pass
    loss = self.loss_fn(predicted.float(), batch_labels.float())  # Compute loss
    loss.backward()  # Backward pass
    self.optimizer.step()  # Update model parameters
    return loss.item(), predicted
```

Gradient Clearing: self.optimizer.zero_grad() clears any previously accumulated gradients. This is necessary because gradients are accumulated by default in PyTorch, and if not cleared, they would sum up with each new batch.

Data Transfer to GPU: Both images and labels are transferred to the GPU using .cuda(), which ensures that all operations are performed on the GPU for faster computation.

Forward Pass: self.model(batch_imgs) performs a forward pass through the model, generating predictions for the current batch of images.

Loss Computation: The predicted outputs are compared with the true labels using the loss function (self.loss_fn), which calculates the error (loss) between the predicted values and the ground truth.

Backward Pass: loss.backward() computes the gradient of the loss with respect to the model's parameters. This is the core of the backpropagation algorithm, which allows the optimizer to know how to adjust the model's weights to minimize the loss.

Parameter Update: self.optimizer.step() updates the model's parameters based on the computed gradients, effectively learning from the data.

Return: The method returns the loss for the current batch (used to calculate the average loss for the epoch) and the predicted values (which could be used for further analysis if needed).

Validation Epoch (valid_epoch method)

The valid_epoch method is similar to train_epoch but is used for evaluating the model's performance on the validation dataset. It differs in several key aspects:

```
def valid_epoch(self, loader, name="valid"):
    self.model.eval()
    tqdm_loader = tqdm(loader)
    current_loss_mean = 0
    for batch_idx, (imgs, labels) in enumerate(tqdm_loader):
        with torch.no_grad():
            batch_imgs = imgs.cuda().float()
            batch_labels = labels.cuda()
            predicted = self.model(batch_imgs)
            loss = self.loss_fn(predicted.float(), batch_labels.float()).item()
            current_loss_mean = (current_loss_mean * batch_idx + loss) / (batch_idx +
1)
    score = 1 - current_loss_mean
    print('metric {}'.format(score))
    return score
```

Evaluation Mode: self.model.eval() sets the model to evaluation mode. This disables dropout and batch normalization, ensuring that the model's behavior is consistent during validation.

No Gradient Computation: The with torch.no_grad(): block ensures that no gradients are computed during validation. This reduces memory usage and speeds up the validation process since the model parameters are not being updated.

Loss Calculation: Similar to the training loop, the loss is calculated for each batch and averaged across all batches to obtain the overall validation loss.

Score Calculation: A simple metric, score = 1 - current_loss_mean, is calculated. This is used to monitor validation performance, with lower losses indicating better performance.

Return: The method returns the validation score, which is used to determine if the model is improving and to potentially trigger early stopping if performance degrades.

Running the Training Process

Finally, the training process is executed across multiple epochs. During each epoch, the model is trained on the training set (train_epoch) and evaluated on the validation set (valid_epoch). The best model (based on validation performance) is saved for future inference or testing.

```
Trainer = trainer(loss_fn, model, optimizer, scheduler)
Trainer.run(train, val)
```

The trainer class's methods are called in a loop across several epochs, and additional features like early stopping, model checkpointing, and logging can be integrated into this process as needed.

3.3.4. TRAINING RESULTS

Training is conducted over a predefined number of epochs (epochs=10). In each epoch, the model is trained using the training set and then validated using the validation data.

For each epoch training time is 45-60 minutes each one making the entire training process 450 – 600 minutes overall, wich is the same, 7.5 – 10 hours.

During training, both the loss and the accuracy of the model were monitored on the training and validation sets to ensure that the model was learning effectively and to avoid overfitting. The best model was selected based on its performance on the validation set, which showed consistent improvement across the epochs.

3.3.1. EVALUATION METRICS AND VALIDATION TECHNIQUES

In this section, the performance of the trained model is evaluated using various metrics and validation techniques to assess how well the model generalizes on unseen data. The process can be summarized in the following steps:

1) Loading the Best Model

The model that performed best during training is loaded before evaluation. This ensures that the evaluation metrics are calculated using the most optimal model, which has been saved based on validation performance.

2) Evaluation Strategy

The evaluation process involves generating predictions for the validation dataset and comparing them to the true labels. The model is set to evaluation mode, disabling layers like dropout and batch normalization to maintain consistent behavior during inference. Predictions are generated for the validation set, and both the predictions and true labels are stored for further analysis.

3) Classification Report and Confusion Matrix

After obtaining the predictions, they are binarized (using a threshold of 0.5) to match the format of multi-label classification. The classification report provides detailed metrics such as precision, recall, F1-score, and support for each class. These metrics are critical in understanding how well the model is performing on each category individually. Additionally, a confusion matrix is plotted to visually represent the classification results, showing the distribution of correct and incorrect predictions across different classes. This matrix helps identify common misclassifications and areas where the model may be struggling.

4) ROC Curve and AUC (Area Under the Curve)

For a more comprehensive evaluation, the ROC (Receiver Operating Characteristic) curve and AUC are calculated. The ROC curve is particularly useful in binary classification tasks, as it plots the true positive rate against the false positive rate across various thresholds. The AUC provides a single value representing the model's overall performance, with higher values indicating better discrimination ability. The curve and AUC help assess the trade-off between sensitivity and specificity, providing insights into how the model behaves under different classification thresholds.

Together, these evaluation techniques give a comprehensive view of how well the model generalizes to new data, ensuring that performance is analyzed from multiple perspectives, which is crucial for high-stakes applications like medical image classification.

4. IMPLEMENTATION

In this section, I describe the technical aspects of implementing the deep learning models for detecting pulmonary thromboembolism (PE) in CT scans. I cover the software and hardware used, the specific implementation steps taken, and the challenges encountered along with the solutions adopted to address them.

4.1. TECHNICAL DETAILS

4.1.1. SOFTWARE AND HARDWARE USED

We used Kaggle as our environment due there we had all the dataset uploaded and we worked with the GPU P100 and 2 GPU T4 from Kaggle, but always time limited to 30h per week on GPU computational time.

4.1.2. IMPLEMENTATION STEPS

The implementation process consisted of several key steps:

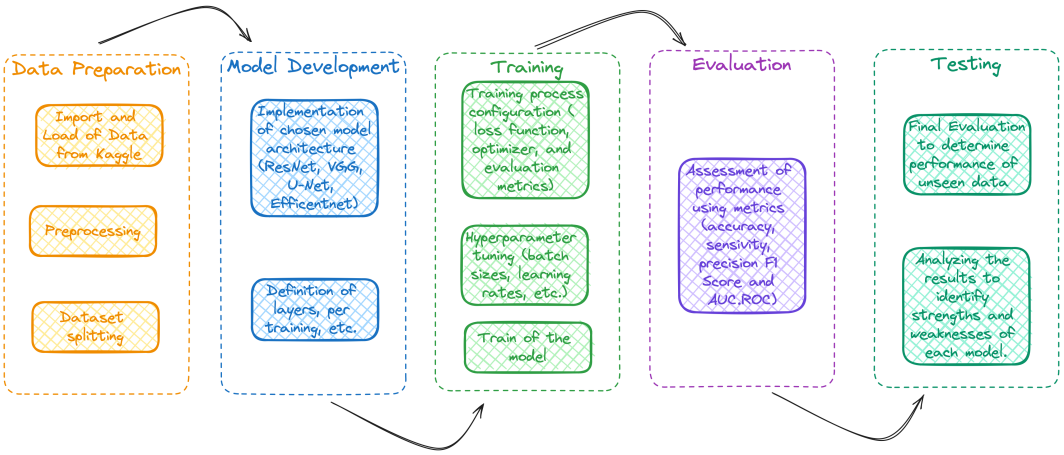


Figure 49: Implementation steps schema.

1) **Data Preparation**
   - Importing and loading the CT scan dataset from Kaggle.
   - Preprocessing the images, including normalization, resizing, and augmentation.
   - Splitting the dataset into training, validation, and test sets.
2) **Model Development**
   - Implementing the chosen deep learning architectures (ResNet, VGG, U-Net, EfficientNet) using PyTorch.
   - Defining the model layers, activation functions, and other architectural components.
3) **Training**
   - Configuring the training process, including setting the loss function, optimizer, and evaluation metrics.
   - Performing hyperparameter tuning to optimize learning rates, batch sizes, and other parameters.
   - Training the models on the GPU resources provided by Kaggle, ensuring efficient use of the 30-hour weekly limit.
4) **Evaluation**
   - Assessing the performance of the models on the validation set using metrics such as accuracy, sensitivity, specificity, precision, F1 score, and AUC-ROC.
5) **Testing**
   - Evaluating the final models on the test set to determine their performance on unseen data.
   - Analyzing the results to identify strengths and weaknesses of each model.

4.1.3. MODEL ARCHITECTURE

In this section, we describe the architecture of the model used for the project, which is EfficientNet-B0. EfficientNet is a family of convolutional neural networks that is designed to achieve high accuracy while maintaining computational efficiency. The model family, particularly EfficientNet-B0, is well-suited for image classification tasks due to its balance of performance and resource requirements.

4.1.3.1. OVERVIEW OF EFFICIENTNET-B0

EfficientNet-B0 is the base model in the EfficientNet family, designed with a specific focus on scaling the model in an optimal way. The scaling approach used in EfficientNet is known as compound scaling, which uniformly scales the depth (number of layers), width (number of channels), and resolution (input image size) of the network in a balanced manner. This allows EfficientNet-B0 to achieve better performance with fewer parameters and lower computational cost compared to traditional models like ResNet or VGG.

4.1.3.2. MODIFICATIONS FOR THE PROJECT

For this specific project, the following adjustments were made to the original EfficientNet-B0 model:

- The fully connected layer was modified to output the desired number of classes, allowing the model to handle multi-label classification for medical images.
- Data augmentation and preprocessing steps were carefully integrated to maximize model generalization, taking advantage of EfficientNet's ability to handle diverse input variations.

