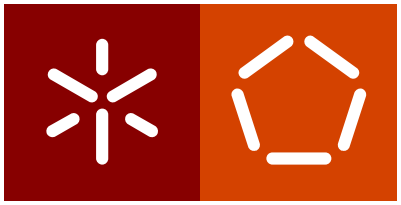
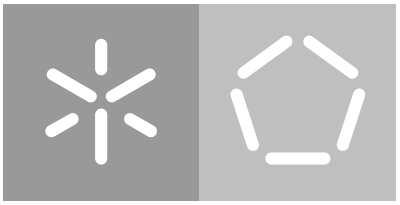


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Nuno Rafael Boto Carlos

Development of a deep learning-based algorithm to predict pneumonia cases from chest X-ray images

Janeiro de 2020



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Nuno Rafael Boto Carlos

Development of a deep learning-based algorithm to predict pneumonia cases from chest X-ray images

Master dissertation
Master Degree in Bioinformática

Dissertation supervised by
Nuno Miguel Sampaio Osório

Janeiro de 2020

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

ACKNOWLEDGEMENTS

Ao longo desta caminhada encontrei apoio e ajuda em várias pessoas, sem as quais este trabalho não teria sido concluído. Dedico assim este espaço para agradecer devidamente a todos aqueles que de uma forma ou de outra me apoiaram e estiveram do meu lado. Sem a vossa colaboração e ajuda a realização deste trabalho não teria sido possível.

Começo por agradecer ao meu orientador Nuno Osório por todo o apoio, paciência e orientação ao longo deste trabalho difícil. Agradeço também a oportunidade concedida de abraçar este desafio e a sua enorme vontade e disponibilidade em me ajudar. Devo a si acima de tudo a conclusão com sucesso deste trabalho.

Aproveito também para agradecer aos companheiros de gabinete e da equipa EvoBiomed, Pedro Araújo, Bernardino Souto, Joana Martins, Carlos Magalhães e Ana Pereira. Obrigado pela receção, ambientação e por toda a ajuda que me prestaram e dúvidas que me esclareceram.

Aos meus dois parceiros do ICVS, Ana e Carlos um grande, mas grande obrigado. A ti Carlos agradeço por todas as conversas, por todo o apoio e bons momentos passados ao longo do último ano. A ti Ana agradeço sobretudo a tua amizade, a tua preocupação e a tua boa disposição que me fez encarar este trabalho de uma maneira bem mais divertida. Agradeço-te os conselhos e lições de vida. Muito obrigado aos dois, eternos companheiros!

Uma enorme gratidão para com o todo o meu grupo de amigos da Guarda e de Braga, por todos os momentos especiais e fantásticos que me proporcionaram e por estarem presentes nesta etapa da minha vida. Desde os momentos mais calmos ás noitadas, foram vocês que estiveram sempre lá para me aturar, animar e descontrair ao longo desta caminhada. Muito obrigado meus meninos, a história não acaba aqui.

A todos os membros do Mestrado, incluindo professores e colegas, obrigado por toda a transmissõ de conhecimento, por todos os momentos partilhados ao longo destes últimos 2 anos que nunca esquecerei, mas, sobretudo pelas novas experiências que me ofereceram ao longo de todo este percurso académico.

Antes de terminar só tenho de agradecer a toda a minha família, nomeadamente aos meus pais por acreditarem em mim e me conseguirem proporcionar todos os momentos até ao final desta etapa. Certamente que vou contar sempre com vocês e vocês comigo. Agradeço ao meu irmão pela partilha de conhecimento, boa disposição e experiência que me fez concretizar este trabalho. Por fim agradeço á minha avó por toda a motivação e carinho, com certeza que merece

a conclusão deste trabalho. Obrigado por me terem feito chegar tão longe e por depositarem a vossa confiança em mim. Este trabalho é especialmente dedicado a vocês.

”Anything is possible when you have the right people there to support you.”

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

Interstitial lung diseases (ILD) are defined as a set of more than 200 pulmonary disorders. Among these, the ones broadly termed as pneumonia represent a major cause of morbidity and mortality in the world. The chest radiograph (CXR) was the first x-ray based lung imaging technique to emerge and is still widely used as a diagnostic method for pneumonia and other lung diseases. However, correct interpretation of CXR requires analysis by experts and stays vulnerable to errors and observer-related variation. To counteract these problems, artificial intelligence (AI) methods have been applied for the automated analysis of CXR and other medical images. The deep learning (DL) branch of AI and in the particular the methods based on convolutional neural networks (CNN), recently obtained impressive results in these tasks.

This dissertation presents a DL approach to classify pneumonia from medical CXR image datasets. Two different models based on the development of CNN were trained from a preprocessed dataset of CXR images obtained from 8562 individuals classified as normal (n=7214) or with pneumonia (n=1348) (Dataset XP1'). Model 1 applied a normal cross entropy loss function, and model 2 an alternative loss function aiming at counteracting the unbalance in normal/pneumonia class frequency. For performance enhancing both models underwent a hyper optimization procedure. The optimized model 1 and 2 were tested on a test set from XP1'. To better understand the predictability and generalization potential we then tested both models on an unrelated test set of 624 images (Dataset XP2).

Interestingly, model 1 obtained better performance when tested on XP2 than in XP1', scoring an accuracy of 85%, recall of 93% and precision of 85% for the detection of the pneumonia class. The higher homogeneity present on dataset XP2 compared with dataset XP1' could be a plausible justification. As for model 2, it correctly predicted more pneumonia cases on test set XP1' than model 1. However, on test set XP2 the results were poor, predicting most cases as pneumonia and scoring a recall value of only 26% for the pneumonia class. Testing the DL models on unseen data is a relevant but not always performed validation. Overall, the higher accuracy, recall and precision levels of model 1 in XP2 suggests it has a higher potential to be applied for real-word application although its performance should be further improved and evaluated. This work opened promising new lines of research for the future development of a high-performance CNN-based automated method to classify CXR and assist in the diagnostic of pneumonia.

Keywords: interstitial lung diseases, pneumonia, chest radiographs, artificial intelligence, deep learning, convolutional neural networks.

RESUMO

Doenças intersticiais pulmonares são definidas como um conjunto de mais de 200 doenças pulmonares. Dentro deste grupo de doenças, as doenças denominadas como pneumonia ou pneumonite representam uma condição inflamatória que afecta o interstício pulmonar e representam uma das principais causas de morbidade e mortalidade no mundo. A radiografia torácica foi a primeira técnica de imagiologia pulmonar baseada em raios-x a surgir sendo, ainda amplamente utilizada como método de diagnóstico de pneumonia e outras doenças pulmonares. No entanto, a correcta interpretação de radiografias torácicas requer uma análise de pessoal especializado e encontra-se vulnerável a erros e variações relacionadas com o observador. De modo a contrariar estes problemas, métodos de inteligência artificial têm sido aplicados na análise automatizada de radiografias torácicas e outro tipo de imagens médicas. Métodos de ”Deep learning” e em particular, métodos baseados em redes neuronais convolucionais, obtiveram recentemente resultados impressionantes quando aplicados nesta área de estudo.

Esta dissertação apresenta uma abordagem de ”deep learning” que permite classificar imagens de pneumonia a partir de ”datasets” de radiografias torácicas. Dois modelos diferentes baseados no desenvolvimento de redes neuronais convolucionais foram treinados a partir de um ”dataset” preprocessado de radiografias torácicas obtido a partir de 8562 indivíduos classificados como normais (n=7214) ou como doentes de pneumonia (n=1348) (”Dataset” XP1’). Ao modelo 1 foi aplicada uma função ”loss” de entropia cruzada normal, e ao modelo 2 foi aplicada uma função ”loss” alternativa que visa contrariar o desbalanceamento entre casos normais e de pneumonia presente no ”dataset XP1’ ”. Para melhorar o desempenho, ambos os modelos foram submetidos a um procedimento de hiper optimização. Os modelos 1 e 2 otimizados foram de seguida testados no conjunto de teste do ”dataset XP1’. Para entender melhor a capacidade de previsão e generalização, os dois modelos foram também testados num conjunto de teste não relacionado de 624 imagens (Dataset XP2).

Curiosamente, o modelo 1 obteve melhor desempenho quando testado no XP2 do que no XP1’, obtendo uma ”accuracy” de 85%, sensibilidade de 93% e precisão de 85% durante a deteção de casos de pneumonia. A maior homogeneidade de informação presente no XP2 em comparação com o XP1’, é vista como a justificação mais plausível. Quanto ao modelo 2, ele previu correctamente mais casos de pneumonia no XP1 do que o modelo 1. No entanto, quando testado no XP2, os resultados ficaram abaixo das expectativas, prevendo a maioria dos casos como pneumonia e obtendo um valor de sensibilidade de apenas 26% para a classe de pneumo-

nia. Testar os modelos de ”deep learning” em dados não relacionados é uma técnica de validação relevante, no entanto nem sempre é realizada. Os níveis elevados de precisão, sensibilidade e ”accuracy” do modelo 1 quando aplicado no XP2 sugerem que possui um grande potencial de utilização em aplicações de carácter real, no entanto, o seu desempenho pode ainda ser melhorado. Este trabalho abriu novas e promissoras vias de pesquisa para o desenvolvimento futuro de um método automatizado baseado em CNN de alto desempenho que seja capaz de classificar radiografias torácicas e auxiliar no diagnóstico de pneumonia.

Palavras-chave: doenças intersticiais pulmonares, pneumonia, radiografias torácicas, inteligência artificial, ”deep learning”, redes neuronais convolucionais.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

2D	Two-Dimensional
3D	Three-Dimensional
AI	Artificial Intelligence
ANN	Artificial Neural Network
AP	Anteroposterior
CAD	Computer-aided Diagnosis
CNN	Convolutional Neural Network
CT	Computed Tomography
CTD	Connective tissue disease
CXR	Chest Radiograph
DICOM	Digital Imaging and Communications in Medicine
DILD	Drug-induced Interstitial Lung Disease
DL	Deep learning
DNN	Deep Neural Network
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
HP	Hypersensitivity Pneumonitis
HPS	Hermansky-Pudlak syndrome
HRCT	High-Resolution Computed Tomography

IIP	Idiopathic Interstitial Pneumonia
ILD	Interstitial Lung Disease
IPF	Idiopathic Pulmonary Fibrosis
LAM	Lymphangioleiomyomatosis
ML	Machine Learning
PA	Posteroranterior
ReLU	Rectified Linear Unit
RSNA	Radiologic Society of North America
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

1

INTRODUCTION

1.1 CONTEXT AND MOTIVATION

Interstitial lung diseases (ILD) can be defined as a set of more than 200 chronic pulmonary disorders, which can cause several degrees of pulmonary fibrosis and inflammation [1]. The ILDs can be classified into two major groups: idiopathic or unknown cause diseases, and known cause diseases [2]. Even if these diseases can be classified into different groups, they have similar clinical manifestations with each other, hindering their differential diagnosis. ILDs typically lead to several complications in the patient affecting the normal function of the respiratory system, such as reduced lung capacity and volume, progressive scarring of lung tissue and decreased amount of available oxygen in the bloodstream [3]. ILDs represents a major cause of morbidity and mortality in the world, being directly linked to respiratory failure [4].

High-resolution computed tomography (HRCT), computed tomography (CT) and chest radiograph (CXR), both based in the principle of x-ray imaging, appear as the most classical diagnostic methods used in the detection of pulmonary diseases, including ILD. The CXR was the first technique to emerge, being widely used nowadays as a first diagnostic method to evaluate the structures of the pulmonary system in most clinical settings [5]. CT and HRCT came latter and are nowadays common in-vivo radiology imaging tools used in the identification of specific types of ILDs by the visualization of abnormal imaging patterns. However, these approaches require analysis by an expert radiologist and stay vulnerable to the inter and intra-observer variation [6]. These variations can lead to incorrect image interpretation ultimately leading to a potential increase in the mortality rates due to treatable ILDs

Over the last 40 years, approaches based on the application of computer systems as complementary tools in the doctors decision-making diagnosis, have been created. Current Computer-aided diagnosis (CAD) systems focused on the detection and classification of many types of lesions obtained from imaging methods, becoming relevant in the detection of lesions that are easily missed [7, 8]. Overall, these CAD systems are gaining importance in the improvement of diagnosis accuracy. Remarkably,the advances in artificial intelligence (AI) techniques and en-

hanced computing power conjoined with a large increase in medical imaging data has opened up new opportunities for the application of CAD systems as focal tools in the modern diagnosis of ILD and other relevant health concerns. AI methods, especially deep learning (DL), as the potential substitute and largely enhance the usefulness of feature extraction and disease classification in the traditional CAD systems [8].

DL is a branch of machine learning (ML), best known by its applications in computer vision, bioinformatics, drug design, and speech recognition. Although the concept of DL has existed since the 1980s, its real impact started only in the early 2000s, being now considered as a tool able to strongly impact vast areas of the society. The most modern DL models are based on artificial neural networks (ANN), which are inspired by knowledge from neurosciences, such as the interpretation of information processing and communication patterns in the nervous system, more properly in neural networks [9]. DL models have recently obtained impressive results in a variety of computer vision problems, which leads to its new application in other areas, such as medical image analysis [3]. Regarding medical image analysis AI technology can be used to perform automatic lesion detection aiming at improved differential diagnostic [10]. Over the last years, many articles based on the application of DL in the medical field have been published [11]. According to the PubMed database, the convolutional neural networks (CNN) are starting to gain advantage over the other DL methods [11]. Some studies even demonstrated, by a comparative analysis of lung pattern classification, that CNN were more effective than other existing DL methods [3, 12].

Given the complexity involved in the medical diagnosis decision of ILDs, this work will be based on the development of a DL model capable of distinguishing CXR images from healthy individuals from the ones acquired in individuals diagnosed with a common ILD broadly known as pneumonia. This highly challenging project will allow extracting practical knowledge on how AI can be useful in the future of medical image analysis and in the diagnosis of pneumonia.

1.2 OBJECTIVES

The main goal of this thesis is the development of a DL-based algorithm for classification of pneumonia images from medical CXR image datasets. In detail, the scientific and technological objectives of this work are:

- Review the relevant literature in lung diseases, their classic and modern diagnosis and how deep learning methods can be applied in related scenarios;
- Explore relevant datasets of CXR image data commonly used in pneumonia diagnosis;
- Implement the automated analysis of chest medical images using Python language;
- Develop deep learning pipelines for pneumonia diagnosis.

1.3 PAPER ORGANIZATION

This dissertation is structured in six chapters.

The current chapter, Chapter 1, highlights the proposed objectives, an overview of the structure of the document and includes the general introduction to the main topics of interstitial lung diseases and artificial intelligence illustrating the context and problematic that motivated this dissertation.

In Chapter 2 a detailed review of the state-of-the-art relevant to this work is presented. A particular focus is given to interstitial lung diseases diagnosis, computer aided diagnosis and detection allied with artificial intelligence, artificial neural networks, deep learning methods and its applications in healthcare.

During Chapter 3 the pratical methods involving both data selection and model development are described in detail. At the same time, the preferences and choices of the work steps are carefully explained to provide a reproducible pipeline.

In Chapter 4 the computational and technical steps of image preprocessing, convolutional neural network models development, hyperparameter tuning, and convolutional neural network final applications are explained in a more detailed context.

During Chapter 5 the main results generated in this thesis are presented and discussed. It starts by the hyper optimization of the developed models and also includes the model performance evaluation.

Finally, at Chapter 6, the general conclusions from this work and topics about its future improvement are drawn.

2

STATE OF THE ART

2.1 PNEUMONIA AND OTHER INTERSTITIAL LUNG DISEASES

ILDs are defined as a set of chronic pulmonary disorders characterized by the inflammation of the lung tissue, which can lead to pulmonary fibrosis. These group of disorders can be clinically characterized by diffuse infiltrates on the CXR and histologically by changes of the gas exchanging portion of the lungs [1]. When in fibrosis, other complications can exist, such as lung stiffness, restriction of lung volumes and impaired oxygenation due to the reduced ability of the air sacs to capture and carry oxygen into the bloodstream. In extreme cases, impaired oxygenation can lead to permanent loss of the ability to breathe [3]. The most common symptoms are the cough and dyspnea, although patients may occasionally have an abnormal CXR in the absence of these symptoms [1].

The group of ILD involves more than 200 chronic lung disorders and accounts for 15% of all cases seen by pulmonologists [4]. These disorders involve the area between the alveolar epithelial and capillary basement membranes, but also frequently involve the alveolar epithelium, alveolar space, pulmonary microvasculature and less usually, the respiratory bronchioles, larger airways, and the pleura [1]. There are some main causes for ILDs, like autoimmune diseases, genetic abnormalities, infections and long-term exposures to hazardous materials. However, in other cases, the cause of ILDs remains unknown and are defined as idiopathic interstitial pneumonia [3]. It is estimated that up to 65% of all ILDs are of unknown causes and only 35% are of known etiology cause [2]. Thus, ILDs can be classified into two major groups: idiopathic diseases and known cause diseases (**Table 1**). This classification is not the ideal but the possible based on the present medical knowledge and tends to help grouping the ILDs based on clinical similarities helping in the differential diagnosis of the different ILDs. The term pneumonia or pneumonitis broadly represents any inflammatory condition involving the lungs, including the pulmonary interstitium. Thus, these terms are commonly used to refer to several ILDs of the idiopathic and non-idiopathic groups (**Table 1**).