Finally, EfficientNet-B0 architecture is designed for a balance between efficiency and accuracy, making it an ideal choice for this project where computational resources are limited, and high performance is crucial. Its sophisticated use of compound scaling, inverted residual blocks, and other modern techniques allow it to outperform many traditional models while being computationally lightweight.

4.2. CHALLENGE AND SOLUTIONS

4.2.1. OBSTACLES FACED DURING IMPLEMENTATION

Several challenges were encountered during the implementation process:

- **Limited GPU Time:** The 30-hour weekly limit on GPU usage on Kaggle posed a significant constraint, especially for training deep learning models that require extensive computational resources.
- **Data Imbalance:** The dataset had an imbalance between PE and non-PE cases, which could lead to biased model performance favoring the majority class.
- **Hyperparameter Tuning:** Finding the optimal combination of hyperparameters was challenging due to the vast search space and the computational limitations.
- **Overfitting:** There was a risk of overfitting, especially with complex models like ResNet and EfficientNet, where the model might perform well on the training data but poorly on unseen data.
- **Original Data size:** The original dataset containing DICOM image files was 980.24 GB which considerable obstacles the time and efforts in order to train efficiently baseline models.

4.2.2. SOLUTIONS AND ADJUSTMENTS MADE

To address these challenges, the following solutions and adjustments were implemented:

- **Efficient Use of GPU Time:** Training processes were carefully planned and executed to make the most of the limited GPU time. This included conducting preliminary experiments to determine the best model configurations before extensive training and utilizing checkpointing to save progress and avoid re-training from scratch. As part of the efficient use of GPU I choose the Efficientnet architecture due the efficiency training time it has and the ratio time-accuracy.
- **Filtering Optimization Techniques:** Knowing that the average number of images per patient is 200, we observed that normally PE appears in the middle layers, knowing that a filter to only train the 80% of the mid slice Z positions were applied to improve training time and better results.

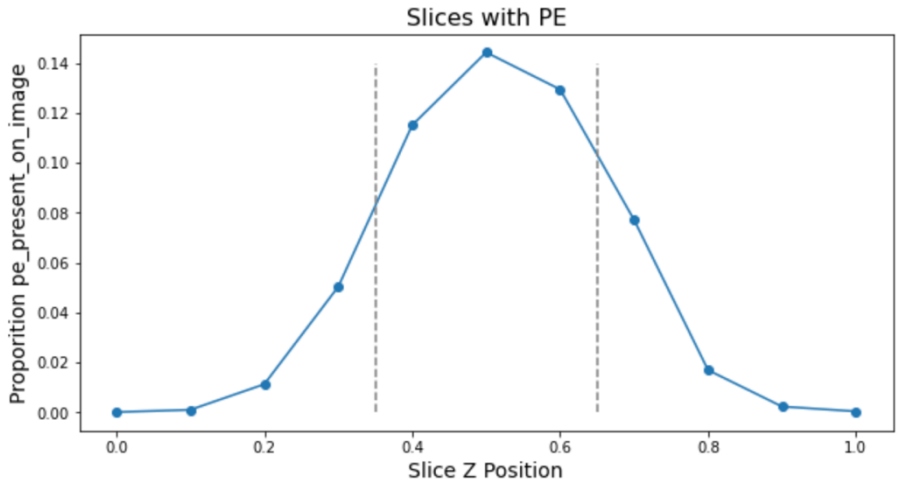


Figure 50: Average Proportion of presence of PE in Slices per patient. Source: (*https://www.kaggle.com/competitions/rsna-str-pulmonary-embolism-detection/discussion/193402*)

- **Data Augmentation and Resampling:** To mitigate data imbalance, data augmentation techniques such as rotations, flips, and shifts were applied to increase the variability of the minority class. Additionally, resampling (cropping the 80% of the image) techniques were employed to ensure a more balanced distribution of PE and non-PE cases.
- **Usage of JPG Data extracted from the original Challenge Dataset:** To optimize processing and reduce workload, the images were converted to JPG format and resized to 256x256 pixels, reducing the dataset size to 56.14 GB. The data was already uploaded by a member of Kaggle community [Add reference]. This adjustment significantly decreased processing time while maintaining relevant data for analysis.
- **Dataset Balancing:** To optimize training accuracy, balancing for 72500 on each class '`pe_present_on_image`' has been made. The decision to consider this value is because is the maximum value of true in `pe_present_on_image`'.

Through these strategies, we were able to overcome the challenges and develop effective deep learning models for the detection of pulmonary thromboembolism, ensuring accurate and reliable results.

This detailed implementation section outlines the technical aspects and challenges encountered, providing a comprehensive overview of the processes involved in developing the deep learning models with the use of PyTorch.

4.3. OVERALL MODEL SCHEMA

The final model schema, developed to address the challenges and incorporate various solutions, is outlined as follows:

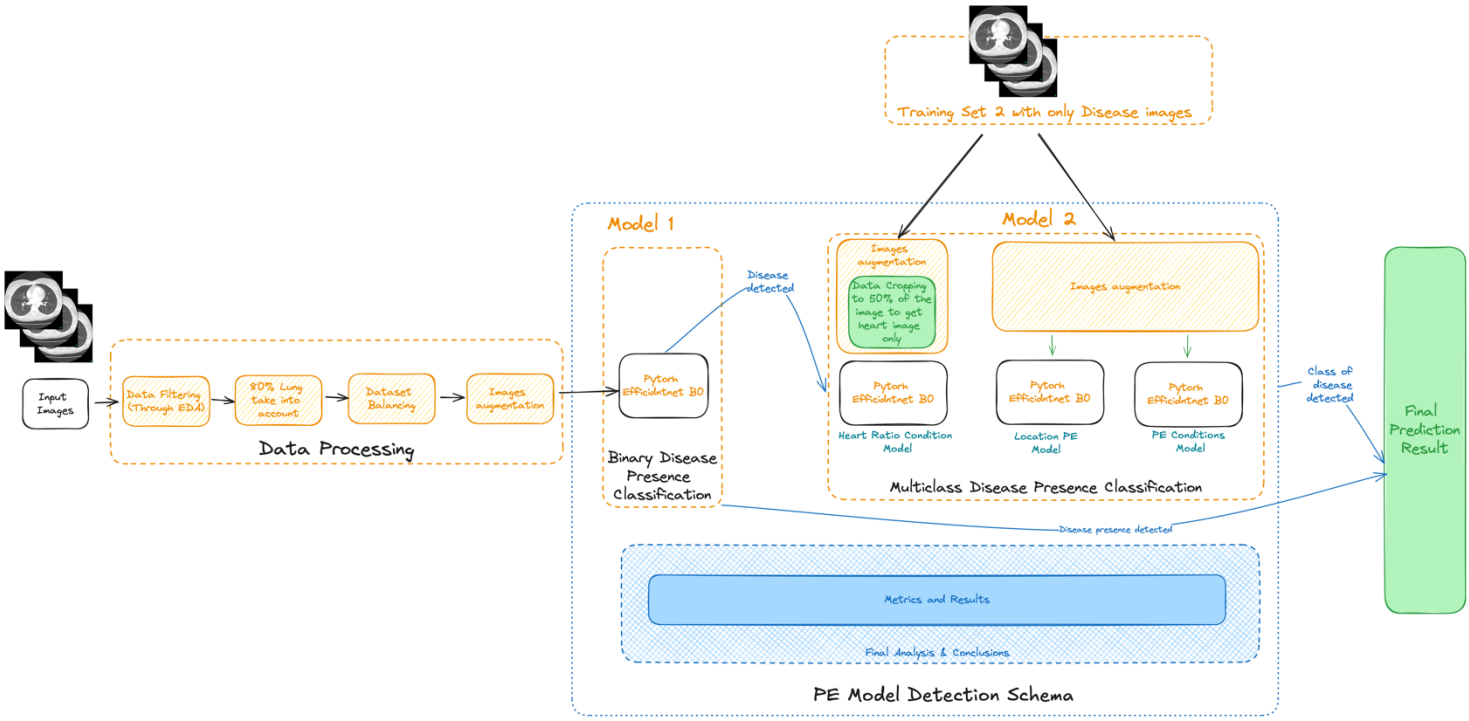


Figure 51: Overall Model Schema

1) Input Data

The process begins with the JPG images of size 256 x 256 pixels.

2) Model Architecture

The schema comprises two main Eficcientnet-B0 models:

- Binary Classification Model: This model determines the presence of disease in the image at exam level.

- Multiclass Classification Model: This model, applied only to images where disease is detected, identifies the specific location and type of pulmonary embolism (PE) at image level.

3) Data Preparation

- Filtering and Reduction: Initially, data is filtered based on the Exploratory Data Analysis phase and cleaned (referenced in Point 3.1.2.3).
- Middle Slice Selection: To optimize model training, only 80% of the middle slices from each patient are used.
- Dataset Balancing: The dataset is balanced to include 72,500 images per class (totaling 145,000 images) from 6,644 patients.

4) Data Augmentation

Training data undergoes various augmentations, including rotations, flips, resized crops, elastic transformations, grid distortions, downscaling, and grayscale transformations.

5) Multiclass Classification

For the multiclass classification in order to obtain better results, the strategy followed was to divide the multiclass classification in 3 different models based on their features proximity.

For this, the first multiclass model was created to predict heart ratio conditions, the second model was developed to predict PE location and the last third model was developed to predict the condition of the PE such if the PE is acute or acute and chronic.

6) Evaluation

Each model performance is assessed using key metrics such as accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics provide insights into how well the models distinguish between different classes and their overall effectiveness in predicting PE.

5. RESULTS

5.1. MODEL PERFORMANCE

In this section, we evaluate the performance of the machine learning models developed for predicting PE. The analysis focuses on two primary models: a binary classification model and a multiclass classification model.

The evaluation is conducted on both the validation and test datasets to ensure that the models are not only accurate on the data they were trained on but also generalize well to unseen data. Additionally, we compare the results of these models with traditional diagnostic methods to understand the potential impact of machine learning in clinical settings. The binary classification model focuses on distinguishing between the presence and absence of PE, while the multiclass model aims to predict specific features and conditions associated with the disease.

This comprehensive performance evaluation is critical for determining the models' readiness for real-world application and identifying areas for improvement in future iterations.

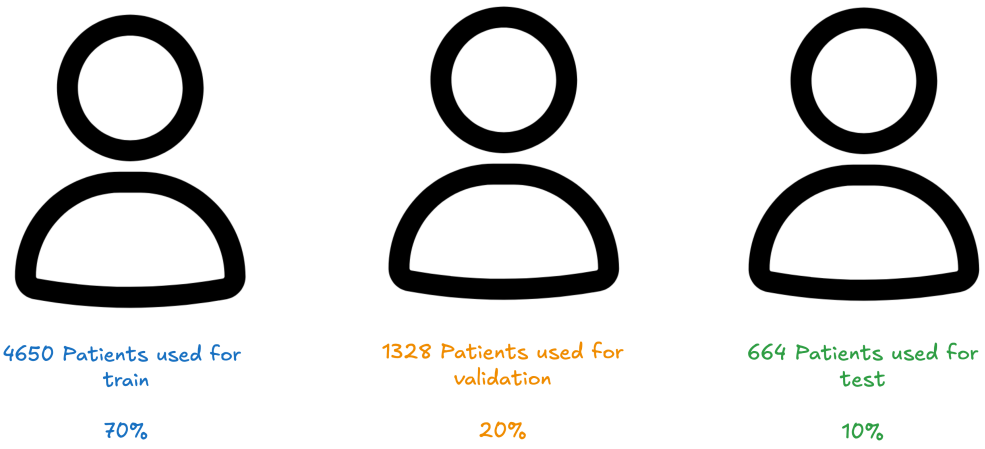


Figure 52: Data Splitting Performed for Binary Model.

5.1.1 EXAM LEVEL BINARY MODEL RESULT

The binary classification model was developed to predict the presence or absence of pulmonary embolism (PE). The performance of the model was evaluated using both the validation set (20% of the data) and the test set (10% of the data). Below are the detailed metrics:

```
Classification Report:
              precision    recall  f1-score   support

     Class 0       0.74      0.75      0.74     14270
     Class 1       0.74      0.73      0.74     13985

    accuracy                           0.74     28255
   macro avg       0.74      0.74      0.74     28255
weighted avg       0.74      0.74      0.74     28255
```

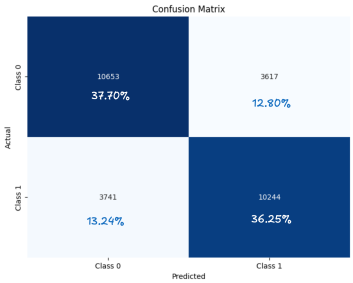


Figure 53: Classification Report 20% Validation Results Binary Model

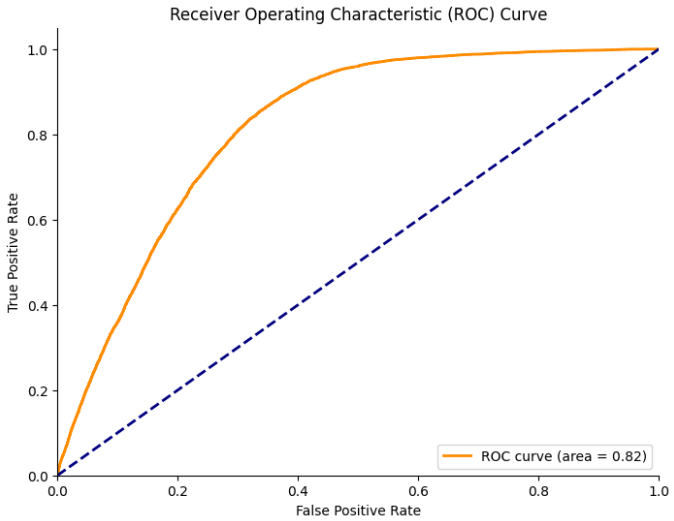


Figure 54: AUC ROC Curve 20% Validation Results Binary Model.

These results indicate that the model has a balanced ability to correctly classify instances of PE and non-PE, with a slight edge towards correctly identifying non-PE cases, as suggested by the higher recall for Class 0 - no presence of PE.

```
Classification Report:
              precision    recall  f1-score   support

     Class 0       0.72      0.74      0.73      7409
     Class 1       0.71      0.68      0.69      6697

    accuracy                           0.71     14106
   macro avg       0.71      0.71      0.71     14106
weighted avg       0.71      0.71      0.71     14106
```

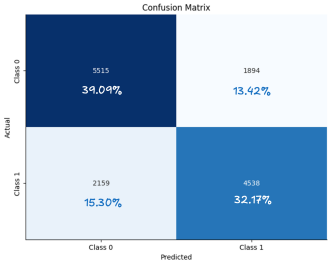


Figure 55: Classification Report 10% Test Results Binary Model.

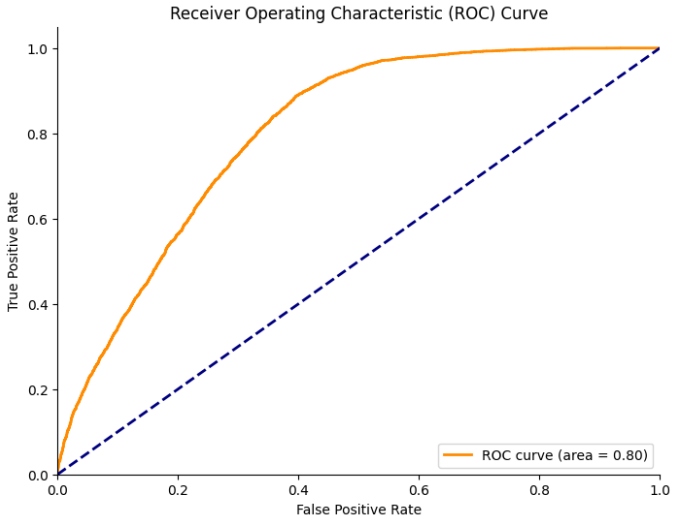


Figure 56: AUC ROC Curve 10% Test Results Binary Model.

The test set results are slightly lower than the validation set, which is not uncommon in machine learning models.

The accuracy of 71% in testing an 74% in validation and an AUC of 0.80 in testing and 0.82 in validation and suggest that the model performs consistently but shows a slight drop in generalizability when tested on unseen data. The lower recall for Class 1 (Presence of PE) in the test set indicates that the model might be more conservative in detecting PE, potentially leading to missed cases, which is a critical consideration in medical applications.

5.1.1.1. COMPARISION WITH TRADITIONAL DIAGNOSTIC METHODS

In clinical practice, the diagnosis of PE typically involves a combination of imaging techniques (like CT pulmonary angiography), clinical assessment, and D-dimer testing. These methods rely heavily on the expertise of radiologists and clinicians, which can introduce variability in diagnosis due to differences in experience and interpretation.

The binary classification model provides a standardized and consistent approach to predicting PE, with an accuracy of approximately 71-74% and an AUC between 0.80 and 0.82. While these results are promising, they suggest that the model could potentially complement but not yet replace the expertise of radiologists.

In particular, the model's ability to quickly flag potential PE cases with a consistent AUC performance could be valuable in high-pressure situations where rapid diagnosis is critical.

Moreover, the model's predictions could serve as a second opinion to reduce human error, particularly in ambiguous cases. However, the relatively lower recall for PE in the test set highlights the need for further refinement before this model could be considered reliable enough for standalone use in clinical settings. It emphasizes the importance of integrating such models into a broader diagnostic workflow where human expertise and machine learning predictions work hand in hand to ensure accurate and timely diagnosis.

5.1.2 MULTICLASS IMAGE LEVEL MODEL RESULTS

To improve the performance of multiclass classification, we adopted a strategy that involved segmenting the classification task into three distinct models based on the similarity of their features.

The first model was dedicated to predicting heart ratio conditions, focusing on the categories `'rv_lv_ratio_gte_1'` and,`'rv_lv_ratio_lt_1'`. The second model aimed to classify the location of pulmonary embolism (PE), distinguishing among `'leftsided_pe'`, `'rightsided_pe'`, and `'central_pe'`. Finally, the third model was designed to predict the condition of the PE, identifying whether it is `'chronic_pe'` or `'acute_and_chronic_pe'`. This approach allowed us to target specific prediction areas more effectively and improve overall classification accuracy.

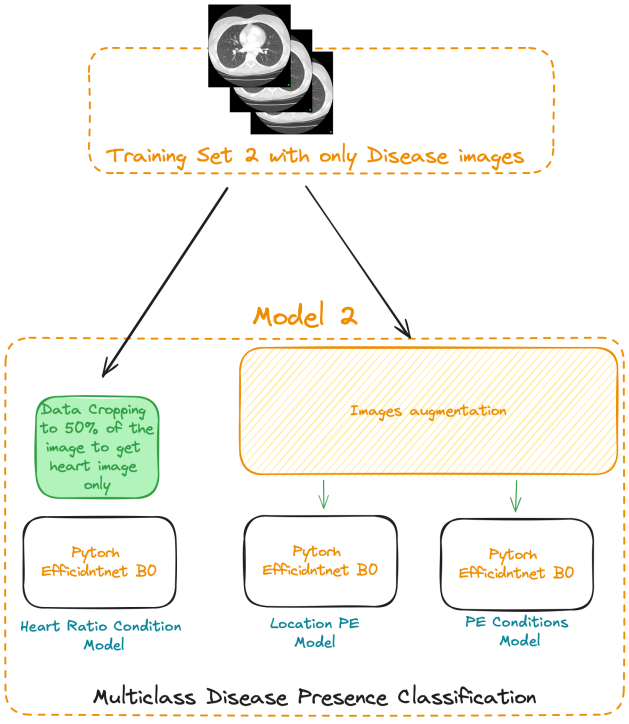


Figure 57: Multiclass Model performed using 3 models for agrupation of categories at exam level.