Table 1: Major categories of interstitial lung diseases. Adapted from [2]

IDIOPATHIC ILDs (65%)	KNOWN CAUSE ILDs (35%)
Idiopathic interstitial pneumonia (IIP): • Idiopathic pulmonary fibrosis (IPF) • Non-specific interstitial pneumonia (NSIP) • Cryptogenic organizing pneumonia (COP) • Respiratory bronchiolitis interstitial lung disease (RBILD) • Desquamative interstitial pneumonia (DIP) • Acute interstitial (AIP) • Lymphoid interstitial pneumonia (LIP)	**Inherited disorders:** • Familial pulmonary fibrosis • Hermansky-Pudlak syndrome
Eosinophilic pneumonia (EP): • Chronic eosinophilic pneumonia (CEP) • Acute eosinophilic pneumonia (AEP)	**Connective tissue disease-associated interstitial lung disease (CTD-ILD)**
Sarcoidosis	**Hypersensitivity pneumonitis**
Primary disorders: • Pulmonary Langerhans cell histiocytosis (HX)	**Latrogneic pneumonitis:** • Drug-induced ILD • Radiation injury
Lymphangioleiomyomatosis (LAM)	**Occupational lung disease:** • Asthma • Bronchiolis obliterans • Pneumoconiosis

2.1.1 *Disease groups*

Idiopathic interstitial pneumonia (IIP) is one type of ILD of unknown etiology that shares clinical and radiologic features, being distinguished primarily by its histological patterns following lung biopsy. All the seven IIP subtypes are characterized by varying degrees of fibrosis and inflammation on the lungs [13]. The main physical symptom is dyspnea. Idiopathic pulmonary fibrosis (IPF) is the most important and common form of chronic ILD, most frequently occurring in older adults [14]. The IPF is associated with a radiologic pattern of usual interstitial pneumonia term, being recently placed in the IPP category [13].

The eosinophilic ILD, commonly known as eosinophilic pneumonia, is an heterogeneous group of lung diseases characterized by a set of infectious and noninfectious pulmonary conditions that induce the infiltration of an increased number of eosinophils into lung areas, such as

alveolar spaces in the interstitium [15]. These eosinophils are immune system cells involved in the response to the infection by multicellular parasites and other pathogens, and responsible to regulate allergy and asthma-associated mechanisms [16]. The most common causes of this disease are parasite infections, allergic reactions, exposure to drugs and toxins [17]. Although some causes for this disease are known, many others remain unknown, like in the case of chronic eosinophilic pneumonia and acute eosinophilic pneumonia conditions, where the massive accumulation of eosinophils in the lungs cannot be clearly explained in the majority of cases [18, 19].

Other categories of unknown cause ILDs are the sarcoidosis, a multisystemic disorder that is characterized by the formation of abnormal collections of inflammatory cells, known as immune granulomas, in a variety of organs [20]. This disease can affect the eyes, liver, heart, and brain, but the lungs and the lymphatic system are the most predominantly affected [21]. It is suspected that sarcoidosis can be caused by an immune reaction trigger, such as an infection or chemicals in those who are genetically predisposed [22, 20]. The symptoms of this disease depend on the involved organ, in the case of lungs the usual symptoms are cough, chest pain and shortness of breath [23].

Langerhans cell histiocytosis, commonly referred to histiocytosis X is a rare and complex disease of unknown cause, characterized by the abnormal proliferation and infiltration of the bone marrow-derived Langerhans cells in multiple organs, being that the lung and bone are the most affected [24]. These cells are specific leukocyte dendritic cells involved in the regulation of the immune system, being capable to migrate from the skin to lymph nodes [25]. This kind of disease provokes a non-specific inflammatory response which includes fever, lethargy and weight loss [24]. In the case of lung involvement, the histiocytosis provokes a swelling of the bronchioles and blood vessels, which leads to breathing problems and increased risk of infection [25].

Lymphangioleiomyomatosis (LAM) is the last major category of idiopathic ILD. LAM is a rare multi-system and progressive disease that commonly results in cystic lung destruction. This disease results from abnormal smooth muscle-like cells (LAM cells) interstitial proliferation in the lungs, kidney and lymphatic channels [26]. That proliferation can, at the early phase, obstruct venules, lymphatics, and small airways and can, at the late phase, develop cystic spaces throughout the lung [27]. LAM disease occurs predominantly in premenopausal women, suggesting the involvement of estrogen and/or progesterone in the disease progress. However, the main cause of the LAM disease is not known for certain. [26]. LAM is considered an atypical or attenuated manifestation of tuberous sclerosis, once its manifestation is associated with people carrying a genetic disorder characterized by the growth of numerous benign tumors in many parts of the body [28].

The connective tissue diseases (CTD) can be defined as a set of disorders caused by autoimmune-mediated damages associated with producing and circulating autoantibodies that target various body organs [29, 30]. In the case of CTD-ILD, the main target are the lungs. CTD can affect some regions of the lungs like chest wall, pleura, vasculature, airways, and parenchyma [31]. In these cases, the ILD is considered a clinical manifestation of CTD, whereby the CTD-ILD category can be defined as a progressive lung parenchymal manifestation of CTD [29]. The overall incidence of CTD-associated interstitial lung disease (CTD-ILD) is 15% and the main symptoms of this disease are fatigue and cough [30].

Another ILD category which cause is known is the hypersensitivity pneumonitis (HP), also designated extrinsic allergic alveolitis. HP represents a group of pulmonary disorders mediated by chronic inflammation reactions involving the lung parenchyma, more properly the bronchi and peri-bronchi tissue. The HP is commonly provoked by the inhalation of an allergen. When the patient is in prolonged contact with a certain type of allergen to which it is sensitive and super-responsive, exaggerated immune reactions will be activated by the organism leading to alveoli damages [32]. These diseases can be caused by inhalation of multiple and varied agents, like microbes, chemicals, and animal and plant proteins [33].

Iatrogenic pneumonitis and fibrosis can be defined as a group of ILD induced by external factors. Many treatment agents have been associated with pulmonary disorders, especially interstitial inflammation and fibrosis [34]. The drug-induced interstitial lung diseases (DILD) can be caused by several medications or treatments like chemotherapeutic agents, antibiotics, antiarrhythmic drugs, and immunosuppressives. The lungs are presented as an easy target for toxic substances and can act as a metabolism site for certain drug compounds. Since some agents are capable to induce specific respiratory reactions, sometimes the toxicity present in these agents can injury the lungs, principally the lung parenchyma, pleura, and pulmonary vasculature. These injuries may result from a direct or indirect drug effect, being that the oral and parenteral administration drugs are the main causes of DILD [35]. About 380 drugs are known to cause DILD and that number tends to increase as new agents are developed [36, 35]. However, therapies using radiation and several medical devices also present various levels of toxicity to the patient lungs and are therefore also considered responsible for DILDs [34].

Occupational lung diseases are another major category of known cause ILDs. These diseases are defined as a variety of lung disorders caused by the inhalation or ingestion of dust particles [37]. These disorders are provoked by people exposed to certain substances in their workplaces. The occupational lung diseases can be caused directly or indirectly by an immunological response to a variety of specks of dust, chemicals, proteins, and even organisms [38]. There is a lot of occupational lung diseases, being asthma, bronchiolitis obliterans and ILDs, like pneumoconiosis, HP, and lung fibrosis, the most common [39].

Relating to the inherited diseases, they are caused by genetic disorders that pass through the offspring [40]. These disorders occur due to one or more abnormalities in the human genome, staying in the family tree for several generations. One of ILDs belonging to inherited diseases category is the familial pulmonary fibrosis [41]. It is characterized by the accumulation of excessive scar tissue in the lungs [3]. When the cause of pulmonary fibrosis is not known, it is classified as IPF. When this disease manifests itself in several people of the same family is designated as familial pulmonary fibrosis and can be attributed to a known genetic cause [41]. During the last years, it was revealed that mutations involving the telomerase complex, the enzyme responsible for the formation and maintenance of the chromosomal ends, are associated with pulmonary fibrosis [40]. Another inherited disease is Hermansky-Pudlak syndrome (HPS). The HPS is an extremely rare autosomal recessive disease characterized by oculocutaneous albinism, a form of decreased pigmentation, and bleeding problems [42]. HPS presents pulmonary fibrosis and immunological deficiencies as major complications [43]. At least six distinct genetic forms of HPS have been identified. Mutations in the HPS1 gene are the main cause of most cases of the HPS in the world [43]. It is also known that the prevailing of pathogenic variants of HPS1 and HPS4 on the genetic basis of the family is related to lethal pulmonary fibrosis [43].

2.2 MEDICAL IMAGING: X-RAY TECHNIQUE

The X-ray radiation was first discovered in the year 1895, by a German physicist named William Conrad Rontgën [44]. It is a form of electromagnetic radiation that can be produced by an energy source and can pass through specific materials. X-ray radiation is similar to visible light but, has more energy that allows itself to penetrate a variety of objects. The X-rays are produced upon a sudden deceleration of fast-moving electrons and at the moment they collide and interact with the target anode [45].The X-ray wavelength can range from 0.01 to 10 nanometers and X-ray energies can range from 100 electro-volts to 100 kilo electro-volts. The X-rays can be classified into soft X-rays and hard X- rays according to the wavelength. The soft X-rays have wavelengths about 10 nanometers, being in the range of the electromagnetic spectrum between ultraviolet light and gamma-rays. The hard X-rays have wavelengths about 0.1 nanometers, being in the same region as gamma-rays in the electromagnetic spectrum [46]. Due to their ability to pass through several materials, X-rays were cautiously applied for various nondestructive evaluation techniques [44].

X-ray imaging, most commonly called radiography, can generate pictures of the inside of the patient body, showing the different parts of the body in different shades of black and white. This contrast is due to the different absorption of X-radiation by the various tissues of the body [47]. For example, the calcium present in the bones absorbs a great number of X-rays, making

bones looks white [48]. The bones present a high radiographic density, opposing the passage of the X-rays through its structure. Contrarly, fatty and soft tissues, muscles and liquids have a low radiographic density, absorbing fewer X-rays. This low absorption makes these tissues look gray [49]. In the case of the lungs, they appear with the blackest color in the radiography, due to the presence of air that absorbs only a minimum amount of X-ray radiation [50]. So, to create a radiography (**Figure 1**), the patient is positioned according to the type of visualization to be obtained, so that the part of the body being imaged is located between the X-ray source and the X-ray detector. When the exam starts, X-rays travel through the body and are absorbed in different scales by different tissues. At the end of the process, the X-ray detector converts the energy transmitted by the X-rays into electronic signals which are digitized and recorded on the computer. These signals present the different white and black contrast of the tissues, being after analyzed by the specialized staff [51].

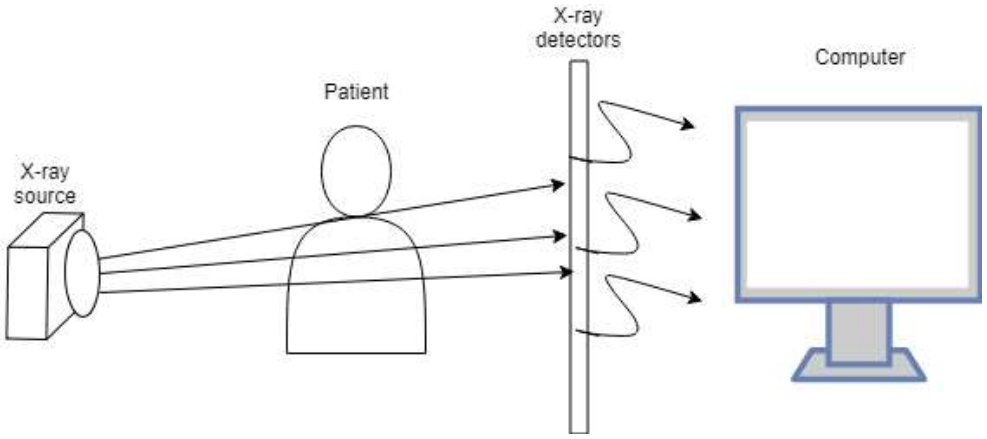


Figure 1: Schematic view of X-ray imaging

X-rays are classified as carcinogenic by the World Health Organization, being capable to cause mutations in the patients DNA and, therefore, increase the probability of developing cancer in later life [5]. The X-rays may also present some short-term effects of lower risk, such vomiting, bleeding and fainting, and at high levels of exposure some tissue damages like cataracts, skin lesions, and hair loss can occur. Nonetheless, the use of X-rays presents large benefits in medicine as a non-invasive method to monitor internal organs of the body potentiating early diagnostic of several health conditions [52]. Due to its health risks, it is not recommended to perform tests involving ionizing radiation when the desired information can be obtained by a non-ionizing method with comparable accuracy. So, X-ray imaging technology should be per- formed after careful consideration of the patients health and only in response to a medical necessity to answer a clinical question or to monitor the treatment of a disease [46].

Several other types of X-ray based medical imaging procedures relying on different technologies and techniques, such as CT, fluoroscopy, radiography and mammography (**Table 2**). All of these medical resources work in the same principle: an energy source emits X-rays that pass through a specific part of the patient's body. Within the body structure, an x-ray portion is absorbed or scattered by the internal tissues, and the other portion is transmitted to a detector, normally a film or a computer screen [51].

Table 2: Different techniques of X-ray imaging.

	Imaging technique	Diagnosis	Resolution
Chest radiograph	Radiography	Bone fractures; Tumors and abnormal masses; Some ILD; Calcifications; Foreign objects; Dental problems	Two dimensional
Mammography	Radiography	Breast cancer; Microcalcifications; Irregular and regular-shaped masses	Two dimensional
Fluoroscopy	X-rays with fluorescent screening	Movement of beating heart; blood flow.	Real-time images. Detailed movements
Computed tomography	X-ray with computer processing	Detailed information: lungs and airways; Heart and blood vessels Bones; Soft tissues; Head and brain; Large part of tumors.	Cross-sectional images combined to form a three-dimensional x-ray image.

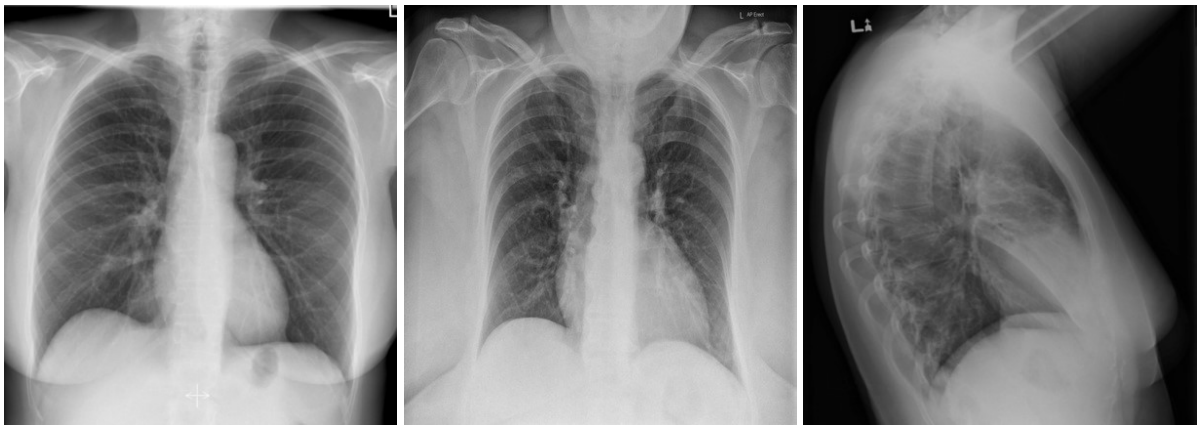
2.2.1 *Chest radiograph*

CXR is a particular form of X-ray imaging, building on the creation of a projection radiograph by using ionizing radiation to generate images of the patient chest [8]. By 1900, five years after the invention of the X-ray technology, the use of the X-ray machine was being considered as essential for medical care, in particular for the diagnosis of foreign bodies and fractures. The fact that the equipment needed to make an X-ray machine was relatively cheap and relatively simple to use has made these devices very popular through various medical and non-medical sites. The sudden ability to observe inside the human body had a very high medical and societal

impact. As X-ray technology was being widely used, new applications such as the access to visual information of the chest were emerging [44].

Nowadays, the CXR are still an important diagnostic method for evaluation of the structures of the pulmonary and cardiovascular system, such as the airways, pulmonary parenchyma and vessels, pleura, chest wall, and heart [53]. Pneumonia, pneumothorax, ILDs, heart failure, bone fracture, and hiatal hernia are some conditions commonly identified by CXR [52].

Different views of the chest can be obtained by changing the orientation of the body and the direction of the x-ray beam. Posteroanterior (PA), anteroposterior (AP) and lateral views are the most common, being that additional perspectives such as supine, lateral decubitus, expiration view, lordotic view and oblique view can also be acquired. The PA view (**Figure 2a**) is typically the preferencial, the x-ray beam enters through the posterior part of the chest and exits the body through the anterior part, where the beam is detected. Briefly, the x-ray detector is positioned behind the patient and the x-ray source is positioned toward the patient. In the case of the AP view (**Figure 2b**), the positions of the x-ray source and detector are reversed, so, the x-ray beam enters through the anterior part of the chest and exits the body through the posterior part [47]. In the lateral views (**Figure 2c**), the patient stands with both arms raised, the left side of the thorax stands adjacent to the x-ray detector and the x-ray beam enters through the right side of the thorax [54].



(a) (b) (c)

Figure 2: Examples of chest radiograph obtained by different views. (a) Posteroanterior CRX view. Case courtesy of Dr Usman Bashir, Radiopaedia.org, rID: 18394; (b) anteroposterior CRX view. Case courtesy of Dr Derek Smith, Radiopaedia.org, rID: 62094; (c) lateral CRX view. Case courtesy of Dr Garth Kruger, Radiopaedia.org, rID: 21938. CXR, chest radiograph.