The train-test-validation split was applied uniformly across all three models, maintaining a consistent ratio of patients: 70% for training, 20% for validation, and 10% for testing. These models were specifically trained using only disease images, as their purpose is to operate at the exam level and predict exam features. In total, the training set comprised 1,447 patients, the validation set included 413 patients, and the testing set contained 224 patients.

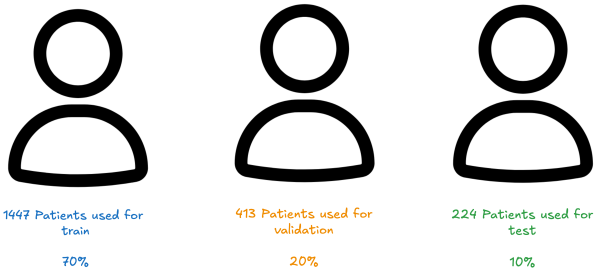


Figure 58: Train-test-validation split ratio and numbers of patients for the multiclass model.

In order to optimize training timing in the multiclassificator, we reduced the training time to 5 epochs to all the models here. The performance for this training where lower than the first binary model, obtaining a training time of aprox. 12 minutes per epoch.

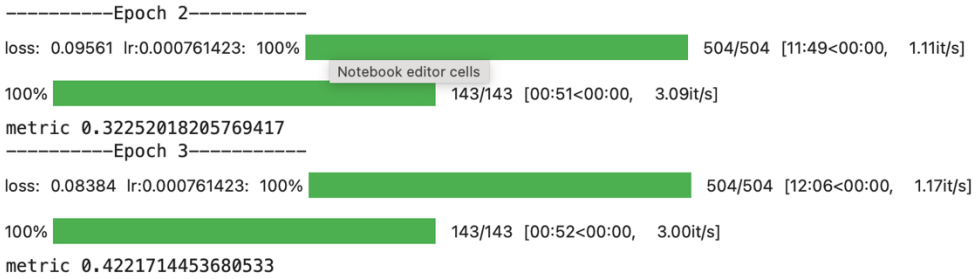


Figure 59: Example of time of exectuion per epoch in one of the multiclass models.

5.1.2.1. MULTICLASS MODEL 1 - PREDICTING LOCATION OF PE (LEFTSIDED, RIGHTSIDED, CENTRAL)

The first model was developed and trained to predict location in order to benefit diagnosis localization. This model predicts three different classes based on where the PE is located.

MULTICLASS MODEL 1 TESTING RESULTS (10% OF THE DISEASED SAMPLE)

After training and predicting with testing, the results of the model one are the following:

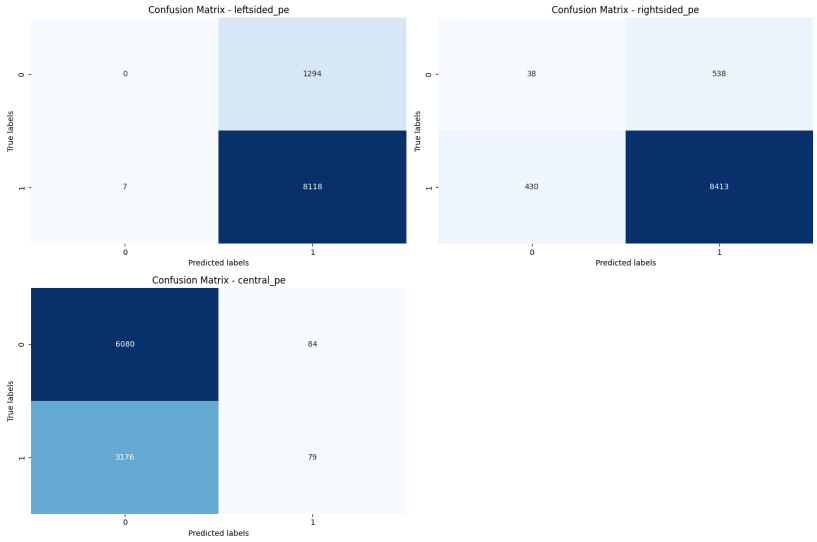


Figure 60: Confusion Matrix Results for first model of the multiclassificator model for test sample.

	precision	recall	f1-score	support
leftsided_pe	0.86	1.00	0.93	8125
rightsided_pe	0.94	0.95	0.95	8843
central_pe	0.48	0.02	0.05	3255
micro avg	0.90	0.82	0.86	20223
macro avg	0.76	0.66	0.64	20223
weighted avg	0.84	0.82	0.79	20223
samples avg	0.89	0.86	0.85	20223

Figure 61: Classification Report Results for first model of the multiclassificator model for test sample.

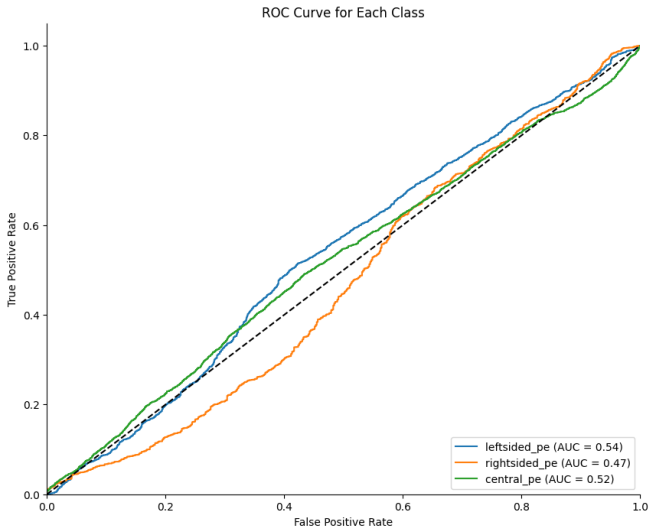


Figure 62: AUC ROC Curve Results for first model of the multiclassificator model for test sample.

MULTICLASS MODEL 1 VALIDATION RESULTS (20% OF THE DISEASED SAMPLE)

In order to verify the predictions we predicted using the validation sample, the results obtained are the following:

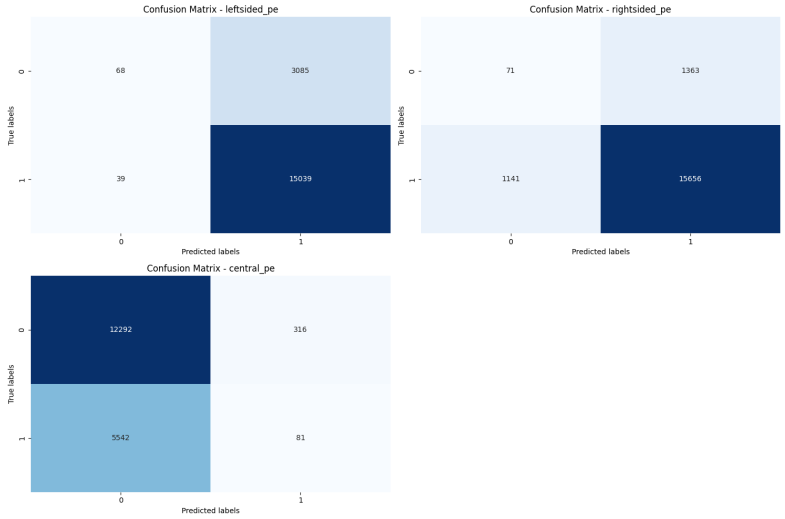


Figure 63: Confusion Matrix Results for first model of the multiclassificator model for validation sample.

	precision	recall	f1-score	support
leftsided_pe	0.83	1.00	0.91	15078
rightsided_pe	0.92	0.93	0.93	16797
central_pe	0.20	0.01	0.03	5623
micro avg	0.87	0.82	0.84	37498
macro avg	0.65	0.65	0.62	37498
weighted avg	0.78	0.82	0.78	37498
samples avg	0.87	0.86	0.83	37498

Figure 64: Classification Report Results for first model of the multiclassificator model for validation sample.

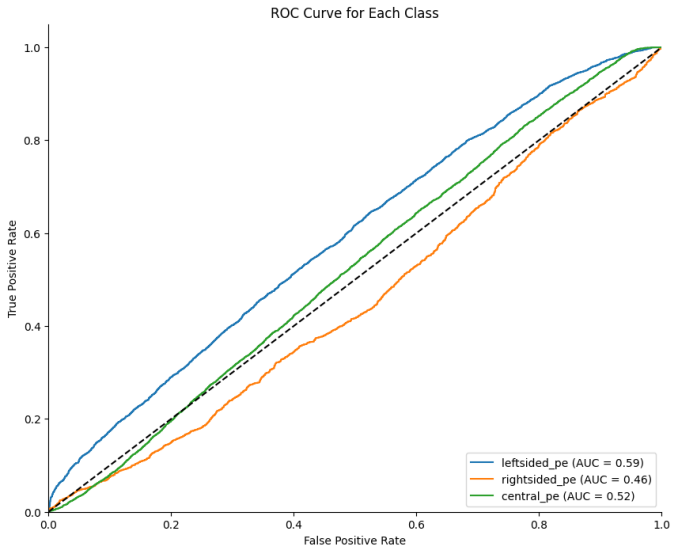


Figure 65: AUC ROC Curve Results for first model of the multiclassificator model for validation sample.

5.1.2.2. MULTICLASS MODEL 2 - PREDICTING DISEASE CONDITION OF PE (CHRONIC PE / ACUTE AND CHRONIC PE)

The second model was developed and trained to predict disease condition in order to benefit detection of critics diagnosis . This model predicts two different classes based on where the PE is chronic or chronic and acute.