2.2.2 *Computed tomography*

The CT method is defined as a computerized imaging procedure using the ionizing radiation in the form of X-rays [55]. This type of X-ray imaging has some advantages over the other X-ray procedures. The CT scan can be performed in minutes and allows more assertive confirmation or exclusion of a certain diagnosis [56]. The CT method has grown to a significant level in the last years as a result of technological advances and new clinical applications [57].

The first application of CT dates between 1957 and 1963, when [58] applied this technology to improve radiotherapy planning. A few years later, the first successful implementation of CT was performed by [59], surprising the entire medical society. In the earlies of 1970, the official introduction of the CT scan in the medical world was made. From that moment, the number of CT examinations began to extremely increase and, nowadays, is widely used for both diagnostic and therapeutic procedures [57]. CT has revolutionized the diagnostic decision, leading to better surgery, better diagnosis and treatment of cancer, and better treatment of injuries, stroke and cardiac conditions [60, 56].

The CT scanner uses a motorized X-ray source that rotates around a circular opening called gantry. During the CT scan, the patient stays on a bed that slowly moves through the gantry while the X-ray tube rotates around the patient, releasing narrow beams of X-rays through the patient's body [59]. The digital X-ray detectors are located directly opposite the X-ray source and when the X-rays leave the patient's body, electronic signals are produced. These electronic signals are collected by the detectors and transmitted to a computer. In the computer, the electronic signals are processed to create cross-sectional images, commonly called slices, of the body. These slices are called tomography images and contain more detailed information than classic X-rays images. The collection of successive slices by the computer can be digitally together to form a three-dimensional image of the patient's body, allowing the easier identification and location of basic structures as well as possible tumors or abnormalities. However, the image slices can also be displayed individually but, unlike the three-dimensional (3D) imaging, cannot rotate the image in space or view slices in succession, being more difficult to find the exact place where the abnormalities may be located [61].

Being an imaging procedure that uses ionizing x-radiation, a CT scan has the potential to cause biological effects in living tissues, such as the development of cancer. This risk increases with the number of exposures added up over the life of a person [62]. Sometimes intravenous contrast agents are applied to better visualization and contrasting of the CT images. These agents can lead to some allergic reactions in the patients or, in rare cases, temporary kidney failure [63]. In the cases of pregnant women, if the CT scan is applied in the abdomen and pelvis regions and if the exposure is delayed or accumulated, injuries to the fetus may occur [56].

CT scans can be used for the identification of diseases and injuries within various regions of the human body. In that way, CT scans can give detailed information about several organs. In the case of the lungs, CT images can reveal the presence of tumors, excess fluid, ILDs, interstitial lung abnormalities, and pulmonary embolisms. CT images of the heart give information about abnormalities and diseases, detailing the structure of the heart and blood vessels. CT can also be used to image the head in order to detect injuries, hemorrhages, and tumors. The complex bone fractures and bone tumors are also located by this imaging procedure. In addition, CT scanning allows the identification of soft tissue abnormalities and the detection of part of organ tumors [64].

2.2.3 *High-resolution computed tomography*

HRCT is a scanning branch of computed tomography, which involves specific techniques to enhance image resolution, being obtained from a set of improvements in the scanner hardware and software that are used in the reconstruction of the scan images. The HRCT is based on a chosen set of imaging parameters that allows maximizing spatial resolution. One of these parameters is the thickness of the slices, being that in these cases narrower slices allow greater spatial resolution. A high spatial frequency algorithm that allows to decrease contrast resolution and to increase the visibility of image noise, improving the spatial resolution, is also an important feature of the HRCT mechanism. Faster scans are used to reduce the appearing motion artifact in the images. Targeted reconstruction is also used in necessary conditions. The quick scans and the target reconstruction allow selected areas to be viewed close to the maximal spatial resolution of the scanning system. The scan data are manipulated in digital form by appropriated software to produce the final image [65].

HRCT is a widely used technique in the diagnosis of various pathology. However, its application is higher in lung diseases by allowing the access to the lung parenchyma through the thin slices [65]. This CT protocol produces extremely high definition images of lung alveoli, airways, interstitium, and pulmonary vasculature [66]. It is used for diagnosis and assessment of ILDs and has at present an important role in the investigation of diffuse parenchymal lung disease and bronchiectasis [67, 68].

As the CXR and the common CT, the application of HRCT scanning presents some setbacks, being one of them the exposure to radiation [69]. In addition, the complexity of HRCT data analysis may be a problem requiring advanced technical support for the correct identification of certain clinical conditions. The fact that this technique is unsuitable for assessing the soft tissues and blood vessels is also a disadvantage [65].

2.3 COMPUTER-BASED SYSTEMS AS MEDICAL ASSISTANTS

2.3.1 *Computer aided-diagnosis and automated computer diagnosis*

In the 1960's, early studies on quantitative analysis of medical images by computer were reported [70], with the scientists assuming that computers could replace radiologists in detecting abnormalities and making the final diagnosis. This idea still defines the present conception line of though underling automated computer diagnosis. More recently another approach gained popularity being based on the use of he computer outputs by radiologists, helping them and enhancing their capability and effectiveness without the intention to replace them. This idea formed the current concept of computer-aided diagnosis (CAD), which spread quickly and widely [71].

At present, the concepts of automated computer diagnosis and CAD are both under strong development particularly in what regards to medical image analysis algorithms. The biggest difference between CAD and automated computer diagnosis is the way in which the output produced by the computer is utilized for the diagnosis. In the case of CAD, the output is integrated by the radiologists as a "second opinion", complementing its final decision on the case. The output helps the radiologist to form its opinion with a higher or lower level of confidence. The potential medical gain with CAD is provided by the synergistic effect obtained by the radiologists competence and computer performance. In the case of automated computer diagnosis, the computer output is used to define the final diagnosis. So, while the level of performance of CAD is achieved by the final decision of the physician, gathering the complement of computers output, the performance level of the automated computer diagnosis is solely based on the computer output. In this sense, the computer levels of sensitivity and specificity must be equal or surpass that of a specialized physician, which is in most cases still hard to achieve [71].

2.3.2 *Computer-aided diagnosis applications on medical imaging*

CAD systems can be defined as a technology that assists doctors in the interpretation of medical images, both in detection and diagnosis. So, CAD systems are divided into two different groups: Computer-aided detection systems and CAD systems. Computer-aided detection systems are engines geared for the location of lesions in medical images. On the other hand, CAD systems perform the characterization of the lesions, for example, the distinction between benign and malignant tumors [72].

From the moment that systematic research of various CAD schemes was begun in the early 1980's, its applicability was emerging as one of the major research subjects in medical imaging and diagnostic radiology [71].

Nowadays, CAD can be defined as a research field widely dedicated to medical image analysis, based on the development of algorithms that complement the diagnosis decision of the physicians. Current research in CAD explores the detection and classification of images of many types of lesions obtained from various imaging methods, such as CXR and CT [7].

CXR is still the most used diagnosis method on lung diseases, such as pulmonary nodules, tuberculosis or other ILDs. A CXR contains a lot of information about the health of the individual under scrutiny but the correct interpretation of this complex information keeps a major challenge for the physicians due to the complexity to distinguish specific lesions or, in some cases, even healthy from non-healthy tissues. Thus, it is conceivable that pertinent information about lung disease or abnormalities could be missed by routine visual examination of CXR. The intensive use of CXR created the demand for the development of CAD systems to enhance the detection of lesions that are easily missed, improving the accuracy of diagnosis while continuing to take advantage of the experience of a trained radiologist [8].

Over the last few years, the development of AI techniques combined with the accumulation of large sets of medical imaging data has opened up new opportunities for the construction of new CAD and computer-aided detection systems for medical applications. Some CAD schemes for detecting, classifying and diagnosing lung diseases have been developed by using CXR and CT images. [73].

2.4 ARTIFICIAL INTELLIGENCE

AI can be defined as a technology branch of conjugating science and engineering aiming at the computational understanding of intelligent behavior and creating mechanisms that exhibit such behavior [74]. The definition and development of AI begun after World War II, with the disclosure of Alan Turing's article "Computing Machinery and Intelligence" [75]. The AI field draws on knowledge from computer science, information engineering, mathematics, psychology, linguistics, philosophy, biology, and many other scientific fields. Although the term has existed for many years, it has only recently begun to popularize due to the development of new technologies, the increasing amount of data, advanced algorithms and improvements in both power and computational storage [76]. With the appearance of the modern computer, more sophisticated, AI began to gain the means and critical mass to establish itself as an integrated science with its own methodologies.

AI systems achieved some forms of reasoning ability, learning, pattern recognition, and inference [77]. As for reasoning ability, the AI methods intend to apply logical rules to a set of data to obtain a certain conclusion. When it comes to learning, the goal is to learn from mistakes and hints in order to act more effectively throughout the process. In pattern recognition, AI systems

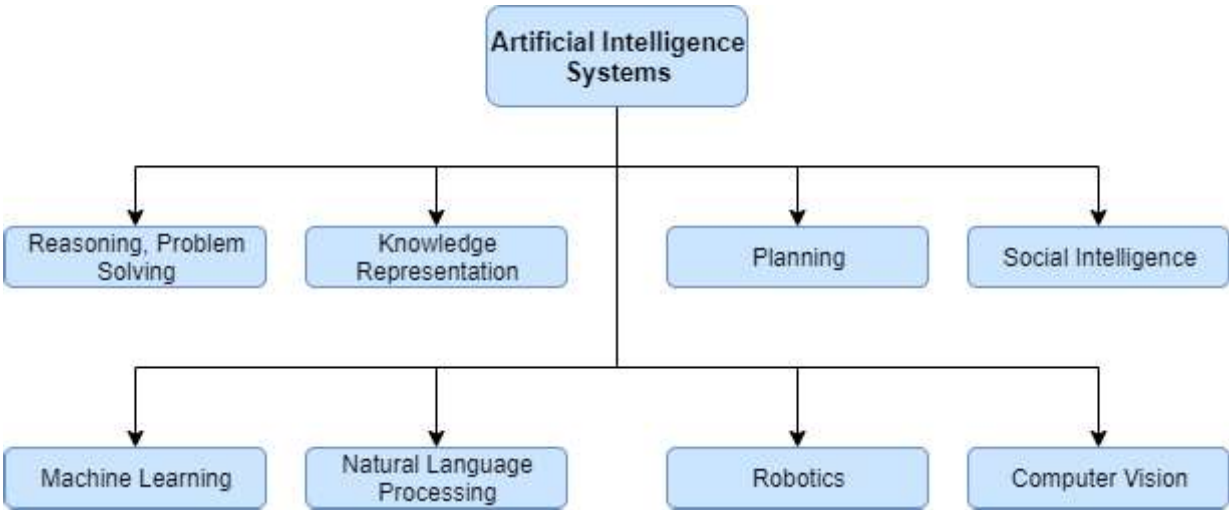


Figure 3: Classification of AI systems. Adapted from [78]. AI, Artificial Intelligence.

aim to identify and classify specific patterns from images and other signals. The AI inference is intended to apply the reasoning to various situations of the diversified human every day. This set of features can divide the AI systems into a set of various AI classes (**Figure 3**). AI systems have recently evolved in some strands, mainly the areas of computational vision, voice analysis, diffusive logic, and artificial neural networks. Many of these strands have the potential to be applied in the most diverse areas of society, always involving human behavior.

The modern medicine presents as a major problem the solving of complex clinical problems, being this due to the existence of the enormous amount of knowledge that needs to be acquired, analyzed and applied to resolve such problems. AI is capable to analyze complex medical data, presenting a great potential to diagnosis, treatment and predicting outcomes in many clinical scenarios. The medical AI development must assist the medical staff in the formulation of a diagnosis, leading to assertive therapeutic decisions. AI systems are projected to support healthcare workers, assisting them with tasks that depend on the manipulation of data and knowledge [79].

The increasing amount of healthcare data and the fast development of methods that can analyze big data have promoted the enthusiasm around AI applications in healthcare. In order to overcome complex clinical problems and, besides presenting high potential in the diagnosis, treatment and prediction, powerful AI techniques harbor the potential to unlock clinically relevant information hidden in a large amount of data, assisting clinical decision making [80]. From here, these systems can help to reduce diagnostic errors that are inevitable in the normal human clinical practice [11].

The development of AI methods into healthcare applications requires a training process through the use of real data generated from clinical activities, such as diagnosis, screening and

treatment [11]). In the diagnosis stage the largest part of the AI literature data comes from medical imaging, genetic testing, and electrodiagnosis (**Figure 4**). This training process allows AI systems to learn similar groups of subjects and associations between subject features and outcomes of interest [11].

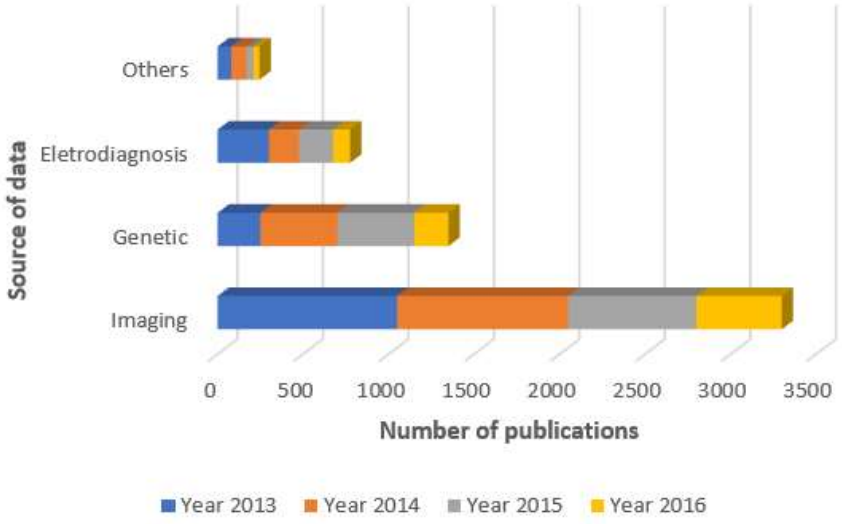


Figure 4: Origin of medical data used in artificial intelligence literature in the last years. Adapted from [11]

Presently, the AI techniques translated into useful medical applications can be divided into two major categories (**Figure 5**). The first category includes ML techniques that analyze structured data, mainly imaging, genetic and electrophysiological data [11, 81]. These techniques can be further sub-grouped into classical techniques, such support vector machines (SVM) and neural networks, and into the more recent technique, DL [11]. The second category includes the natural language processing methods that allow extracting information from unstructured data, such as clinical notes and medical journals, to complete and enrich structured medical data [82].

The IBM Watson system [83] is the pioneer in the application of AI systems in healthcare, based both in ML and natural language process. The IBM Watson system obtained promising results in cancer research, in which 99% of the treatment system recommendations are coherent with the physician decisions [84]. Nowadays, Google is taking a crucial role on healthcare domain, by applying AI to disease detection, new data infrastructure, and potentially insurance. In health care, Google has designed tools to assess heart diseases risk, predict patient's overall risk of premature death and, more recently, detect breast cancer [85, 86].

Since this work is aimed at the classification from the analysis of medical imaging data, we focused on the first aforementioned category and do not address in detail natural language processes.

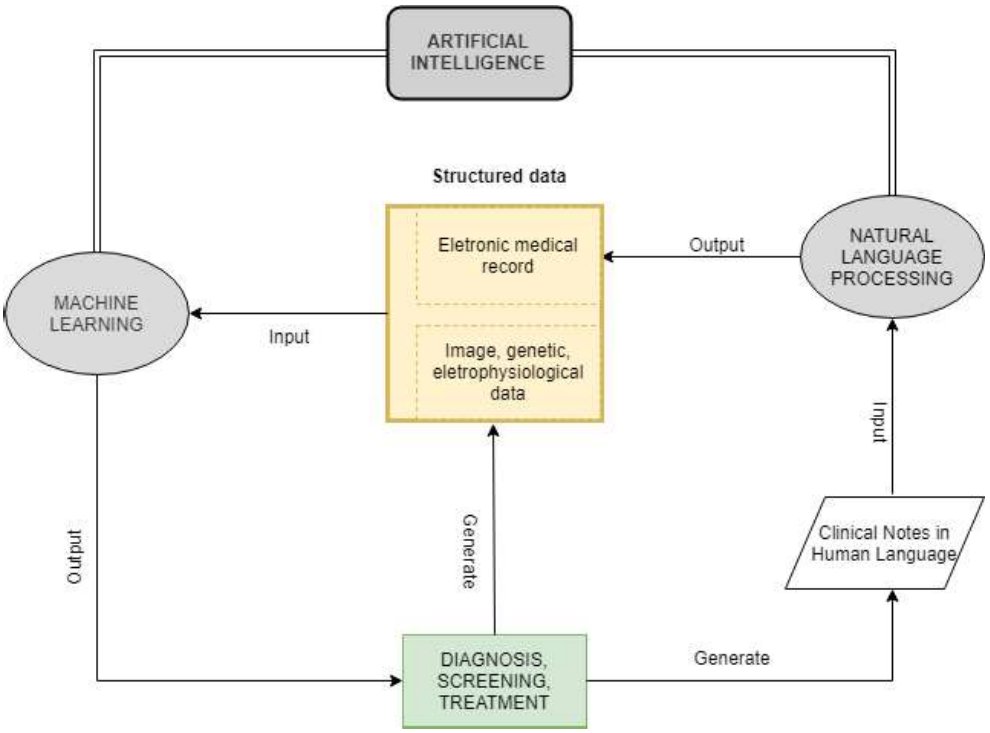


Figure 5: The road map from clinical data generation to ML analysis and natural language processing to data enrichment. Adapted from [11]. ML, machine learning.

2.5 MACHINE LEARNING

The ML can be defined as a set of automatic computing procedures based on logical or binary operations that learn tasks from a set of examples [87]. So, ML studies the computer algorithms that can learn complex relationships or patterns from empirical data and make accurate decisions [88]. It is considered a branch of AI, enabling the collection and extraction of patterns from examples, which is a component of human intelligence [89]. In ML, the training procedure uses specific sets of instructions together with large amounts of data and algorithms that confer to the machine the ability to learn how to perform a specific task. The ML methods can be classified, according to the type of learning, in supervised learning methods and unsupervised learning methods [90]. How a ML algorithm is trained to recognize certain features and thereby become able to make accurate predictions on new examples depends on the type of data and the adopted algorithms.

2.5.1 *Supervised learning*

The supervised learning method aims at the prediction of a known output or target, at the same time that the computer algorithm tries to approximate human performance [11]. Supervised learning focuses on classification, involving the choice of subgroups to best describe a new data instance, and prediction, which involves estimating an unknown parameter [91]. The input consists of a set of training examples with recognized labels or features (**Figure 6**). Because input data and response values are already identified, the algorithm is improved to can make iterations until it reaches an agreed-upon result. A supervised learning algorithm analyzes the training data and produces an inferred function. Supervised learning is the most commonly applied ML approach in radiology, more properly in medical image analysis [11]. Inside radiology and other medical procedures, supervised learning is commonly used to diagnose or predict disease outcomes. Each case, in the input dataset, is characterized by a category label. The algorithms generate a function that maps the dataset to the predefined categories by minimizing the classification error. Both supervised and unsupervised learning tasks can be combined for the detection and prediction of disease outcomes [92].

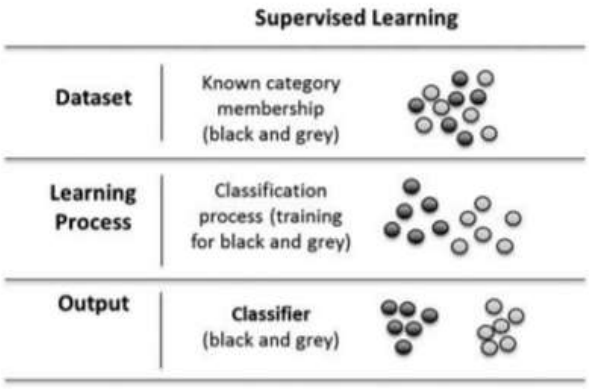


Figure 6: Illustration of supervised approach, when the category membership is known. Adapted from [92]

The supervised learning techniques most frequently used are logistic regression, random forest, linear regression, linear discriminant analysis, SVMs, ANNs, bayesian classifiers, k-Nearest Neighbor, and decision and classification trees. In medical applications SVMs and ANNs are the most commonly used [11].

2.5.2 *Support vector machines*

In ML, SVMs are supervised techniques with associated learning algorithms that analyze data used for classification and regression analysis [93]. Through a set of training examples, where each example is labeled as belonging to one of two categories, a SVM training algorithm builds a model that assigns new examples to one of the categories. A SVM can be defined as a representative model of the examples as points in space, mapped so, that the examples of the different categories are separated by a hyperplane. The hyperplane is defined as a plane that exists in the hyperspace (the new way of the space) that can optimally separate the different classes. More properly, it is the separation line between the categories, found by the SVM algorithm. This plane attempt to maximize the distance between the two nearest points of each category. Finally, the new examples are mapped in the same space and predicted to belong to one of the predefined categories, based on which side of the gap they fall into [94].

In medical application, SVMs can transform input data in a way that produces the widest plane, or support vector, that allows the separation between the two categories. Basically, they are used for classifying the subjects into two groups, where the outcome Y_i is a classifier: Y_i = -1 or Y_i = 1 [11]. This value represents whether the i_{th} patient is grouped in category 1 or 2, respectively. The basic assumption is that the subjects can be separated into two different categories through a decision boundary defined on the traits X_{ij}, which can be determined by the following equation:

$$a_i = \sum_{j=1}^{p} w_j X_{ij} + b,$$

where W_j is the weight placed on the j_{th} trait to manifest its relative importance on affecting the outcome among the others. The decision rule states that if a_i >0, the i_{th} patient is considered a member of group 1, that is, labelling Y_i = -1; If a_i <0, the patient is classified to group 2, that is, labelling Y_i = 1 (**Figure 7**).

The category memberships are indeterminate if the spacial points have an $a_i = 0$. The training goal of SVMs is to find and set the optimal W_{js} in order that the resulting classifications agree with the outcomes as much as possible, that is, with the smallest misclassification error, i.e. the error of classifying a patient into the wrong category. The best weights must allow two important facts: first, the sign of a_i must be equal as Y_i so the classification is correct. Second, $|a_i|$ must deviate from zero to minimize the ambiguity of the classification [11]. These can be obtained by selecting W_{js} that minimize a quadratic loss function [95]. If the new patients come from the same training population, the obtained W_{js} can be applied to classify these new patients based

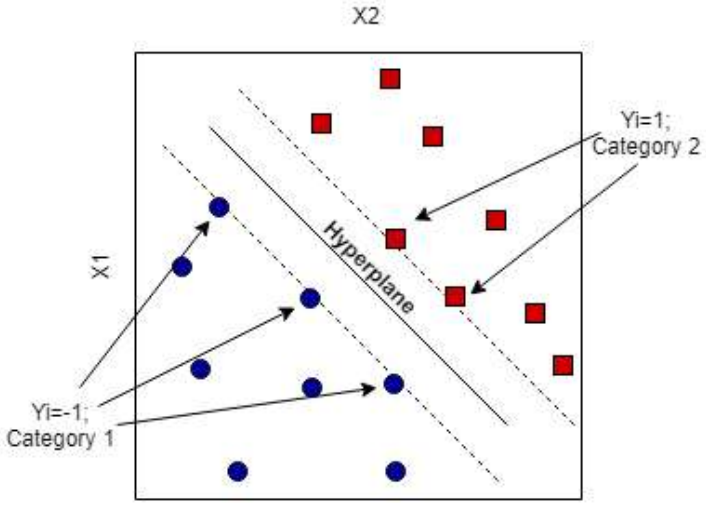


Figure 7: Illustration of SVM classification. SVM, support vector machine.

on their traits. An important estate of SVM is that the determination of the model parameters is considered a convex optimization problem, so the resulting solution is always global optimum. There are many convex optimization tools that can be readily applied in SVM [11].

SVMs can be considered as a powerful method for building a classifier and, compared to other ML methods it is very powerful at recognizing patterns in complex datasets. As an AI method, SVMs can help to recognize specific genomic features or patterns that may represent some genetic diseases or be useful in classifying cancer subytpes [96]. Regarding the application of this method in detection of thoracic diseases, several studies have been done, always obtaining a high predictive efficiency, within 80% [8].

2.5.3 *Neural Networks*

The neural networks, most commonly named artificial neural networks (ANN), is an approach of supervised learning that mimics the human brain in processing input signals and transform them into output signals [97, 98]. They are inspired by advances in neuroscience, such as the interpretation of information processing and communication patterns in the nervous system, more properly in the connection between neurons, i.e. neural networks in the sense of biology [9]. The input signals are received by dendrites of a neuron from environmental stimulation or other up-stream neurons. Then, the signal is processed in the cell body and transmitted along the axon to the output terminal. Finally, the output signal is received by downstream neurons or by the function organs to make a reaction. The work of a single artificial neuron is based on this principle [99]. This single neuron, firstly called perceptron, was first proposed by Frank

Rosenblatt in 1957 [100] (**Figure 8**). The perceptron consists of one or more inputs, a processor, or "neuron", and one output. If a network only contains two nodes, the input, and the output nodes, is designed single-layer networks. Yet, such networks can be handled in order to do more complex tasks, by adding one or more hidden layers, being then called multilayer network [98].

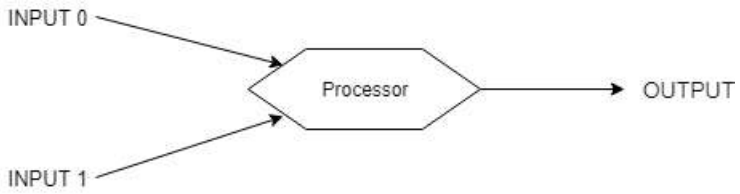


Figure 8: Single neuron: perceptron. Reading from left to right: inputs come in, output goes out.

The ANN consists of several individual elements, designated by nodes or "neurons" [92]. Basically, these "neurons" have as function the read of the input, processing it, and the generation of an output [99]. In ANN topology, input nodes receive feature variables from raw data and the output node applies an activation function to combined information from input nodes [98]. "Neurons" are connected to each other into different layers: one input layer to input data, one or more hidden layers with different neuron connection weights and one output layer to produce the classification (**Figure 9**).

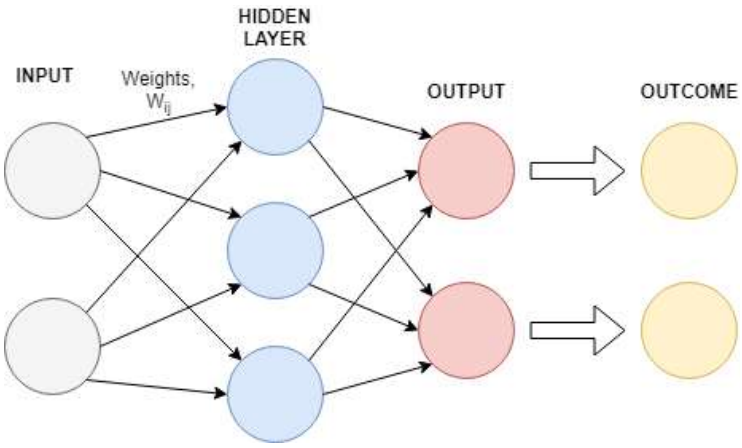


Figure 9: Artificial neural network with one hidden layer.

The associations between the outcome and the input are depicted through multiple hidden layer combinations of prespecified functionals [11]. These kinds of connections between "neurons", input, output, and hidden layers allow the collective processing of the information, in parallel throughout the network of "neurons". By this basic principle, the ANN can be considered as a connectionist computational learning system. In a network of many "neurons", rich and intelligent-like behaviors and faster processing systems can be obtained [99].

The ANN aims to find the ideal weights through input and outcome data so that the average error between the outcome and the resulting classifications is minimized, i.e. the prediction error: $\sum_{i=1}^{n}(Y_i - a_i)^2$ [11]. The ANN is trained with the input data, being the obtained output compared with the known input labels. If the new data come from the same training population, the resulting weights can be applied in the prediction of the new outcomes based on their specific traits [95].

One of the key elements of an ANN is its ability to learn patterns during the training process. So, an ANN can be classified not just as a complex system, but also as a complex adaptative system. Its adaptative ability means that the network can change its internal structure based on the information that receives and flows through it [99]. If some classification errors occur, the process is repeated until the errors are minimized in order to improve subsequent results [11]. This can be done by adjusting the weights, i.e. the number that controls the signal between two neurons, that minimize the prediction error [99]. The error minimization can be performed through standard optimization or gradient descent optimization algorithms [11]. In short, the neural network is designed to adapt to itself, by changing the weight values according to the expected quality of the output.

The use of neural networks in health care and other areas was firstly discouraged due to weak processing power and low sets of data. However, in the past 20 years, due to advances in computational ability, more properly the enhanced computing power, the larger availability of big data and novel algorithms to train the networks, these methods have back into the stage as a relevant viable tool for a variety of tasks [10, 101]. Some studies have been made, suggesting that ANN have a big potential to perform better than human in some visual and auditory recognition tasks, which may be very useful in medicine and healthcare applications [10]. For example, it has been studied the application of ANN variants in the classification of patterns of various diseases, using as input medical images. Dheeba et al [102] developed a neural network to predict breast cancer, using as input the texture information presented by mammography scans. In the same thought, [3] used CNNs to predict the classification of the most frequent ILDs patterns using CT scans as input.

With the development of more complex ANN architectures encouraged by the increase of computational power, a new set of ANN classes appeared, more commonly known as DL models [101]. These models are a subgroup of ANNs characterized by an increasing number of hidden layers in order to improve prediction from big data. Inside of these hidden layers, a larger set of activation functions are implemented, making deep neural models a good option for multifarious tasks, such as natural language and image analysis [103]. Nowadays, DL techniques are currently gaining a lot of attention for its utilization in big healthcare data, exhibiting impressive results in mimicking humans performance in various fields, such as medical imaging [10].

Overall, with the fast development of computer technology, internet, statistics, ML, ANN and DL models, in addition to the increase in handheld networked devices, AI technology is beginning to apply revolutionary changes in the society and medical and healthcare practices are no exception [104]. It is not expected that medical AI systems will completely replace clinical work but will with certainty play a focal role in patient monitoring, storage and automated analysis of electronic health records, diagnosis and even decision of treatment protocols in some settings. Together with these tasks, the medical AI models have also the potential to be protagonists at patient care, in particular in the areas of robotic surgery and health system management [11].

2.5.4 *Unsupervised learning*

Contrary to the supervised learning task, in unsupervised learning, there are no outputs to predict (**Figure 10**). This task aims at finding naturally occurring patterns or groupings within data. The goal of unsupervised learning is to model the underlying structure or distribution in the data, in order to learn more about the data and discover hidden signals within [91]. In an unsupervised learning model, the input is a set of unlabeled examples without predefined categories [89]. Inside the set of medical applications, unsupervised learning techniques are commonly used to identify patterns of diseases but, in comparison to supervised learning, have low foreseen applicability in radiology [92].

Clustering and principal component analysis are considered the two major unsupervised learning methods. The most popular clustering algorithms are k-means clustering, hierarchical clustering, and Gaussian mixture clustering. The clustering algorithms consist of the formation of natural clusters, based on specific similarity criteria between the input dataset cases. The principal component analysis method is mainly used for data dimension reduction, being this reduction able to prevent the loss of important information on the subjects. The most applications of AI in healthcare are based on supervised learning, but the unsupervised methods are also important and can be used as part of data preprocessing steps leading to a more efficient follow-up supervised learning step [11].

2.6 DEEP LEARNING: ALGORITHMS AND ARCHITECTURES

DL was inspired by the human brain and how the information is transmitted and processed in the nervous system [105]. The early structure of the DL technique was derived from ANNs [106], and since then DL has been developed and used in a wide range of technologic fields, including, image recognition, pattern classification, natural language processing, drug discovery and bioinformatics [107]. Simplistically, DL can be viewed as an artificial neural network with

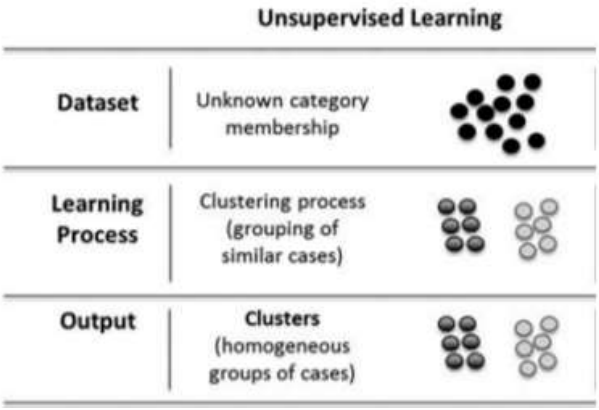


Figure 10: Illustration of unsupervised approach, when the category membership is unknown. Adapted from [92].

more than one hidden layers, attempting to model abstraction from large-scale data, such as images, videos, sounds and texts [11]. A DL network is composed of a set of nodes, but only a few nodes will contribute to the final output and possibly with different weights. DL aims to adapt the weights across the network, by changing their value in order to get the right nodes operating [105].

In the last decades, witnessed a rapid development of modern computing and a massive growth in biomedical data and medical images, due to the advances of high-throughput technologies [107]. At the same time, the big amount of medical data was presented with several problems in terms of storing, analyzing and interpreting demanding effective and efficient computational tools to process the data and overcome these setbacks [108]. The DL techniques presented as a high potential tool for solving these problems. Studies showed that ANN had remarkable performance in various fields but some limitations at the level of the optimization and influence of overfitting. Inside these problems, researchers attempted to apply deep architectures to determine better solutions. However, its complex operation limited the generalized ability to generate successful models, being that DL applications are still under development [10].

As a general rule, DL is based on two important properties: multiple layers of nonlinear processing units and supervised or unsupervised learning of feature presentations on each layer [109]. In terms of architecture, DL networks consist of a series of stacked layers (**Figure 11**): the first layer (input) represents the observed values in which a prediction is based. The layers between the input and the output called hidden layers (no observable data), allow the network to handle complex data. The last layer (output), produces a specific value or class prediction. This layers-based structure allows the production of more complex decisions based on a combination of simpler decisions [10]. The greatest advantage of DL networks is that each layer produces a certain representation of the input data, which in turn is also used as input for the next represen-

tation level. From here it is possible to pass through several different layers to combine all these representations in order to perform any kind of task. By the way, the fact that a high number of hidden layers are present allows the algorithm to handle complex data with various structures [105].