MULTICLASS MODEL 2 TESTING RESULTS (10% OF THE DISEASED SAMPLE)

In order to verify the algorithm performance we predicted using the testing sample, the results obtained are the following:

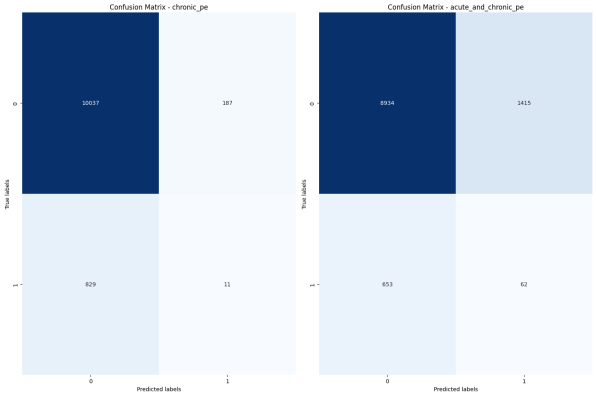


Figure 66: Confusion Matrix Results for second model of the multiclassificator model for test sample.

	precision	recall	f1-score	support
chronic_pe	0.06	0.01	0.02	840
acute_and_chronic_pe	0.04	0.09	0.06	715
micro avg	0.04	0.05	0.05	1555
macro avg	0.05	0.05	0.04	1555
weighted avg	0.05	0.05	0.04	1555
samples avg	0.01	0.01	0.01	1555

Figure 67: Classification Report Results for second model of the multiclassificator model for test sample.

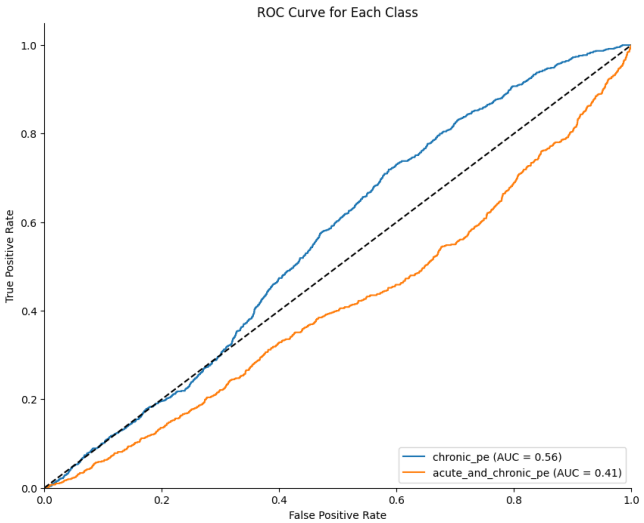


Figure 68: AUC ROC Curve Results for second model of the multiclassificator model for test sample.

MULTICLASS MODEL 2 VALIDATION RESULTS (20% OF THE DISEASED SAMPLE)

In order to verify the algorithm performance we predicted using the validation sample, the results obtained are the following:

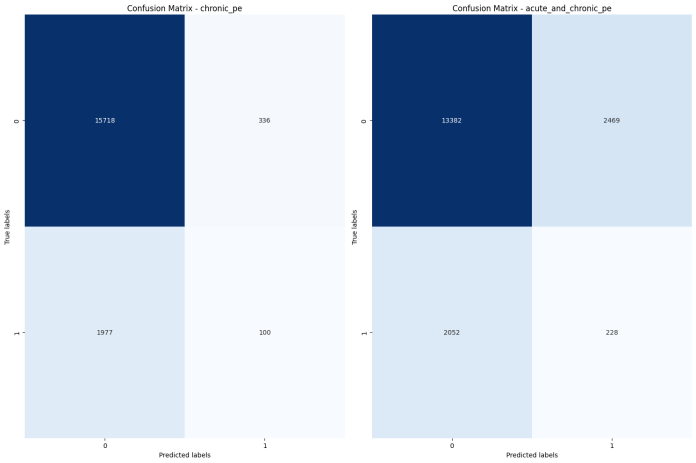


Figure 69: Confusion Matrix Results for second model of the multiclassificator model for validation sample.

	precision	recall	f1-score	support
chronic_pe	0.23	0.05	0.08	2077
acute_and_chronic_pe	0.08	0.10	0.09	2280
micro avg	0.10	0.08	0.09	4357
macro avg	0.16	0.07	0.09	4357
weighted avg	0.15	0.08	0.09	4357
samples avg	0.02	0.02	0.02	4357

Figure 70: Classification Report Results for second model of the multiclassificator model for validation sample.

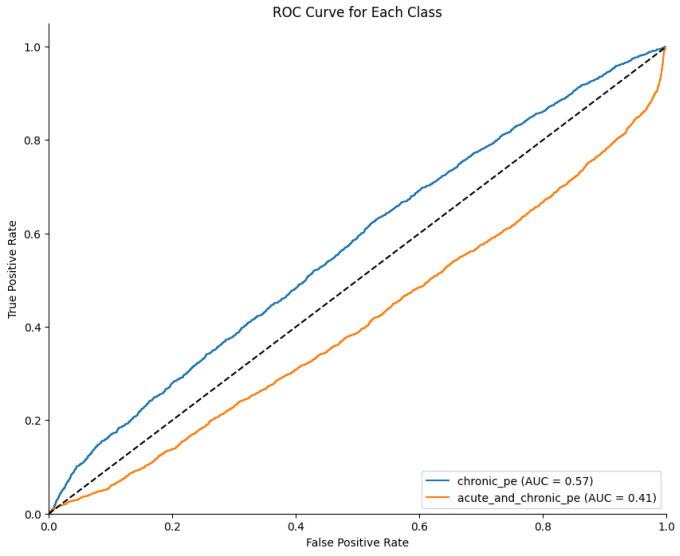


Figure 71: AUC ROC Curve Results for second model of the multiclassificator model for validation sample.

5.1.2.3. MULTICLASS MODEL 3 - PREDICTING HEART RATIO CONDITION

Multiclass Model 3, which was developed to predict heart ratio conditions, demonstrates moderate performance based on the testing results. This model classifies pulmonary emboli into two categories: 'rv_lv_ratio_gte_1' and 'rv_lv_ratio_lt_1'. Accurate predictions of conditions such as elevated heart ratios help in identifying severe cases of PE, enabling timely and appropriate interventions.

MULTICLASS MODEL 3 TESTING RESULTS (10% OF THE DISEASED SAMPLE)

In the following figures I present the testing results for the third model about heart ratio conditions:

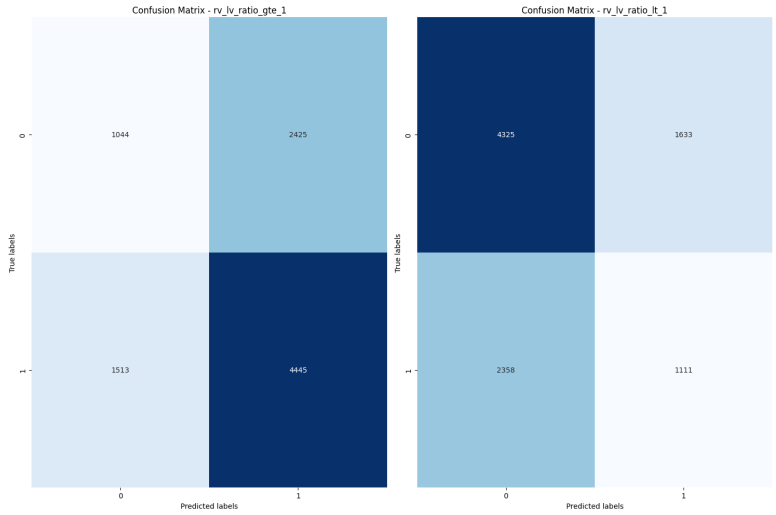


Figure 72: Confusion Matrix Results for third model of the multiclassificator model for testing sample.

	precision	recall	f1-score	support
rv_lv_ratio_gte_1	0.65	0.75	0.69	5958
rv_lv_ratio_lt_1	0.40	0.32	0.36	3469
micro avg	0.58	0.59	0.58	9427
macro avg	0.53	0.53	0.53	9427
weighted avg	0.56	0.59	0.57	9427
samples avg	0.58	0.59	0.58	9427

Figure 73: Classification Report Results for third model of the multiclassificator model for testing sample.

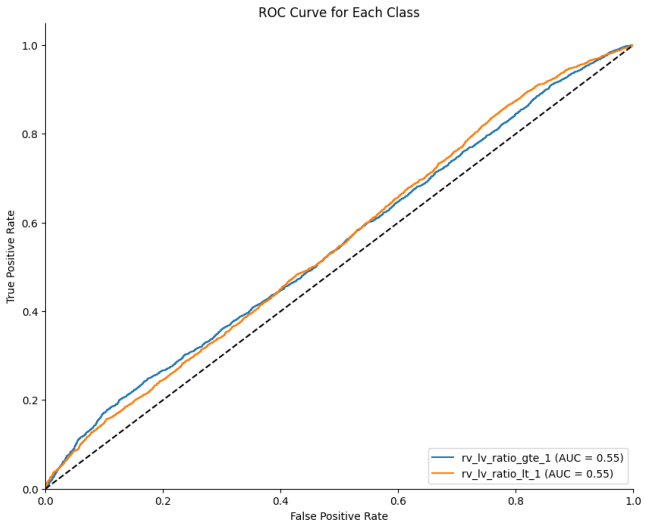


Figure 74: AUC ROC Curve Results for third model of the multiclassificator model for testing sample.

MULTICLASS MODEL 3 VALIDATION RESULTS (20% OF THE DISEASED SAMPLE)

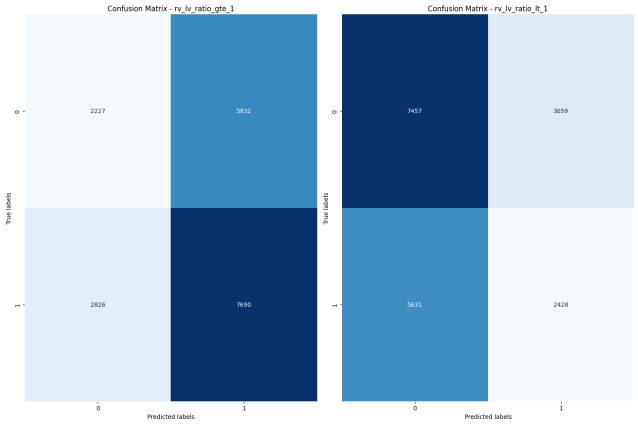


Figure 75: Confusion Matrix Results for third model of the multiclassificator model for validation sample.