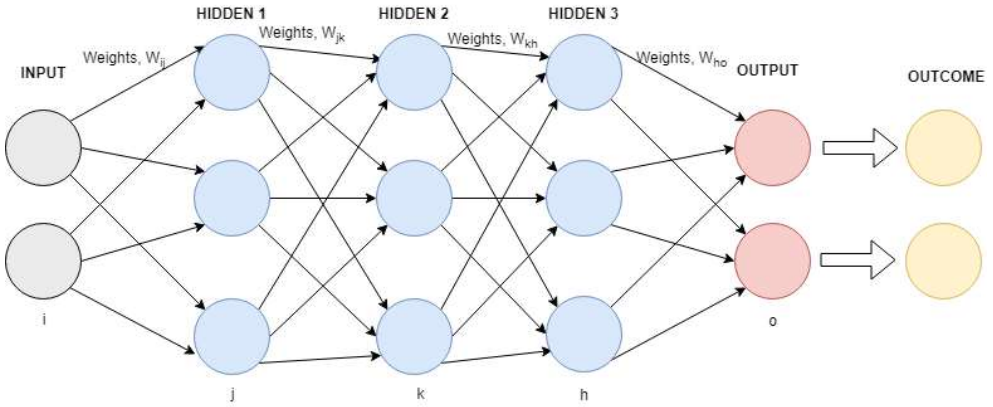


Figure 11: Illustration of a deep learning network with 3 hidden layers.

A DL network must learn by itself and only from the input data. The learning of a network is divided into three main stages: training, validation, and testing. In the training phase, the weights are updated at each step and layer, in order to reach the best model. The learning process can be supervised or unsupervised, according to if the given input presents labels that should be considered in the output. The validation phase assesses the performance of the network by inputting different data from the initial training. In this phase there is an understanding of how well the network behaves with unknown and new data, allowing to find the best model. Finally, in the testing phase, once the best model is chosen, its general performance is evaluated with previously unseen data [105].

The fact that the DL networks or deep neural networks (DNN) present more hidden layers offers a much higher ability for feature extraction from large scale and complex data [110]. The basic presupposition for feature extraction is similar in all types of DNN, in which the network is activated by an input, which then spreads the activation to the final layer along with the weighted connections. In the final, prediction results are generated. Along the process, the network weights are tuned by minimizing the average error between the outcome and predicted data [107].

The build-up of a DL neural network structure is based on the choice of some specific conditions, parameters and algorithms, such as activation functions, optimization objectives and methods, and proper architecture, taking always in account the type of data to be handled [107]. The activation functions form the non-linear layer in all DL structures, which in combination

with the other layers allows the simulation of the non-linear transformation from the input to the output [103]. The selection of appropriate activation functions allows a better feature extraction. The most frequently used activation function in DL is the rectified linear unit (ReLU). This specific function and its variants show superior performance compared with others in many cases [111]. There are other popular activation functions, namely the sigmoid functions, hyperbolic tangent, softmax, soft plus, absolute value rectification, and max out. A DL network also has an optimization objective, composed of a loss function and a regularization term. The loss function measures the discrepancy between the output of the network and the expected results [107].The regularization process involves a set of strategies to reduce the test error. Relating to the optimization methods, they present different advantages and disadvantages to different architectures and loss functions [103]. There are some optimization methods used in DL networks, being the stochastic gradient descent and its variants the most commonly used.

The proper architecture should be selected according to the considered data guiding the choice among the main existing DNN architectures:

- Auto-encoder: is similar to principal component analysis, extracting features from unlabeled data and setting target values to be equal to the inputs. When the number of hidden units, i.e. the dimension of features, is smaller than the input dimension, is performed a reduction of data dimensionality [107];

- Restricted Boltzmann Machine: is a generative stochastic ANN that can learn the distribution of training data by generating the probability distribution and optimizing parameters. Restricted boltzmann machines can also be used in unsupervised learning [107];

- Deep belief network: is built by stacking restricted boltzmann machines [112] or auto-encoders [113] and can learn the distribution of the data or learn to classify the inputs according to given class labels;

- Convolutional neural network (**Figure 12**): within deep neural networks by layers, CNN is the most complex. Its ”neurons” extract features from small areas of the input, which are called as receptive fields. This mechanism of feature extraction was inspired by the visual system in living organisms, where cells in the visual cortex are sensitive only to small regions of the visual field [114]. Beyond the basic structure of DL networks, CNN has implemented more two different types of layers: the convolutional layers, in which the receptive fields change and the number of hyperparameters reduced. These layers are composed of many neurons and are typically used in series after the input; the pooling layers, which function is to reduce the computational requirements progressively through the

network. This kind of layer is used after or interleaved with multiple stages of convolution layers or non-linear layers [3, 107];

- Recurrent neural network: presents itself as the most different DL architecture, being built-in circuit. They represent a hidden-to-hidden recurrence and are applied in sequential data [107].

According to the PubMed database, CNN, recurrent neural network, and deep belief network are the most used DL algorithms in medical applications, being that, in the last years, CNN started to gain an advantage over the others [11].

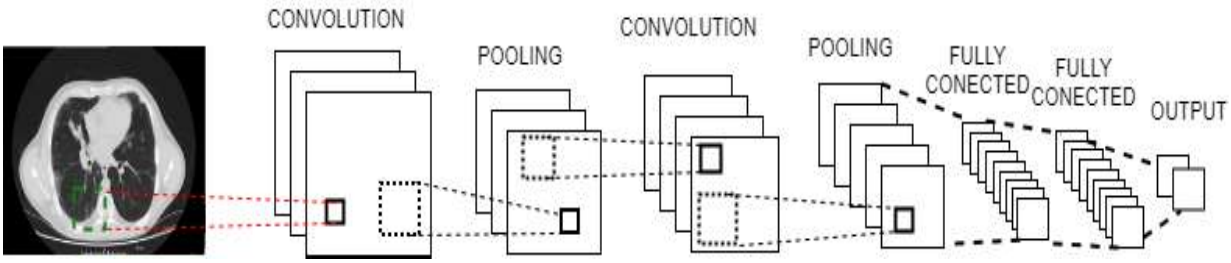


Figure 12: Illustration of a CNN structure: CT image as input; convolutional layers interleaved with pooling layers before applying fully-connected or dense layers. CNN, Convolutional neural network.

2.7 CONVOLUTIONAL NEURAL NETWORKS

2.7.1 *Pipeline*

CNN algorithms adopt the majority steps of typical ML pipeline, including several steps that must be meticulously implemented to achieve the best predictions (**Figure 13**). In the first step, it is necessary to define the problem and how it can be solved. Then, some data preparation must be done in order to build a good model. This data preparation is divided into three main steps: data selection, data pre-processing and data transformation. Data selection allows selecting the relevant samples of the entire data. The pre-processing step consists of removing undesired data and sampling the data. The last step, data transformation, consists of transforming the pre-processed data to a specific format that can be used during the model development and subsequent application [115].

After the data preparation process, the dataset is divided into training, validation and test sets. The data must be correctly partitioned to avoid biased evaluations. The training set must contain the majority of the samples, once they are used for learning features by the model. These learned

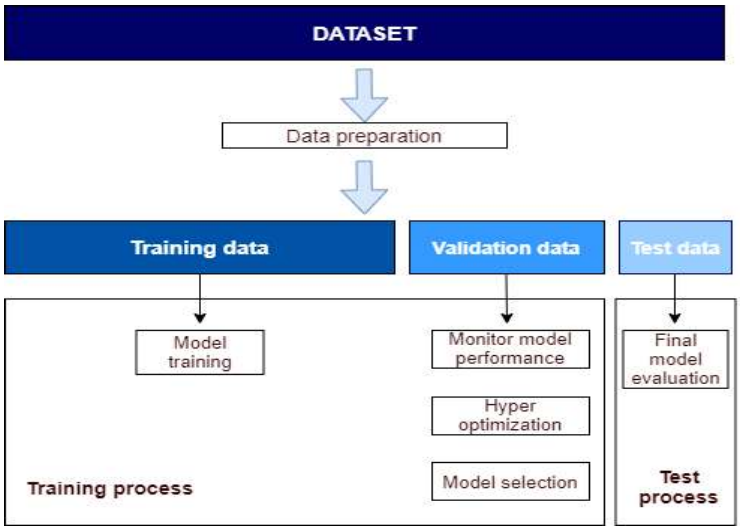


Figure 13: Typical CNN pipeline. CNN, Convolutional neural network.

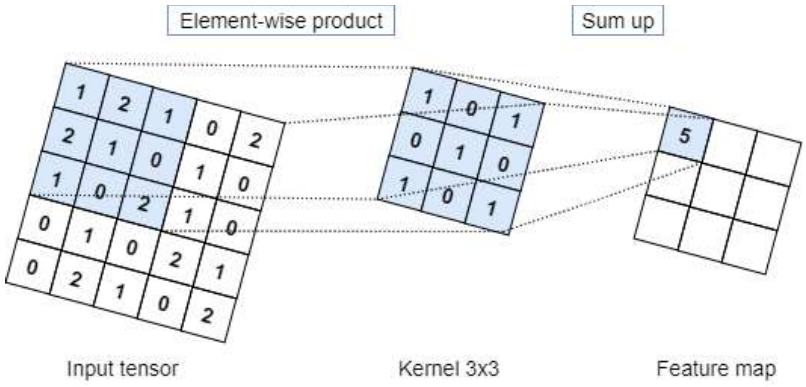
features allow the model to make predictions for new samples. The validation set is used to evaluate the model during the hyperparameter optimization and to select the best model. Algorithms for tuning hyperparameters within a family of models and optimizing model parameters on the training data are needed. Finally, the test set is used for measuring the final model performance. These three sets allow the model to pass through a training and validating process before making an accurate prediction on new examples, for which outputs are unknown.

2.7.2 *Concepts*

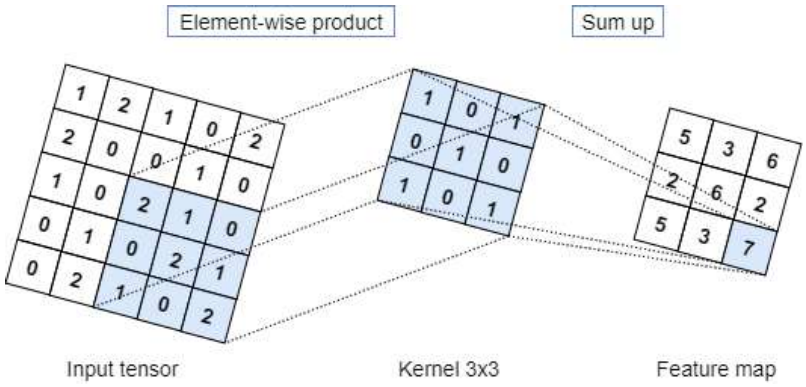
The CNN was first proposed by LeCun et al. [116], becoming at present as one of the most important branches of DL [117]. As said before, CNN is composed of three different types of layers: convolutional, pooling and fully connected layers. The first two, in interleaved series, perform the feature extraction of the input data. Next, the fully connected layers map the extracted features into a final output. These layers work as a classifier, assigning a probability for the object on the image to belong to one of the problem classes [118, 119].

The convolutional layer consists of a combination of a linear operation, namely convolution, and a nonlinear operation, namely activation function [119]. The convolution operation is used on feature extraction, where a small array of numbers, called a kernel, is applied across the input, which is an array of pixel values, called a tensor (**Figure 14**). Then, an element-wise product between each element of the kernel and the input tensor is calculated at each location of the tensor. This product is summed to obtain the output value in the corresponding position of the output tensor, called a feature map [118, 119]. This process is repeated applying multiple kernels

to form an arbitrary number of feature maps [120]. These feature maps are representative forms of the different characteristics of the input tensors. The most common kernel form is 3x3, but 5x5 and 7x7 are widely used too. As one convolutional layer feeds its output into the next layer, the extracted features can hierarchically become more complex [118].



(a)



(b)

Figure 14: An example of convolution operation with a kernel size of 3 x 3, no padding, and a stride of 1. Adapted from [118]

Sometimes, the convolution operation does not allow the center of each kernel to overlap the outermost element of the input tensor and reduces the dimension of the output feature map compared to the input tensor [118]. To solve this issue, the padding technique, typically zero padding is used (**Figure 15**). This technique adds rows and columns of zeros on each side of the input tensor, in order to fit the center of a kernel on the outermost element and keep the same dimension through the convolution operation [121, 122]. Without zero padding, each feature map dimension would be successively reduced after the convolution operation. The conservation of the same dimension through this technique allows adding more layers to the model [118].

Another factor that defines the convolution operation is the stride, i.e. the distance between two successive kernel positions. The stride value is commonly set as 1. Higher values of stride

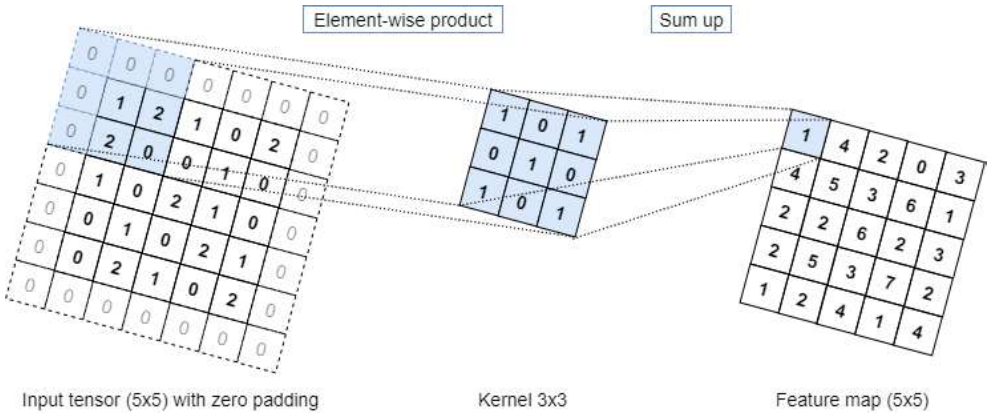


Figure 15: A convolution operation with zero padding. Adapted from [118]

are sometimes used to achieve the downsampling of the feature maps. Yet, pooling operations can also perform downsampling, as described below [103, 118].

After the convolution operation, its outputs are then subjected to a nonlinear activation. This activation allows the model to create a complex mapping between the network's inputs and outputs, leading to a big capacity of learning and modeling complex data, such images and videos [117]. The most common nonlinear activation function used is the ReLU [111]). Other nonlinear functions, such as hyperbolic tangent and sigmoid are also used [123]. After the nonlinear activation steps, pooling layers are usually applied.

The pooling layers consist of a downsampling operation that reduces the dimensionality of the feature maps. This operation allows to introduce translation invariance to small shifts and decrease the number of learnable parameters [117, 124]. The pooling operation can be divided into two different forms: max pooling and global average pooling [103]. The max pooling is the most used form and consists of extract patches from the input feature maps, assign and output the maximum value for each patch and reject all the other values (**Figure 16**). A max pooling with a filter of size 2x2 and a stride of size 2 is commonly used in practice [118]. The global average pooling presents an extreme type of downsampling, where a feature map is downsampled into a 1x1 array by simply taking the average of all the elements in each feature map [125]. At the pooling layers there are no learnable parameters, whereas filter size, stride, and padding are hyperparameters in pooling operations, similar to the hyperparameters presented in convolution operations.

After the last convolutional layer or pooling layer, the output feature maps are flattened. This process transforms the output feature maps into a single one-dimensional array of numbers. The resulting product is then connected to one or more linked fully connected layers. At this point, every input of each fully connected layer is connected to every output by a learnable weight. The

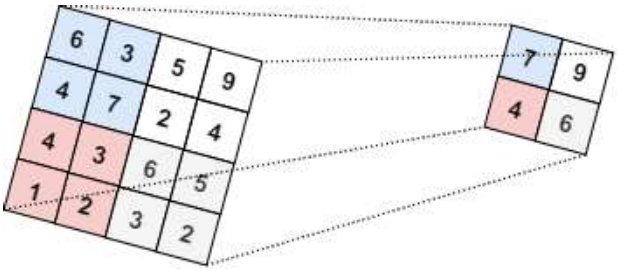


Figure 16: An example of max pooling operation with a filter size of 2x2, no padding, and a stride of 2.

created features on the anterior ”convolutional + pooling” layers are now mapped by a subset of fully connected layers to the final output of the network. This output is commonly expressed as the probabilities for each class in classification tasks, being the output nodes of the last fully connected layer equal to the number of the problem classes [118, 119, 120].

Each of the selected fully connected layers is typically followed by a ReLU function, with the exception of the last layer. In the last fully connected layer, the choice of the activation function must be done according to the original task. Sigmoid and softmax are the most used, being that sigmoid function is applied on binary classification and multi-label classification tasks, and softmax function is applied on multi-class classification tasks [118]. If there is a problem that each input could present more than one correct answer, the sigmoid is more appropriate. On the other hand, if there is a problem that each input can belong to exactly one class, i.e. each input only could present one correct answer, the best choice is the softmax function [120, 126].

2.7.3 *Training and hyperparameters*

The training process of CNN consists basically of finding kernels in convolution layers and weights in fully connected layers which minimize the differences between the predicted outputs and the respective ground truth labels on a training set. This process is commonly unleashed by a backpropagation algorithm. This algorithm allied with a gradient descent optimization algorithm, called optimizer, and a loss function plays an important role in the training process [118]. While the loss function measures the error between output predictions of the CNN and ground truth labels on the training set through forward propagation, the optimization algorithm iteratively updates the learnable parameters, such as kernels and weights, according to the loss value measured by the previous function [103, 127].

The type of loss function and optimizer are one of the many hyperparameters existing on CNNs and need to be defined according to the nature of the task. The cross-entropy loss function is the most used on multi-class classification and the mean squared error is commonly used

on regression problems. As for optimizer, many improved gradient descent algorithms, such as Adam, RMSprop, and stochastic gradient descent (SGD) with momentum are widely used [128, 129, 130].

Inside the optimizer, there are other hyperparameters that have big relevance, namely the learning rate. The learning rate has a crucial role in the training process, once it determines the step size that each learnable parameter is updated [131].

Another important CNN hyperparameter is the batch size, which is commonly used to counteract the memory limitations, more concretely big datasets or big data memory. The whole training dataset is divided into random subsets, called batches, of defined size. Then, for each batch, the gradients of the loss function are measured [118].

Regarding batch size, another important hyperparameter is the number of epochs. One epoch is defined as one forward pass and one backward pass of all the training examples, through the network. It is known that the whole data must pass several times through CNN [131]. However, the number of needed epochs is not fixed, being this dependent on the model capacity, number of examples and complexity of the data.

As shown in the previous subsection, all of CNN's layers have also subsequent hyperparameters, that need to be set before the training process. The selectable hyperparameters for each type of layer are synthesized on **Table 3**.

Table 3: CNN layers hyperparameters

	Parameters	**Hyperparameters**
Convolutional layer	Kernels	Kernel size, number of kernels, stride, padding, activation function
Pooling layer	None	Pooling method, filter size, stride, padding
Fully connected layer	Weights	Number of weights, activation function

The complexity and the high number of hyperparameters are still a problem on CNNs. Their values are set empirically, as they are linked to the problem, the dataset, and the model architecture. There are no good predefined values, as they must pass through a tuning process based on the model's performance [132]. The validation data is hugely important on this step, once it can be used for accurate the model's performance before the final evaluation.

2.7.4 *Overfitting and underfitting*

A good CNN model can be defined as the one that can be able to generalize any input data, giving correct predictions when tested on unseen data. However, this ability can often be affected by two problematic situations very typical on the ML domain, namely overfitting and underfitting.

The overfitting problem consists of a situation where the network learns too well specific features of the training set, leading to a memorizing process of the irrelevant noise. This problem is one of the major challenges of ML, once it decreases the model capacity to generalize on new data and leads to inaccurate predictions. Underfitting is a frequent problem too, where the model cannot fit the data well enough, leading to low generalization and unreliable predictions [118, 133]. A usual procedure for detect overfitting and underfitting during the training process is the monitoring of accuracy and loss curves on the training and validations sets (**Figure 17**). The training process allows evaluating the model performance on the validation set at the end of each epoch. If the model performs poorly on both training and validation sets, it is a signal of underfitting. If the model performs much better on the training set relating to the validation set, it is a signal of overfitting [118]. In general, the longer a network is trained, the greater is its ability to perform in the training set. However, there is always a point where the network fits too well on the training data but starts to lose its ability to generalize well, which can be observed by the increase of the validation loss values.

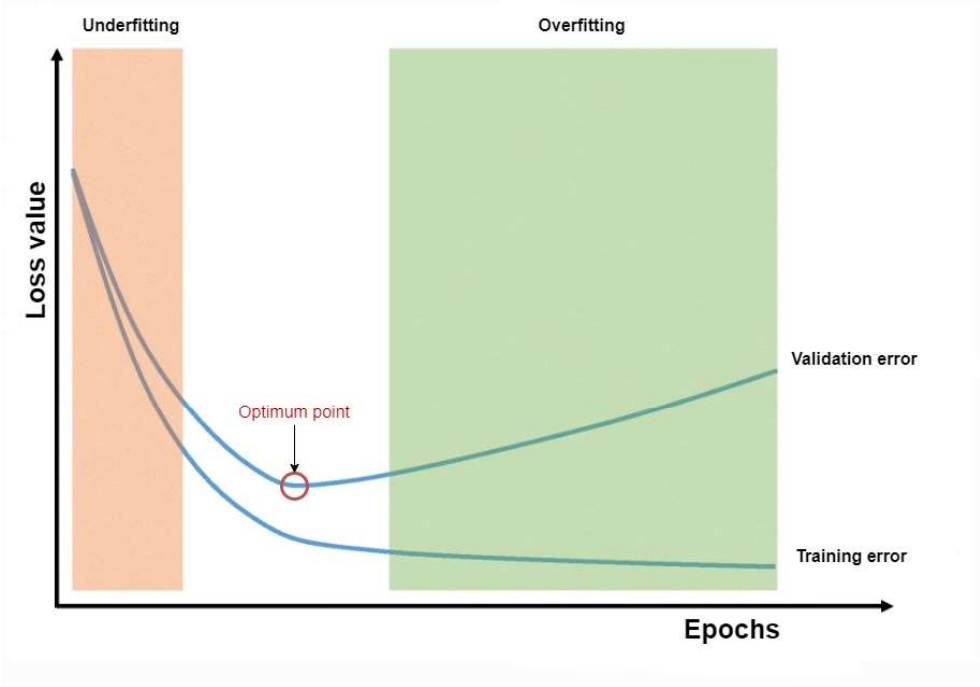


Figure 17: Overfitting, underfitting and optimum point.

Several methods to minimize overfitting were proposed, being adding more training data, data augmentation, regularization, batch normalization and reducing architecture complexity the most common. The most simple and advisable method to combat overfitting is to obtain more training examples. A larger number of training examples leads to a better process of generalizing by the

model, however, in medical imaging, this does not always happen. One way to avoid underfitting problems is to build more complex and deep models [134].

In CNN it is crucial to set the right balance between overfitting and underfitting. This balancing point means that the model must be simple to generalize well but, at the same time, must be not so simple at all in order to fit the data in a perfect way.

2.7.5 *Evaluation*

In DL techniques, such as CNN, there are several approaches and metrics that can be measured in order to evaluate the performance of the models. The choice of the metrics depends on the type of the task, classification or regression and depends on the sample distribution along with the classes.

Regarding classification problems, the most commonly used approach is the confusion matrix, which evaluates the classification accuracy by mapping the predicted outputs with the respective real value. In the simplest case of classification, binary classification, this matrix consists of 2 columns and 2 rows, where columns present the predicted values and rows present the real values (**Table 4**). Each cell of this matrix corresponds to a single group of statistical mean, namely true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The instance of positive cases correctly predicted, and the instance of negative cases correctly predicted are denominated as, TP and TN, respectively. On the other hand, if the predicted instance is positive, but its real value is negative, it is denominated as FP. If the predicted instance is negative but its real value is positive, it is denominated as FN [135, 136].

Table 4: Confusion matrix for binary classification.

Real/predicted	**Negative**	**Positive**
Negative	TN	FP
Positive	FN	TP

From the confusion matrix results, it is possible to compute some important performance metrics such as the accuracy, specificity, recall, precision and f1 score (**Table 5**). The accuracy formula consists of dividing the number of examples correctly predicted by the total number of examples. The specificity reveals the percentage of negative cases correctly identified and is calculated by dividing the number of TN by the total number of real negatives (FP+TN). The recall, also called sensitivity, indicates the percentage of real positive cases that are correctly predicted, and is calculated by dividing the number of TP by the total number of real positives (FN+TP). The precision is defined as the proportion of positive cases correctly predicted in the total of positive cases predicted. Precision is calculated by dividing the TP by the sum of all

positives (TP+FP). Although recall and precision are mostly applied to the positive class, they can also be applied to the negative class for more detailed studies. Finally, the F1 score can be defined as a weighted average measure between precision and recall values [137].

Table 5: Classification metrics formula.

Accuracy	$\frac{TN+TP}{TN+FP+TP+FN}$
Specificity	$\frac{TN}{FP+TN}$
Recall	$\frac{TP}{FN+TP}$
Precision	$\frac{TP}{FP+TP}$
F1 score	2 x $\frac{precision \times recall}{precision + recall}$

2.8 CONVOLUTIONAL NEURAL NETWORKS ON MEDICAL IMAGING

Millions of people suffer from chest diseases every year, being tuberculosis, pneumonia, and lung cancer the most common [71]. As said anteriorly, DL is actually emerging in several fields, being its application on healthcare one of the most promising. Detection of lesions and abnormalities is the major issue in medical image analysis [107]. This renewed AI area can be applied to perform automatic lesion detection and differential diagnoses [10]. Over the last years, many articles related to the application of DL in the medical field have been published. According to data from PubMed, the application of DL in medical research nearly doubled in 2016 (**Figure 18**). Most of these publications involve the application of DL on detection and classification of abnormalities and segmentation of regions of interest, such tissues and organs [11].

The recent development of AI combined with the accumulation of medical images brings new opportunities for building CAD systems in medical applications. DL emerges as a big potential theme that can be applied in these cases, replacing the feature extraction and disease classification stages in traditional CAD systems. Specially CNNs, which automatically learn image features to classify chest diseases, have become a mainstream trend [8]. There are some developed CAD systems for pulmonary diseases, but most of them only focus on identifying single patterns, such as nodules [138].

In some cases, it is difficult to distinguish between normal and abnormal tissue based on lung texture. This difficulty combined with the hardest abnormalities detection by classic methods leads to a need for improvements in medical tools capability to overcome these problems, being the researches on intelligent detection systems, such AI-based CAD systems, a promising

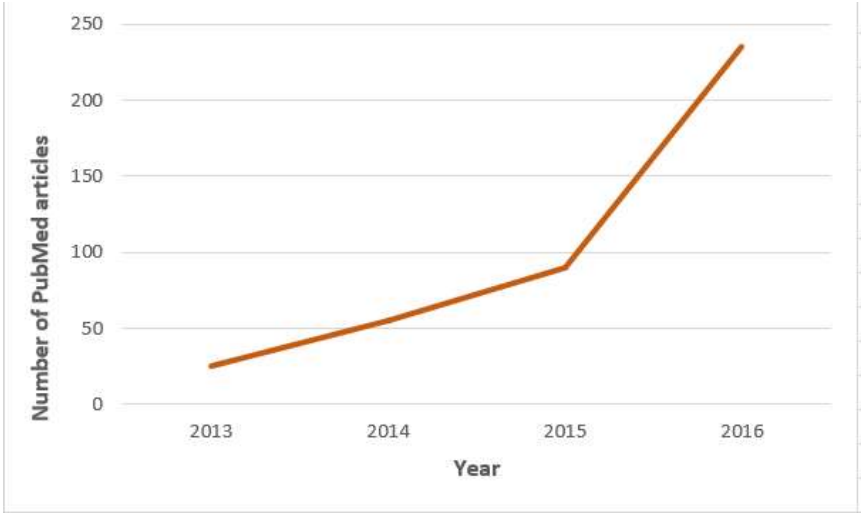


Figure 18: Current evolution of deep learning articles in medical applications. Adapted from [11]

approach in medical image future analysis. Some CAD systems have been build to detect ILD in CXR through changes in lung texture. The Kun Russman Laboratory in Chicago [139] developed a CAD system that divides the lung into multiple regions before analyzing it, to determine the presence of abnormalities. Then, an neural network was implemented to classify suspicious abnormalities. Another example is the work done by Plankis et al [7], which developed a CAD system for ILD detection. This system was developed to detect a variety of pathological features of lung tissue based on an algorithm that can divide the lung region into different regions of interest. However, with the extensive application of DL on lung disease detection, the related literature has more preference on CT datasets rather than CXR datasets [8]. In large medical image datasets, DL methods are designed to classify each pixel to be a lesion point or not, and this can be made by using a DNN or a fully-connected convolutional network [107]. For example, [3] proposed a CNN method to classify the 7 most relevant patterns present in ILDs, obtaining an efficiency of 86%, which outperformed other literature methods. As a future projection, it is intended to extend this method to three-dimensional data and to integrate it into a CAD system as a supportive tool in the differential diagnosis of ILDs. Other similar works have been done, Gao et al [12] proposed a holistic classification of ILDs imaging patterns, using entire CT slices as input [12]. The authors also used a CNN and, despite being different from other image patch-based algorithms, have demonstrated some promising advantages, addressing a more practical and realistic clinical problem.

2.9 PYTHON LIBRARIES FOR DEEP LEARNING

There are many DL frameworks and software libraries in open source for ready use. The most used ones are Torch, Caffe, Theano, MX net and TensorFlow. In addition to these, there are others that were designed to promote ease of use, such as Keras, Lasagne and Blocks. All of these frameworks and software are free and facilitate the application of DL algorithms and methods in several fields. Regarding python language, the most used framework is Keras, which consists of a neural networks library [107]. Besides DL libraries, there are other important packages that play an important role in the preprocessing data and model development steps.

NumPy is a python package that allows efficient manipulation and transformation of n-dimensional arrays and matrices. This package consists of a large collection of mathematical functions to work with these structures. NumPy is the fundamental package for scientific computing with Python [140].

Scikit-learn or Sklearn is a library that consists of many unsupervised and supervised learning algorithms for Python language. It provides a user-friendly interface and a simplest user guide for both software experts and non-experts. Scikit-learn is built upon other useful python libraries such as NumPy, pandas and Matplotlib [141, 142].

Matplotlib is a python package that consists of the produce of high-quality two-dimensional (2D) graphics. This package provides several tools for visualization and design of plots, being very useful for interactive graphing, scientific publishing and user interface development [143].

Pandas is another important python library widely used on data manipulation and analysis. It provides fast, flexible and expressive data structures (e.g. data frame) designed to make working with labeled data in an easy way. It is very important in the pre-building of ML and DL models. Furthermore, this library is very useful in handling CSV and excel files and can provide efficient summary statistics [142].

OpenCV or cv2 is a python package widely used for image processing. It consists of programming functions aimed at real-time computer vision, being able to do all the operations related to images and their pixel data [142].

Talos is a python package and consists of applying hyper optimization methods on Keras models. It allows data scientists and data engineers to achieve complete control of their Keras model. Talos provides grid, random and probabilistic hyperparameter optimization strategies with the aim of maximizing efficiency both on search process and model performance. Among other hyper optimization methods, Talos provides the simplest and most powerful available method for hyperparameter tuning on Keras workflow [144].

3

METHODS

In this section of the dissertation, the practical methods will be described in detail at the same time as their choice is explained. The programming language used in this work was Python resorting to the Spyder [145] environment to ensure reproducibility and advanced development functionalities. Google Colab [146] was the selected computational resource to accelerate the DL model development, hyper optimization and performance evaluation. All the code, functions, files and data used in this dissertation can be accessed at https://github.com/Nunorbc/Pneumonia_classification.

3.1 DATASETS SELECTION

In a global context, there are not many medical large-size imaging datasets available for research. Most data are highly fragmented in different hospital facilities and dispersed across different and non-interoperable platforms. Furthermore, access to medical data even without compromising the privacy of individual patients is still a slow and highly bureaucratic process. As a relevant exception to this scenario was a collaborative effort by the Radiological Society of North American (RSNA), US National Institutes of Health, The Society of Thoracic Radiology and MD.ai that collected and made available a large CXR dataset including images from individuals with and without pneumonia. This was one of the datasets selected for this work (referred in the thesis as dataset XP1) since the success of DL model development is largely dependent on the use of datasets that are large and well-controlled.

Regarding XP1 structure, it is composed of 29 684 images with a resolution of 1024x1024 pixels in the Digital Imaging and Communications in Medicine (DICOM) format. The DICOM format standard is a non-proprietary digital image format used for the communication, management, and storage of medical imaging information that is widely adopted by the medical community [147]. In addition to the pixel image data optional tags to store, patient and exam information can also be included as metadata in the DICOM file. Images in this dataset are annotated with several useful patient and exam-related tags (**Figure 19**). Most importantly, all images

corresponding to a single patient are identified with the same and unique patient id. Images can be divided according to the patient's gender and also according to age. Exam-related tags include the information if the images refer PA view or AP view.

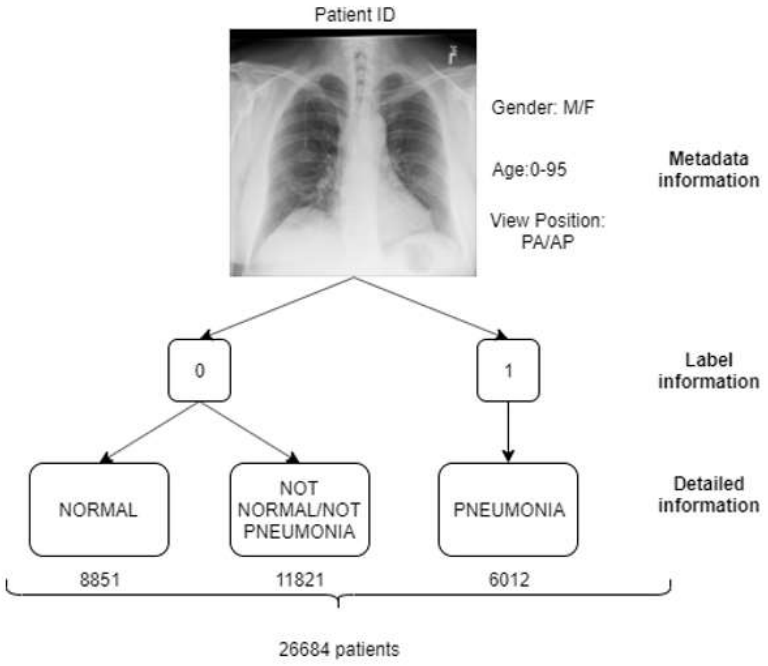
```
(0008, 0005) Specific Character Set            CS: 'ISO_IR 100'
(0008, 0016) SOP Class UID                     UI: Secondary Capture Image Storage
(0008, 0018) SOP Instance UID                  UI: 1.2.276.0.7230010.3.1.4.8323329.12323.1517874364.297773
(0008, 0020) Study Date                        DA: '19010101'
(0008, 0030) Study Time                        TM: '000000.00'
(0008, 0050) Accession Number                  SH: ''
(0008, 0060) Modality                          CS: 'CR'
(0008, 0064) Conversion Type                   CS: 'WSD'
(0008, 0090) Referring Physician's Name        PN: ''
(0008, 103e) Series Description                LO: 'view: PA'
(0010, 0010) Patient's Name                    PN: '00ec0d87-8e2a-44bd-b79c-4b91cfabcacf'
(0010, 0020) Patient ID                        LO: '00ec0d87-8e2a-44bd-b79c-4b91cfabcacf'
(0010, 0030) Patient's Birth Date              DA: ''
(0010, 0040) Patient's Sex                     CS: 'M'
(0010, 1010) Patient's Age                     AS: '49'
(0018, 0015) Body Part Examined                CS: 'CHEST'
(0018, 5101) View Position                     CS: 'PA'
(0020, 000d) Study Instance UID                UI: 1.2.276.0.7230010.3.1.2.8323329.12323.1517874364.297772
(0020, 000e) Series Instance UID               UI: 1.2.276.0.7230010.3.1.3.8323329.12323.1517874364.297771
(0020, 0010) Study ID                          SH: ''
(0020, 0011) Series Number                     IS: "1"
(0020, 0013) Instance Number                   IS: "1"
(0020, 0020) Patient Orientation               CS: ''
(0028, 0002) Samples per Pixel                 US: 1
(0028, 0004) Photometric Interpretation        CS: 'MONOCHROME2'
(0028, 0010) Rows                              US: 1024
(0028, 0011) Columns                           US: 1024
(0028, 0030) Pixel Spacing                     DS: ['0.168', '0.168']
(0028, 0100) Bits Allocated                    US: 8
(0028, 0101) Bits Stored                       US: 8
(0028, 0102) High Bit                          US: 7
(0028, 0103) Pixel Representation              US: 0
(0028, 2110) Lossy Image Compression           CS: '01'
(0028, 2114) Lossy Image Compression Method    CS: 'ISO_10918_1'
(7fe0, 0010) Pixel Data                        OB: Array of 118558 elements
```

Figure 19: DICOM metadata view presents all the exam-related tags of each patient. DICOM, Digital Imaging and Communications in Medicine.