	precision	recall	f1-score	support
rv_lv_ratio_gte_1	0.57	0.73	0.64	10516
rv_lv_ratio_lt_1	0.44	0.30	0.36	8059
micro avg	0.53	0.54	0.54	18575
macro avg	0.51	0.52	0.50	18575
weighted avg	0.51	0.54	0.52	18575
samples avg	0.53	0.54	0.54	18575

Figure 76: Classification Report Results for third model of the multiclassificator model for validation sample.

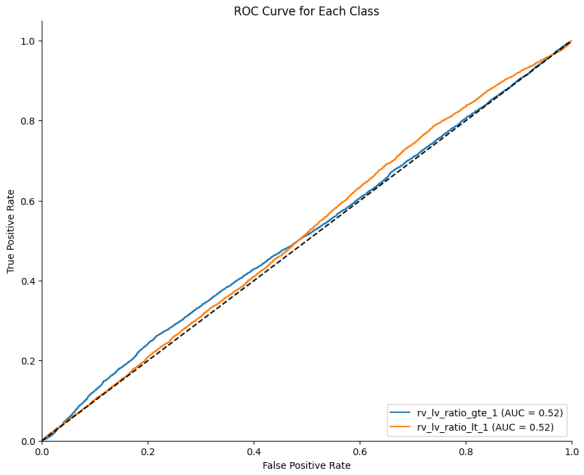


Figure 77: AUC ROC Curve Results for third model of the multiclassificator model for testing sample.

5.2. ANALYSIS OF RESULTS

5.2.1. ANALYSIS FOR THE BINARY MODEL

The binary classification model, designed to distinguish between the presence and absence of PE, demonstrates strong overall performance with accuracy rates of 71% on the test set and 74% on the validation set, and AUC-ROC values of 0.80 and 0.82 respectively. These metrics suggest that the model is generally reliable for identifying PE, though its recall for PE cases is lower in the test set, indicating that it may miss some instances of PE. While the model shows promise as a standardized tool that can complement traditional diagnostic methods, it is not yet capable of replacing the nuanced judgment of experienced clinicians.

Its role would be most beneficial in rapidly flagging potential PE cases and providing a consistent second opinion, particularly in high-pressure situations. Future enhancements should focus on improving recall for PE cases to reduce the risk of missed diagnoses.

5.2.2. ANALYSIS FOR THE MULTICLASS MODEL

The analysis of the multiclass models provides crucial insights into the strengths and limitations of our approach to classifying different categories of pulmonary embolism (PE). These models were designed to assist in the diagnosis and localization of PE, a condition that can manifest in various forms with significant implications for patient treatment. By evaluating the performance metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC, we can better understand how well the models perform across different PE classes, including left-sided, right-sided, and central locations, as well as chronic and acute conditions. This analysis not only highlights areas where the models excel but also reveals significant challenges, particularly in dealing with class imbalances that affect the models' ability to accurately detect less frequent PE categories.

5.2.2.1. MODEL 1 PREDICTING THE LOCATION OF PE (PULMONARY EMBOLISM)

The model demonstrated acceptable performance in identifying left-sided and right-sided PE, achieving accuracy rates of 86% and 94% in the test sample, and 83% and 92% in the validation sample, respectively. These results indicate that the model is effective at distinguishing between these two types of PE. However, when it comes to detecting central PE, the model struggles significantly. In the test sample, the accuracy for central PE was only 48%, with a notably low F1-score of 0.05 and a recall of just 0.02, indicating frequent failure to correctly identify true positive cases. Similarly, in the validation sample, precision and recall for central PE remained extremely low at 0.20 and 0.01, respectively.

The AUC-ROC curves further underscore these challenges. For central PE, the AUC-ROC value was around 0.52 in both the test and validation phases, revealing the model's limited ability to differentiate central PE from other categories. While the AUC-ROC values for left-sided and right-sided PE were somewhat higher, they still indicate a need for improvement, with validation AUC-ROC values of 0.59 for left-sided PE and 0.46 for right-sided PE.

The primary issue identified in this model is the class imbalance, which skews the model's predictions towards the majority classes. This results in high sensitivity (recall) and a reasonably good F1-score for the majority classes, but comes at the expense of specificity, particularly for central PE. If specificity were calculated for central PE, it would likely be near zero, highlighting a significant limitation of the model.

To address these deficiencies, particularly in detecting central PE, future iterations of the model should focus on improving feature extraction, incorporating additional data, and exploring more sophisticated model architectures. Addressing the class imbalance through techniques such as

oversampling, undersampling, or synthetic data generation could also be crucial in enhancing the model's performance across all PE categories.

5.2.2.2. MODEL 2 PREDICTING THE DISEASE CONDITION OF PE (CHRONIC PE / ACUTE AND CHRONIC PE)

Multiclass Model 2 was designed to predict the disease condition of pulmonary embolism (PE), specifically distinguishing between chronic PE and a combination of acute and chronic PE.

The model's performance was notably poor across both the test and validation phases. In the test sample, the model achieved very low precision, recall, and F1-scores for both categories. For chronic PE, the values were 0.06, 0.01, and 0.02, respectively, while for acute and chronic PE, they were slightly higher at 0.04, 0.09, and 0.06. These metrics suggest that the model struggles significantly to differentiate between these conditions. The validation phase echoed these results, with F1-scores of 0.08 for chronic PE and 0.09 for acute and chronic PE, further confirming the model's inadequate predictive capability.

The AUC-ROC values reinforce this conclusion. In the test sample, the AUC-ROC for chronic PE was 0.56 and for acute and chronic PE was 0.41, indicating minimal discriminative power. The validation results showed similar AUC-ROC values, with 0.57 for chronic PE and 0.41 for acute and chronic PE, demonstrating that the model has only a marginal ability to distinguish between these disease conditions.

The consistently low performance across both testing and validation phases suggests that Multiclass Model 2 is not effectively capturing the differences between chronic and acute PE conditions. The poor precision, recall, and F1-scores, along with the low AUC-ROC values, indicate that the model is inadequate in its current form. This underperformance may be attributed to inadequate feature extraction, insufficient data diversity, or suboptimal model architecture.

Significant improvements are necessary to make this model viable. Future efforts should focus on enhancing feature extraction techniques to better capture the nuances of chronic and acute PE conditions. Additionally, incorporating more diverse and relevant data could help the model learn more distinct patterns, while exploring more advanced modeling approaches might improve its predictive capabilities. Addressing these issues is crucial for developing a reliable model that can accurately differentiate between chronic PE and the combination of acute and chronic PE.

5.2.2.3. MODEL 3 PREDICTING HEART RATIO CONDITION (RV/LV RATIO)

Multiclass Model 3 was designed to predict the ratio between the right and left ventricles (RV/LV), which is crucial for assessing the severity of PE. The model showed moderate performance for predicting an RV/LV ratio greater than or equal to 1, achieving a precision of 0.65 and a recall of 0.75 during testing. For an RV/LV ratio less than 1, however, the model's performance was significantly lower, with a precision of 0.40 and a recall of 0.32. Validation results reflected a similar pattern, with precision of 0.57 and recall of 0.73 for RV/LV ≥ 1, and lower values for RV/LV<1.

AUC-ROC values in both testing and validation phases were low, indicating limited discriminatory power and suggesting substantial room for improvement in the model's sensitivity and specificity. To enhance the model's effectiveness, particularly in identifying cases with RV/LV<1, improvements in feature engineering and adjustments to the model architecture are recommended.

The consistently low performance of Multiclass Model 3 across both testing and validation phases highlights its challenges in accurately predicting the RV/LV ratio. The model's moderate performance for predicting an RV/LV ratio ≥ 1 is overshadowed by its significantly poorer performance for RV/LV < 1, as evidenced by the lower precision and recall values. The low AUC-ROC scores in both testing and validation phases further indicate that the model struggles to effectively discriminate between the RV/LV ratio categories. These issues may stem from inadequate feature engineering, a lack of sufficient data diversity, or limitations in the current model architecture.

To enhance the model effectiveness, it is crucial to focus on several key areas. Improving feature extraction techniques could help capture more relevant information related to the RV/LV ratio. Additionally, increasing the diversity and volume of the training data might enable the model to better differentiate between the RV/LV ratio categories. Exploring and implementing advanced modelling approaches and architectures could also improve the model's predictive performance. Addressing these areas will be essential to developing a more accurate and reliable model for assessing RV/LV ratios in the context of pulmonary embolism.

5.2.2.4. CLASS IMBALANCE IN MULTICLASS MODEL AND FUTURE STEPS TO AVOID IMBALANCING

Class imbalance is a critical factor affecting the performance of the models presented. When a dataset is imbalanced, with one class significantly outnumbering the others, the model tends to learn patterns that favor the majority class. This happens because most machine learning algorithms optimize for overall accuracy, leading the model to "take the easy route" by predicting the more frequent class to minimize its apparent error rate. As a result, while the model might show good precision and F1-scores for the majority classes, its ability to correctly identify the minority classes is severely compromised.

In the case of Multiclass Model 1 for example, we observe that the model performs adequately in predicting left-sided and right-sided PE, which are the more represented classes. However, its ability to predict central PE is significantly limited, with extremely low recall and F1-scores. This is a clear indication that the model is biased towards the majority classes and is not sufficiently learning about the less frequent class.

We understand that class imbalance has negatively impacted the model's performance, particularly in detecting central PE. However, due to time and resource constraints, it was not feasible to fully address this issue during the current phase of the project.

We acknowledge the importance of balancing the classes, whether through techniques such as oversampling the minority class, undersampling the majority class, or generating synthetic data for underrepresented classes. Adjustments to the model architecture or implementing penalty techniques that account for class imbalance during training could also be considered.