The dataset XP1 is divided into two groups: training with 26684 images and test with 3000 images. Beyond image data, the dataset is supplemented by two .csv files with additional information for each patient id: the train image labels file and the detailed information file. The first file classifies the images in a binary fashion where 0 corresponds to normal class and 1 corresponds to the pneumonia class. The second file contains supplementary information relative to the image labels, being the images subclassified into 3 different classes: ”normal”, ”pneumonia”, and ”not normal/not pneumonia”. Basically, the images labeled with 0 are originally subdivided into ”normal” and ”not normal/not pneumonia”, while the images labeled with 1 remain in the ”pneumonia” group. This classification was performed by experts based on the clinical diagnostic of each patient and on the inspection of X-rays for visual signals specifically termed lung opacities. Lung opacities refer to areas that attenuate the x-ray beam and therefore appear gray and opaquer than the surrounding area in the acquired image. Pneumonia is a lung infection that can be caused by bacterial, viral or fungal infection being a common cause of lung opacity. The bodys immune response to these infections leads to localized fluid accumulation in the lugs. These fluids spread and accumulate within the lung airways in regions normally filled with air.

These fluid sacks alter the penetrative ability of x-rays leading to the appearance of opacities on the CXRs [148].

For more perception and observation of the non-imaging data, one global data frame was created from the information of the first file, second file and DICOM relevant metadata of each patient id, being synthesized on (**Figure 20a**). It should be noted that the patient id links the image to the information of that specific patient corresponding to a unique data frame line (**Figure 20b**).



(a)

patientId	PatientAge	PatientSex	BodyPartExamined	ViewPosition	Label	AgeGroup	Condition
001031d9-f904-4a23-b3e5-2c088acd19c6	57	M	CHEST	PA	1	Adult (18-64)	Pneumonia
0010f549-b242-4e94-87a8-57d79de215fc	56	M	CHEST	PA	0	Adult (18-64)	Normal
0022995a-45eb-4cfa-9a59-cd15f5196c64	54	M	CHEST	PA	0	Adult (18-64)	Normal
0025d2de-bd78-4d36-9f72-e15a5e22ca82	28	F	CHEST	PA	0	Adult (18-64)	Normal
00293de0-a530-41dc-9621-0b3def01d06d	63	F	CHEST	PA	0	Adult (18-64)	Normal
...	...	...	...	...	...	...	...

(b)

Figure 20: Global dataframe created. (a) The information from the three different files was obtained in order to create a global dataframe. For each patient id, first the dicom metadata was accessed, then the labels file and lastly the detailed information file; (b) view of the global dataframe created with information relative to patientId, age, sex, body part examined, view position, label, age group and condition.

In order to better understand the predictability of the generated models on independent data, we decided to include in this thesis an unrelated dataset of X-ray images (referred in the thesis as dataset XP2). This additional dataset was proposed by Kermany et al [149] and, like the dataset XP1, has as its purpose the classification of X-ray images into ”normal” or ”pneumonia” category.

The dataset images are in JPEG format and were limited to pediatric pneumonia patients between 1 and 5 years of age collected at the Guangzhou Women and Childrens Medical Center, China. All the images were classified as Normal or Pneumonia. In total, the dataset contains 5871 images (1590 ”normal” and 4281 ”pneumonia”), being divided into train (5231), validation (16) and test (624) folders. Similarly to the dataset XP1, the dataset XP2 is also unbalanced, but in this last case, the ”pneumonia” class represents the majority of the cases. Unlike the dataset XP1, the images have no supplementary patient and exam informations. Furthermore, the classification into ”normal” or ”pneumonia” class is not available from tags to embed in the image file and additional .csv files but, from the title of the folder in which the images were grouped.

3.2 DATASETS PREPROCESSING

Image classification can be used for predicting normality or disease, based on the CXR images and respective label information. For this, the dataset XP1 undergone filtering during the preprocessing and selection stage (**Figure 21**). As for the dataset XP2, no further selection steps were made, keeping the number of images intact.

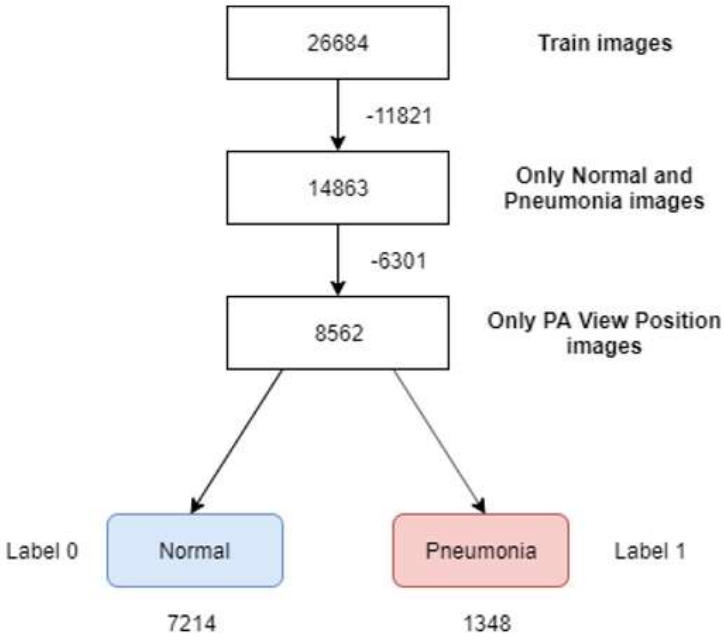


Figure 21: Dataset XP1 image selection.

There are a few key data characteristics that we considered throughout the dataset XP1 selection, with one of them being the division of the labeled 0 images into two subgroups: ”normal” and ”not normal/not pneumonia”. The last one does not refer to any specific medical condition.

So, in order to have a dataset-specific of pneumonia and to prevent possible interferences of this third class in the collection of characteristics of the totally normal images, we decided to remove all images belonging to the ”not normal/not pneumonia” subgroup. Finally, to avoid having more than one image per unique patient we selected only the PA view images. The PA view preference over the AP was due to the fact that the first is the most common in radiology exams [150]. Resuming, the dataset XP1 was filtered from 26884 images to 8562 images, being classified into two classes: ”normal”, labeled as 0, and ”pneumonia”, labeled as 1.

In order to understand the preprocessed dataset XP1 (referred to in the thesis as dataset XP1') we are working with, some data analysis were conducted. For the data analysis, some Matplotlib and Pandas functions were used. While Pandas functions allow selecting specific columns of the data frame, the Matplotlib package permits to do some graphical analysis of the selected data. The **Table 6** shows the number of patients grouped by gender and age group.

Table 6: Dataset XP1' number of patients grouped by gender and group.

	Young	Adult	Elderly	**Total**
Female	174	3221	389	3784
Male	219	3890	669	4778
Total	393	7111	1058	

It's possible to infer that from the chosen 8562 patients the majority is male and adult (**Figure 22a**, **Figure 22b**), which also happens in the original dataset XP1. Relating to the incidence of pneumonia in the samples, it is higher in males (0.167) than in females (0.145), and higher in the elderly population (0.206) than in the other group ages (0.153 for young and 0.15 for adult patients) as would be expected.

The ratio between healthy cases and pneumonia cases is 5.35 on dataset XP1, which indicates that exists approximately 5 healthy images for each pneumonia image (**Figure 22c**). That is a superior ratio compared with the original dataset XP1, (1,5 for 1). Even though it is an unbalanced dataset, it is clearly simple and well processed, with enough image data for many types of experiments.

Regarding misplaced values, namely outliers, there are three patients over 140 years old. This is probably due to an age collection mistake or the patients were born 140 years ago but passed away meantime. As this information does not affect the work to be done, and there is no way to know the exact age of these three patients, they were grouped in the age group of over 64 years old.

Relating to the dataset XP2, no additional information about the imaged patients, such as gender or age, could be obtained and therefore no detailed statistical analysis could be performed.

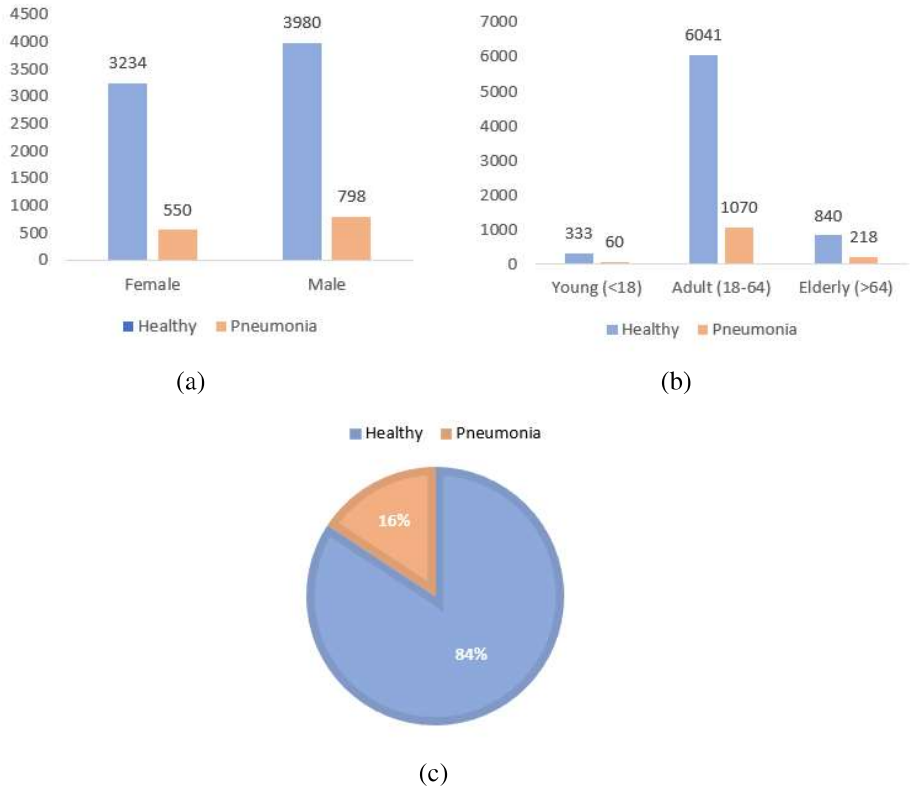


Figure 22: Dataset XP1' statistical analysis. (a) number of patients by gender and clinical condition; (b) the number of patients by age group and clinical condition; (c) proportion of healthy and pneumonia patients.

It can only be mentioned that the ratio between normal cases and pneumonia cases is 0.37, which indicates that exists approximately 2,7 pneumonia images for each normal image.

3.3 MODEL

With the recent improvement of computational power and availability of big data, DL has become the most used AI approach in medical imaging, since it can learn much more sophisticated patterns than conventional ML techniques. DL techniques are increasingly used to improve clinical practice, and the list of examples is getting longer and longer. Regarding the interest of DL in medical imaging, this is mostly triggered by convolutional neural networks, a class of artificial neural networks. They are a powerful approach for learning useful image representations and exploiting local connectivity patterns efficiently. Inside of medical imaging, there has been a surge of big interest in the potential of CNN in radiology, being that several articles have already been published in diversified areas such as image classification, image segmentation, image reconstruction, and lesion detection.

Focusing on the classification task, since it is the DL task covered in this work, it is based on two important terms: the ”X” variable, more properly the features, such as patterns, colors or forms obtained from the medical images; and ”Y” variable, commonly called label, which refers to the target lesions or clinical conditions present in the medical images. These clinical situations are annotated and classified by expert radiologists into two or more classes, and then saved as labels.

The main aim of this work is the classification of CXR into two clinical conditions: healthy or pneumonia. To achieve that we chose to develop and use a CNN model since its the most used DL technique in this field of study. The construction of the CNN model was based on the traditional architecture, where a set of convolutional layers is interleaved with pooling layers, ending with a few fully-connected layers. For CNN development Keras modules such as layers, activation functions, loss functions, optimizers, and models were used.

The further usage of the model will be divided into two different stages: hyperparameter optimization, or hyperparameter tuning, and testing. The first stage is sequentially subdivided into two important steps:

- Training, where all the training data will be loaded into the model. From the training images the CNN will extract and learn differential features associated with the respective label;

- Validation, despite being a step-in tune with training, is used for monitoring the model learning. This step is important for selecting hyperparameters and choose the best combination that obtains the best model performance.

The testing is the final stage, where new data, not seen by the model previously, is used for making predictions. The final performance of the model can be then measured by the application of appropriate metrics on the predicted results.

3.4 HYPERPARAMETERS OPTIMIZATION

In order to set the model for usage, it’s important to look at some important hyperparameters that influence the model behavior. Hyperparameters like the number of epochs and batches, batch size, model optimizer, learning rate, learning rate decay and model loss function, are important pieces of a CNN model. To get the best performance out of a CNN model, it is important to achieve the best combination of hyperparameters that, together, provides better performance. This process is not very simple, since the greater the number of hyperparameters, the greater the number of possible combinations for the model. Sometimes the time required, and the lack of

computational refusals make this process unfeasible. So, once the number of hyperparameters present in CNN algorithms is too high, we opted for using a random search optimization algorithm. For this, we used the Talos package and its scan function with the random search tool enabled.

3.5 PERFORMANCE EVALUATION

According to the medical context of the problem, the main purpose is to correctly identify as many pneumonia cases as possible, being, at the same time, more important to avoid misclassifying a pneumonia case as healthy than misclassifying a healthy case as pneumonia. In other words, the biggest aim is to reach a high sensitivity for positive class. In unbalanced datasets, the accuracy metric is often uninformative once it tends to favor the majority class, which in the dataset XP1' is the healthy class. However, we opted for use accuracy as well as f1 score, recall and precision to evaluate the performance of the CNN model. Although, more importance was given to the last 3 metrics, once these are more appropriate for this kind of problem. Finally, confusion matrices were created for analyzing the number of TP, TN, FP, and FN. The Numpy and Sklearn packages were used during the performance evaluation.

4

DEVELOPMENT

In this section of the dissertation, the computational and technical steps of image preprocessing, CNN model development, hyperparameter tuning, and CNN final testing will be explained in more detail.

All the 8562 images present on the dataset XP1' were further submitted to a preprocessing and transformation steps, where the pixel array and respective labels were collected and saved. This step consists of preparing the data for posterior usage on the CNN model. When all the data is prepared, the CNN model started to be built with defined architecture. In order to see the response of the model according to different loss functions, two different models were considered: one with a classical loss function and another with a loss function that can counteract unbalanced datasets. The two models were both subject to a training process with the training data of dataset XP1'. This process was conducted together with hyperparameters tuning through defined hyperparameter values and a limited number of combinations. Next, the best model 1 and 2 were selected. Finally, the best model 1 and 2 were tested in unseen data.

4.1 IMAGE PREPROCESSING

Before defining the architecture and building the CNN model, the data should be prepared, transformed and stored in the ideal format for later use. For that, some data adjustments were made on dataset XP1 and dataset XP2, both at the image content level, more specifically at the pixel array, and at the level of the respective labels.

The construction of DL models based on imaging data requires, initially, a set of data adjustments. All the image preprocessing steps done in this work for dataset XP1', are schematized in **Figure 23**. Python packages cv2, Numpy, Skimage, tqdm and random were used along this process. For the dataset XP1', we first access each DICOM file, get the image pixel array and convert to JPEG format. The JPEG format allows us to see the image from a first perspective instead of DICOM format. Each image has a resolution of 1024 height by 1024 width, representing a total of 1048576 pixels. The high amount of information present in each image led us to

consider decreasing image resolution. This option is justified by the fact that the larger the image resolution greater the computational resources required to exploit it and the longer the time required for such exploration. It is also known that high image resolution implies more complex learning models, and sometimes the model performance turns out to be worse [134]. Therefore, we decide to resize the resolution of each image from 1024 by 1024 pixels to 200 by 200 pixels.

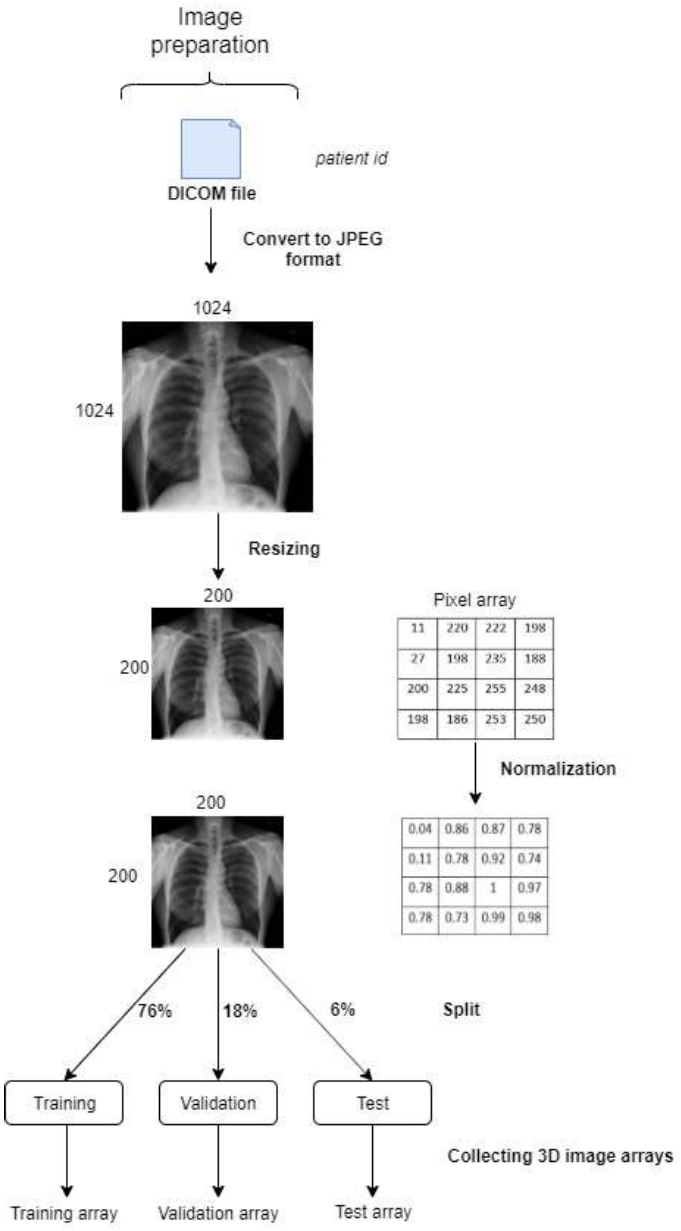


Figure 23: Image preparation pipeline for dataset XP1'. All dataset XP1' images pass through some preprocessing steps: convertion of DICOM file to JPEG format; resizing from 1024x1024 to 200x200 pixel resolution; normalization at pixel array content; split to training, validation and test folders; lastly, the pixel arrays of all the images in each folder are collected and stored in three different files.

Afterward, a normalization procedure was implemented, where all the image pixel values were divided by the maximum pixel value of each image. The pixel values are defined as integers with values between 0 and 255, and the usage of large integer values can disrupt or slow down the learning process on neural networks. So, we chose to normalize all data, obtaining for each pixel a value between 0 and 1, with certainty that the image display remained normal.

Most DL models need different sets of input data in order to effectively train, monitor and measure the performance. So, the dataset XP1' images were split into three different sets and saved in different folders, according to the following percentages:

- training set (76%): set of 6500 images used for training the model;
- validation set (18%): set of 1562 images used for control and measurement of the model performance during the training. It's important for model hyper optimization before testing new unseen data;
- test set (6%): set of 500 images used for measuring the final model performance.

The dataset splitting was done randomly, being the images divided into three different folders: train, validation, and test. For this, a Python function able to randomly select and move the images to each of the folders according to the defined proportions was developed. The proportion of healthy: pneumonia cases remained similar in each set.

Next, for each folder, the pixel array of each image was collected with the aid of the Numpy package, resulting in three different 3D arrays: training array of dimension 6500,200,200, validation array of dimension 1562,200,200 and test array of dimension 500,200,200. These arrays were posteriorly saved in three different files.

Furthermore, label information was collected from the global data frame. After the images were divided, the respective labels were stored in different arrays also with aid of the Numpy package. As such, three different label arrays with zeros and ones were created, one for each folder: training, validation, and test. However, being a classification task and in order to facilitate the usage of the data in the CNN model, we opted for one hot encode the label arrays. The one-hot encoding process consists of transforming the labels into a categorical format by converting the class vector to a binary class matrix. This process was conducted by a Keras auxiliary function called "to categorical". So, each label referring to an image with a normal condition was transformed from 0 into [1,0] and each label referring to an image with pneumonia clues, was transformed from 1 into [0,1]. The label arrays were saved in three different files.

During the construction of both image and label arrays, there was special attention to the order of each image and its respective label, avoiding possible mismatches and wrong information.

Since the dataset XP2 was used to measure the final model performance, only the test set, composed of 624 images, was considered in this work. For the test set, the array of each test

image was also resized to 200 x 200, normalized and posteriorly collected and saved in a global array. For label information, the label of each test image was collected, one hot encoded and saved in a global array too. The two arrays were shuffled in tune, in order to undo the folder order. Finally, two different files have resulted: the dataset XP2 test set array of size 624,200,200, that will be used for new predictions, and the corresponding labels array of size 624,2, that will be used for confirming and evaluate the final predictions.

After all these procedures, the image data is ready to be further loaded on the CNN model in the format of arrays of pixel arrays, where a pixel array corresponds to a single image. As for the label data, it will be loaded as a list of arrays, where an array corresponds to a single label matching the respective image. The 1's position in each array indicates the presence of normality if in the first column, or disease if in second column.

4.2 MODEL ARCHITECTURE

CNN's are constructed to learn spatial hierarchies of features in an adaptative and automatic way. A standard CNN design starts with feature extraction and ends with classification. The feature extraction process is performed by alternating the convolution layers with pooling layers while classification is performed with some dense layers followed by a final output dense layer. In the case of the image classification task, this kind of architecture performs better than an entirely fully connected network [103, 151]. The number of layers is determined by the complexity of the problem and the amount of data. So, the greater the number of data and the more information present in them, the deeper CNN must be to keep up with all the data information. The CNN architecture can be applied either in 2D data, like CXR or in 3D data like CT scans and magnetic resonance imaging [152, 153].

Regarding the proposed CNN architecture, it is composed of five blocks of two 2D convolutional layers, being each block separated by one pooling layer. Then a flatten layer is applied to create a unidimensional vector that can be used on the following layers. Finally, three fully connected layers are added, completing the CNN model (**Figure 24**). The input data is provided from the previously saved arrays, being the input shape of 200,200,3 for each image. Once the images are in RGB format, the number 3 is indicative of the total channels in the image: red, green and blue. The model was developed through the Keras library and its CNN construction modules. The model's architecture, layers, filters and kernel's size were defined as shown on **Table 7**.

For all the convolutional layers, zero-padding were used, meaning that the output size remains the same as the original input. In order to achieve this, one-pixel padding is applied around the

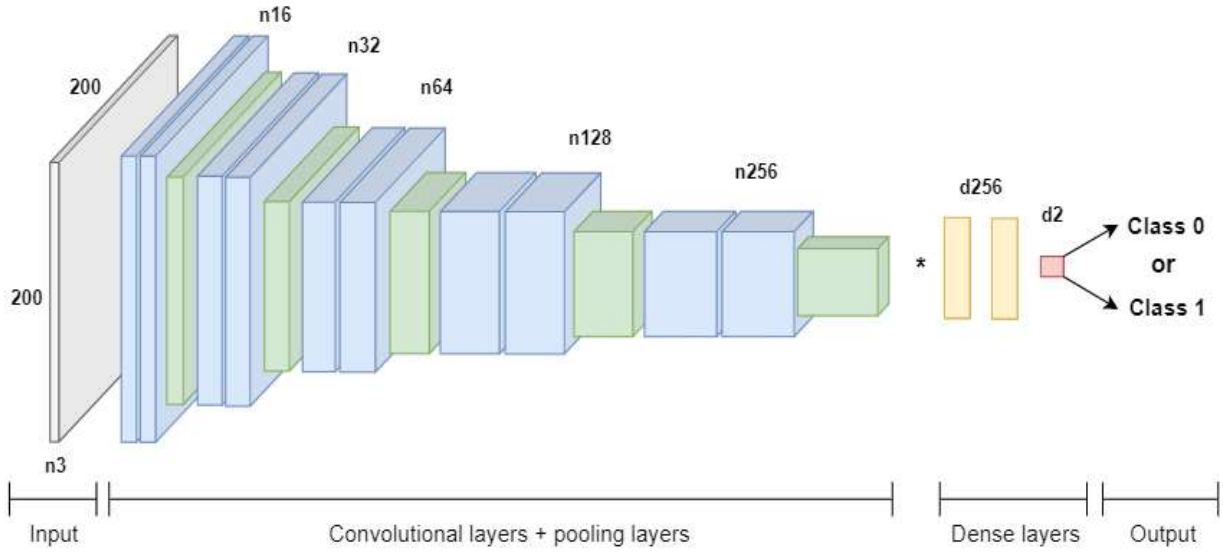


Figure 24: Proposed CNN architecture based on the classical arrangement. The feature extraction process is composed by five blocks of two 2D convolutional layers, each one interleaved with a pooling layer. Finally, the classification process is composed by three dense layers, being the last one responsible for the final output. The referred output results on one of the two different classes: 0, "Normal" or 1, "Pneumonia". Blue boxes refers to convolutional layers, green boxes refers to max pooling layers, yellow boxes refers to dense layers and the red box refers to the output softmax layer. More information about the composition of the layers is provided in table **Table 7**. n, number of filters by layer; d, number of dense units; *, flattening operation.

image array and the filter slides outside the array into this padding area. The stride value used was the default by Keras, i.e. 1, both for convolutional layers and pooling layers.

Regarding the activation functions used in the CNN architecture, ReLU function was used in the five blocks of convolution layers and in the first two dense layers. The softmax function was used in the output layer. This function follows a probability distribution, where the input value is normalized into a vector of two values, corresponding to both classes of the problem. This output vector is composed of one probability value between [0,1] for each class, whose total sums up to 1. The position of the largest value indicates the output class. For example, an output vector of [0.90,0.1] means that the input is more likely to belong to class 0 and is classified as such.

Dropout function was applied between the first and second dense layers and between the second and third dense layers. The use of dropout function allows reaching better performances by reducing model capacity. The first dropout value was set as 0.6 and the second as 0.5.

Table 7: Model layers description.

		Description	Input shape	Output shape
Feature extraction	First block	Number of filters equal to 16; Kernel size of 3x3.	200x200x3	200x200x16
	Max pooling layer	Window size of 2x2.	200x200x16	100x100x16
	Second block	Number of filters equal to 32; Kernel size of 2x2.	100x100x16	100x100x32
	Max pooling layer	Window of size 2x2.	100x100x32	50x50x32
	Third block	Number of filters equal to 64; Kernel size of 2x2.	50x50x32	50x50x64
	Max pooling layer	Window of size 2x2.	50x50x64	25x25x64
	Fourth block	Number of filters equal to 128; Kernel size of 2x2.	25x25x64	25x25x128
	Max pooling layer	Window of size 2x2.	25x25x128	12x12x128
	Fifth block	Number of filters equal to 256; Kernel size of 2x2.	12x12x128	12x12x256
	Max pooling layer	Window of size 2x2.	12x12x256	6x6x256
Classifier	First dense layer	256 units	256	256
	Second dense layer	256 units	256	256
	Third dense layer	2 units: class 0 and class 1.	256	2

4.3 MODEL HYPERPARAMETERS

In order to set the model for usage, it is important to look at some important hyperparameters that influence the model behavior during the training process. Hyperparameters like the number of epochs and batches, batch size, model optimizer, learning rate, learning rate decay and model loss function, are crucial in the model development.

Epoch is defined as the number of times that the learning algorithm will work through the entire training data. That means that one epoch equals one forward pass and one backward pass of all the training samples. This value was firstly defined as 80 in order to better monitor the model learning. The batch size defines the number of training samples per batch. The training data will be learned separately in groups of x random samples. This value was defined as 128. In the same line, the batch defines the number of training samples to work through before updating

the internal model parameters. A training dataset can be divided into one or more batches. This value was defined as 51, the result of dividing all training samples (6500) by the batch size (128). The optimizer is one of the most important hyperparameters needed on CNN, having as function updating the model response to the most accurate form. Choosing the most appropriate optimizer was more difficult, once there are several optimizers that present good results in the literature. However, comparing their performance is complicated as they are applied in different image classification problems. In general, ADAM and SGD show better results in different problems and, as such, were chosen. Within the optimizer, there are other hyperparameters that must be defined, such as the learning rate and the learning decay rate. The learning rate controls how much to change the model in response to the loss function each time the model weights are updated. It is known that learning rate values must be not too high and not too low, in order to achieve the best learning for the model. This value is commonly used between 0.01 and 0.00001. For our work, learning rates of 0.01 and 0.001 were used. The decay rate allows reducing the learning rate during the training process. Low values dont have an effect on learning rate and high values lead to a learning rate blunt drop. It was defined as 1e-04 and 1e-06. Finally, the loss function is responsible for measure the error between the real value and predicted value. The loss function is one of the most important hyperparameters alongside the optimizer. For choosing the loss function, some particularities were considered. The fact that the data is unbalanced could interfere in the equal learning of the two classes. So, we opted for defining two different loss functions: categorical cross-entropy and weighted categorical cross-entropy. The first was chosen because it is the standard loss function used in multi-class classification problems and the one that gets the best results along with using of softmax function on the output layer. The second is a variation of the first and was chosen to try to offset data class imbalance, giving more weight to the minority class, that is, the pneumonia class. With this weight gain, is expected that the model can capture more information from the sick images. The class weights were defined as 0.25 for Normal class and 4 for Pneumonia class.

In order to compare how these different loss functions can influence the model behavior and performance, two different models were considered, one for each loss function. The model that has categorical cross-entropy as loss function, will be named as model 1. The model that has weighted categorical cross-entropy as loss function will be named as model 2. These two models were then separately subjected to a process of hyperparameter optimization, according to the selected hyperparameters.

4.4 MODEL HYPER OPTIMIZATION AND MODEL SELECTION

The choice of the hyperparameters options was reduced by a pre-observation of the model's performance, excluding options that add no observable effects. The selected hyperparameters are shown in **Table 8**. Since the possible combinations of hyperparameters remain too high (64), we cannot explore all the experiments and their results. So, in order to get around this problem, we opted for a simpler and more appropriate optimization method, the random search provided by the Talos package. This technique consists of picking some random combinations from the total number of possible combinations. The number of random combinations is defined by a percentage value. In our case, the percentage value was set as 10%, which gives a total of 6 hyperparameters combinations for each model. For more simplicity, the term hyperparameters combinations are now denominated as experiments.

Table 8: Selected hyperparameters for model optimization.

Batch size	64, 128
Optimizer	Adam, SGD
Learning rate	0.01, 0.001
Decay rate	1E-04, 1E-06
Dense layer 1 units	256, 512
Dense layer 2 units	128, 256

The optimization process was based on two steps: training and validation. For each model, the 6 experiments were trained with the same input data, i.e, the 6500 samples present in the training array. For that, the Keras "compile" and "fit" functions were used. While the "compile" function allows configuring the model for training, the "fit" function trains the model for a defined number of epochs, at the same time as the model parameters are saved as an internal object.

As each train went through, the model's learning capacity was monitored. For that, the training accuracy and loss were measured for each epoch. As for tracking the quality of learning, the validation accuracy and validation loss were measured for each epoch. These measures are very important as they are the most direct quality learning markers of CNN models. Finally, the experiments were tested on validation data in order to draw conclusions about which one gets the best performance. For that, a confusion matrix was created and accuracy, precision, f1 score, and recall values were calculated. From the optimization results, the best model 1 and 2 were chosen to test on new data.

5

RESULTS AND DISCUSSION

To find out the response of the model according to different loss functions, two different models were considered: one with a standard cross-entropy loss function and another with a loss function that can counteract unbalanced datasets. The effect of hyperparameter tuning on both model performance was assessed by a random search method and the best model 1 and 2 were selected from a predefined number of combinations. After choosing the best models, a testing comparison of model 1 and model 2 performance was done on two different test sets: the test set of dataset XP1' and the test set of dataset XP2

5.1 CHALLENGES AND IMPORTANT CONSIDERATIONS IN DATA COLLECTION

There are not many medical imaging datasets available, being, for now, the access to these datasets the biggest challenge on DL tasks. Furthermore, except for a few datasets, the accessible datasets only contain a low number of patients and/or samples. Comparing to the datasets used on general computer vision tasks where the number of data typically ranges from hundreds of thousands to millions, the size of medical imaging datasets is too small. However, and given the scarcity or even lack of accessibility of this data type, the free usage of the RSNA dataset is an important step for many studies of this kind. The number of data provided by this dataset is a good start point for improving the application of ML and DL techniques in medical imaging.

On the other hand, the heterogeneity present in this dataset, principally due to high variances on age, weight, body mass, local of the exam and chest ray machine, is always a difficult problem to reverse. When looking at the CXRs present in this dataset is possible to note that the quality of each image is good but looking in a general context many images differ from the standard point (**Figure 25**). Some variances on brightness, in chest distance to the x-ray detector and rib cage field of view, may lead to the collection of conflicting information or even wrong information.

The unbalanced data was another challenge for dataset XP1'. The number of negative samples is much higher than the positive samples. However, this problem encompasses almost every medical imaging dataset, once the number of healthy