As part of future steps, we plan to address this challenge. Our focus will be on developing and implementing strategies to mitigate the effects of class imbalance, thereby improving the model's ability to accurately identify all PE categories. This will involve not only enhancements in data collection and preparation but also exploring more sophisticated models that can more effectively handle imbalanced datasets. Correcting this issue is critical to ensuring that our models can provide accurate and reliable diagnostics in clinical settings, where correctly detecting all classes is essential for patient health.

6. DISCUSSION

6.1. IMPLICATIONS OF FINDINGS

In clinical practice, the research demonstrates that the application of AI in the detection of pulmonary embolisms is not only feasible but also promising. It opens a significant opportunity for enhancing prediction capabilities in scenarios where investments can be made in researching and improving detection methods. This involves employing more powerful and dedicated servers, rather than relying on a limited academic sandbox environment like Kaggle.

The findings indicate that AI can substantially enhance medical detection capabilities for pulmonary embolisms. However, it is crucial to note that these AI-driven methods are not a substitute for healthcare professionals. Instead, they serve as a complementary tool that provides a secondary opinion, thereby assisting in reducing the number of false negatives. This has the potential to prevent mortality and improve patient outcomes by ensuring timely and accurate diagnoses.

The implementation of AI-based detection systems for pulmonary embolisms holds significant implications for clinical practice. One major benefit is the potential for increased diagnostic accuracy. AI algorithms are capable of identifying subtle patterns in CT scans that might be overlooked by the human eye, thereby enhancing the overall precision of diagnoses. This increased accuracy helps ensure that patients receive the correct diagnosis, which is crucial for appropriate treatment and management.

Early detection is another critical advantage of AI-driven prediction models. By facilitating the earlier identification of pulmonary embolisms, these models enable timely intervention, which is essential for effective treatment and improved patient outcomes. Early detection can significantly reduce the risk of severe complications and mortality associated with pulmonary embolisms.

AI systems also contribute to resource optimization within medical settings. By prioritizing cases that require immediate attention, AI helps streamline clinical workflows and reduce the workload on healthcare professionals. This optimization not only improves the efficiency of the diagnostic process but also alleviates the pressure on medical staff, allowing them to focus on the most urgent cases.

The integration of AI in clinical environments fosters continuous improvement through ongoing learning and refinement of algorithms. As these systems are exposed to new data and evolving clinical scenarios, they can adapt and enhance their predictive capabilities over time. This continuous evolution ensures that the AI models remain effective and up-to-date with current medical knowledge and practices.

Furthermore, AI-driven diagnostic support can lead to better patient outcomes by reducing the incidence of missed diagnoses. Enhanced detection capabilities ensure that conditions like pulmonary embolisms are identified more reliably, leading to timely and appropriate treatment. This improved accuracy directly translates into better health outcomes and increased patient safety.

Lastly, AI solutions offer scalability and accessibility benefits. They have the potential to extend high-quality diagnostic support to under-resourced or remote areas where specialized medical expertise may be limited. By making advanced diagnostic tools more widely available, AI can bridge gaps in healthcare access and ensure that more patients benefit from accurate and timely diagnoses.

In conclusion, while AI methods are not intended to replace healthcare professionals, they offer substantial supplementary value. By enhancing diagnostic processes, improving patient safety, and supporting medical staff, AI-driven systems hold the promise of significantly advancing the field of medical diagnostics and treatment.

The integration of this algorithm into clinical workflows can potentially reduce diagnostic errors by assisting radiologists in interpreting CT scans more accurately and efficiently. The algorithm not only automates the diagnostic process but also demonstrates better consistency compared to traditional methods, which are prone to human errors due to variability and heavy workloads. By reducing these errors and enhancing efficiency, the proposed model could significantly improve patient outcomes while alleviating the burden on healthcare systems.

In summary, while AI methods are not meant to replace healthcare professionals, they offer a valuable supplementary tool that can enhance diagnostic processes, improve patient safety, and potentially save lives by reducing the incidence of false negatives.

6.2. LIMITATIONS

The study presents several constraints and limitations that impact the overall findings and their applicability.

Firstly, the use of DICOM files and the segmentation of lung regions could be enhanced to improve the algorithm's focus on critical areas within the images, specifically regions where pulmonary embolisms are likely to be located. Better segmentation would provide more precise input data, potentially improving the algorithm's ability to detect and predict pulmonary embolisms. Since egmentation has been done, the limitation of computational power and time per week limitation has compromised this approach.

The computational resources available through Kaggle presented notable constraints, including service interruptions and limitations on computational usage due the weight of the original DICOM images sizing 950GB. Transitioning to more robust and dedicated servers could alleviate these issues, allowing for more iterations and optimizations within the model. This shift could substantially improve result quality and speed up the research process.

Another limitation is related to the version of EfficientNet employed. The study used EfficientNet-B0 due to its stability and compatibility with the image resolution of 256 pixels. However, more recent versions of EfficientNet could offer significant improvements in prediction accuracy. Future work could benefit from exploring these advanced versions to leverage their enhanced efficiency and accuracy.

Also Grad-CAM analysis was performed to Model Prediction for Binary Prediction without results due the Efficientnet architecture obtaining results like the following Figure. The results obviusly failed and left the correction for future steps.

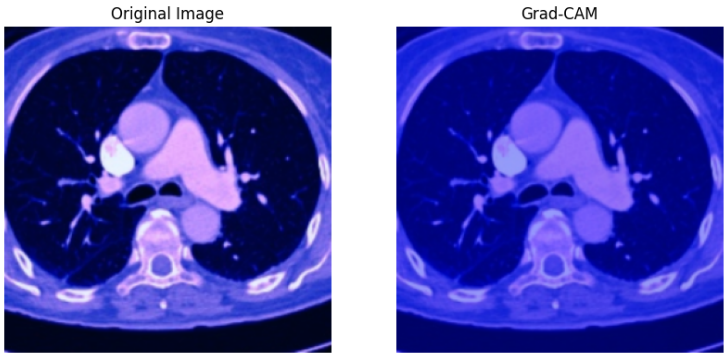


Figure 78: Grad-CAM Analysis for a PE Diseased Image.

Additionally, incorporating more comprehensive patient data through Electronic Health Records (EHR) could refine the predictive models. Access to detailed patient histories would enable the development of personalized prediction algorithms, enhancing the precision of diagnoses tailored to individual patient profiles.

Furthermore, we could have obtained better results by trying different architecture approaches as envisaged in the state of the art, such as the MONAI 3D or U-Net architectures, but again, limited computational resources and time compromised this approach.

Finally, the work carried out has been limited by the lack of in-depth clinical knowledge in the field, where I personally believe that this type of research should be accompanied by a specialised healthcare professional, making technical and clinical integration possible and opening up better prospects for development and assistance in current research or other clinical topics.

6.3. ETHICAL-SOCIAL IMPACT, SUSTAINABILITY, AND DIVERSITY ASPECTS

6.3.1. ETHICAL-SOCIAL IMPACT

The use of AI in medical diagnostics raises ethical considerations regarding patient data privacy, algorithmic transparency, and the potential for bias. Ensuring robust data protection measures, transparent AI processes, and equitable access to technology are essential for addressing these concerns. Additionally, AI tools must be validated rigorously to avoid perpetuating existing healthcare disparities.

6.3.2. SUSTAINABILITY

The environmental impact of computational resources and data storage should be considered. Adopting energy-efficient computing practices and sustainable data management strategies can mitigate the ecological footprint of AI research. Furthermore, ensuring that AI tools are developed and deployed with long-term maintenance in mind can enhance their sustainability.

6.3.3. DIVERSITY

Incorporating diverse datasets in AI training models is crucial to avoid biased outcomes and ensure that the tools perform well across various populations. Research should prioritize inclusive data collection practices and account for demographic variations to promote fairness and accuracy in diagnostic predictions.

6.4. FUTURE WORK

Several avenues for further research could significantly advance the field of AI-driven prediction for pulmonary embolisms.

First thing we can do for improving the model is to assess better the results thorugh Grad-CAM and even if possible adding more reliable data. Also exploring another approaches of model architectures such ensembles would contribute to improve model accuracy.

Another more promising direction is the use of 3D CT DICOM files for volumetric prediction. This approach, while computationally intensive, could offer a more detailed analysis of lung structures and potential embolism locations. Implementing 3D image (such the Figure showed below) into models would require substantial computational resources, which were not feasible within the scope of this study conducted on Kaggle. However, future research could explore this method using more powerful computational infrastructure.

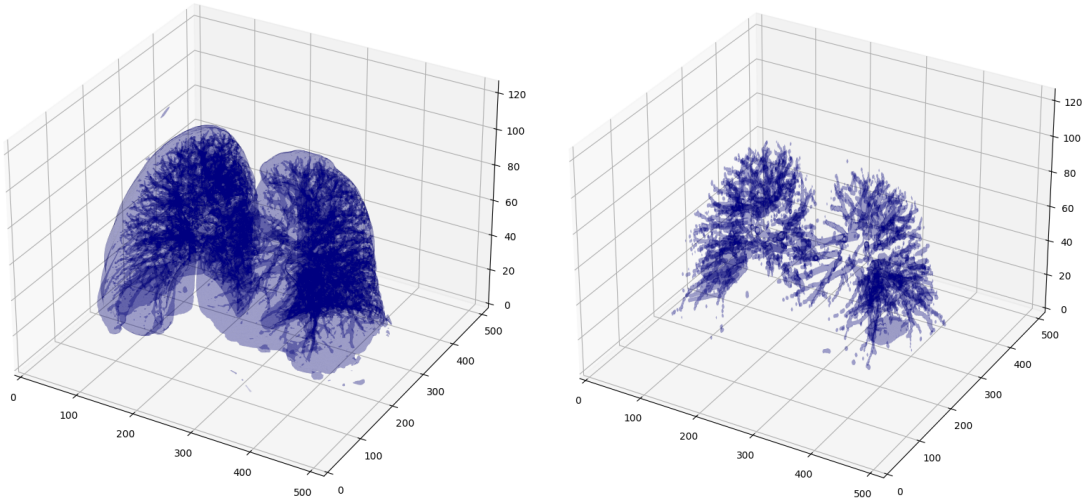


Figure 79: 3D Image Preprocessed using DICOM format.

Additionally, integrating external patient data from EHRs could enhance the algorithms by providing richer and more varied datasets. This integration would allow for the development of models that consider individual patient characteristics, leading to more personalized and accurate predictions.

Exploring these areas will not only address the current limitations but also contribute to the development of more robust and effective AI tools for the early detection of pulmonary embolisms. Enhanced computational resources and comprehensive data integration are key to advancing the capabilities and accuracy of predictive algorithms in clinical settings.

7. CONCLUSIONS

The work conducted has successfully demonstrated the significant potential of AI-driven algorithms in the detection of pulmonary embolisms using CT scans. The research highlights how AI can enhance diagnostic accuracy, enable earlier detection, and ultimately improve patient outcomes. Implementing these technologies in clinical settings represents a major advancement in medical diagnostics, providing powerful tools for radiologists and healthcare professionals.

7.1. OBJECTIVES MEET

The objectives of this study have been met and fulfilled satisfactorily:

i) A set of pulmonary embolism images was obtained, which allowed for effective data exploration, understanding, and preprocessing of the dataset, leading to satisfactory results.

ii) A tailored deep learning architecture and model for detecting pulmonary embolisms was designed and implemented, as demonstrated throughout the project. Several approaches were explored, and models, preprocessing techniques, and the architecture were carefully chosen to address the significant limitations of this study, particularly the constraints of working within the Kaggle environment and managing the large dataset.

iii) The evaluation of the results yielded satisfactory outcomes in detecting the presence of pulmonary embolisms in images. However, the results for the multiclass detection model, while acceptable, leave room for improvement.

iv) In the discussion section, the study addressed the implications and great potential that AI models hold for the future. Given the promising results of this academic study, it is evident that, with further development and professional implementation in clinical environments such as hospitals, these models could achieve highly accurate results, significantly aiding in the diagnosis and prevention of various diseases.

7.2. STREANGTHS AND WEAKNESSES OF THE WORK

One of the key strengths of this study is its focus on integrating AI with medical imaging to address a critical diagnostic challenge. The use of EfficientNet for image classification and the analysis of CT scans exemplifies how advanced algorithms can be applied to improve detection capabilities. Additionally, the research highlights the potential for AI to complement the expertise of healthcare professionals, providing a valuable secondary opinion and reducing diagnostic errors. The study effectively demonstrates the feasibility of AI-driven tools in enhancing diagnostic accuracy and supporting clinical decision-making.

However, the study also faces several weaknesses. The use of a single version of EfficientNet may limit the potential accuracy improvements that could be achieved with newer or more advanced versions. The computational constraints encountered on Kaggle restricted the extent of the analysis, impacting the ability to perform extensive model iterations and optimizations. Furthermore, the study's reliance on a limited dataset may not fully capture the variability and complexity of real-world patient data. This limitation could affect the generalizability of the results and the model's performance in diverse clinical settings.

Additionally, the study's scope was constrained by the Kaggle environment, which may not have been ideal for conducting comprehensive volumetric analysis. Future work could benefit from transitioning out of Kaggle to utilize more powerful computational resources and advanced architectures like MONAI, which are specifically designed for medical imaging tasks. This shift could facilitate more detailed and accurate analysis of 3D CT DICOM files, leading to improved predictive capabilities.

Addressing these weaknesses in future research will be crucial for advancing the field and developing more robust, reliable, and clinically applicable AI tools for the early detection and management of pulmonary embolisms.

7.3. FUTURE STEPS

Future research should focus on several key areas to build upon this study's findings. Exploring more advanced versions of EfficientNet or other cutting-edge models could significantly improve prediction accuracy. Additionally, incorporating 3D CT DICOM files for volumetric analysis presents a promising direction, offering more detailed and precise predictions despite the higher computational demands. Transitioning out of the Kaggle environment and utilizing advanced architectures like MONAI for volumetric prediction would be a particularly promising step. MONAI's specialized capabilities for medical imaging could enhance the model's performance in analyzing 3D image data, leading to more accurate assessments of lung structures and embolism locations.

Furthermore, integrating comprehensive patient data from Electronic Health Records (EHR) could enhance the personalization of AI models, enabling predictions tailored to individual patient profiles. This integration would provide a richer and more varied dataset, improving the models' ability to account for diverse clinical conditions and patient-specific factors. Enhanced models incorporating such data could offer more precise and relevant predictions, potentially leading to better patient outcomes.

In addition, exploring the application of ensemble methods and other novel model architectures could further refine prediction accuracy. Combining multiple models or incorporating advanced techniques like transfer learning could help address current limitations and improve overall performance.

To achieve these advancements, future research should also focus on acquiring more powerful computational resources and optimizing algorithms to handle the increased complexity of 3D image analysis and extensive EHR data integration. Collaborations with medical institutions and technology providers could facilitate access to the necessary infrastructure and data.

7.4. FINAL THOUGHTS

In summary, the work presented underscores the transformative potential of AI in advancing medical diagnostics. The integration of AI technologies into healthcare systems represents a significant technical leap forward, offering enhanced diagnostic capabilities and improving preventative measures across various medical domains.

AI is poised to play a fundamental role in reducing human error, enhancing life expectancy, and addressing mortality rates, not only in the context of pulmonary embolisms but across a range of medical conditions.

It is crucial to continue research in this field and leverage open medical data, while always respecting ethical and privacy considerations. Open data can drive innovation in medical research, biotechnology, and other health-related areas, ultimately improving the quality of life for individuals.

As a fnial though, I am incredibly proud of the project I developed, the challenges I faced, and the results I achieved. Throughout the development process, I came to appreciate the significant impact that AI can have in the medical field, particularly in the prevention and diagnosis of diseases like pulmonary embolism. I take pride in having successfully trained an algorithm that has the potential to reduce fatalities, even though it was developed and tested in a sandbox environment, and I acknowledge that more powerful approaches would yield even greater precision and accuracy.

Reflecting on this work, I truly believe that if such technology had been available years ago, it might have prevented the loss of my grandfather. This personal connection to the project deepens my sense of accomplishment and reinforces my commitment to advancing AI in healthcare. Knowing that the technology I developed could one day save lives makes me exceptionally proud of the work presented in this thesis.

8. BIBLIOGRAPHY

Soffer, S. K. (2021). Deep learning for pulmonary embolism detection on computed tomography pulmonary angiogram: a systematic review and meta-analysis. Nature Sci Rep.

Cahan, N. K. (2023). Multimodal fusion models for pulmonary embolism mortality prediction. . Nature Sci Rep.

Huang, S. K. (2020). PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. Nature - Digit. Med.

Huang, S. P. (2020). Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Nature - Sci Rep.

Ma, X. F. (2022). A multitask deep learning approach for pulmonary embolism detection and identification. Nature - Sci Rep.

Penn Medicine. (2024). Penn Medicine . From Pulmonary Embolism (Pulmonary Embolus): https://www.pennmedicine.org/for-patients-and-visitors/patient-information/conditions-treated-a-to-z/pulmonary-embolus

Mayo Clinic. (2022, December 01). Mayo Clinic . From Pulmonary embolism: https://www.mayoclinic.org/diseases-conditions/pulmonary-embolism/symptoms-causes/syc-20354647

Conrad Wittram, M. K.-A. (206). Acute and Chronic Pulmonary Emboli: Angiography–CT Correlation. American Journal of Roentgenology (AJR), S421 - S429.

Anouk Stein, M. C. (2020). Kaggle. From RSNA STR Pulmonary Embolism Detection.: https://kaggle.com/competitions/rsna-str-pulmonary-embolism-detection

Radiological Society of North America . (2024). Radiological Society of North America . From Main website page: https://www.rsna.org

Society of Thoracic Radiology. (n.d.). Society of Thoracic Radiology. From Main Web site page: https://thoracicrad.org

9. APPENDIX

ANNEX 0 – PROJECT SUPPLEMENTARY MATERIAL – ALL PROJECT UPLOADED IN ZENODO PLATFORM

Joan Pau Gutiérrez Pascual. (n.d.). AI-Driven Prediction for Detection of Pulmonary Embolism in CT Scans - Master Thesis Project - Additional Resources. Zenodo. https://doi.org/10.5281/zenodo.13685334

ANNEX 1 – EDA NOTEBOOK

https://github.com/j0anpau/AI-Pulmonary_Embolism_Algorithm_DL_Kaggle_Sandbox-Project/blob/487166005d832f973f734c09d51fb0c74dae3c85/TFM-Exploratory_Data_Analysis.ipynb

ANNEX II – PREPROCESSING NOTEBOOK

https://github.com/j0anpau/AI-Pulmonary_Embolism_Algorithm_DL_Kaggle_Sandbox-Project/blob/487166005d832f973f734c09d51fb0c74dae3c85/rsna-preprocessing-block.ipynb

ANNEX III – BINARY MODEL NOTEBOOK

https://github.com/j0anpau/AI-Pulmonary_Embolism_Algorithm_DL_Kaggle_Sandbox-Project/blob/487166005d832f973f734c09d51fb0c74dae3c85/TFM-Binary-Classificator-PE.ipynb

ANNEX IV– MULTICLASS MODEL NOTEBOOK

https://github.com/j0anpau/AI-Pulmonary_Embolism_Algorithm_DL_Kaggle_Sandbox-Project/blob/330b83f8513cbafbd020d7211c419e749234ff18/Multiclass__pulmonary_embolism_image_level.ipynb

ANNEX V – MULTICLASS OTHER RESULTS WITH RESNET50 APPROACH

Result before hyperparameter tuning and after hyperparameter tuning: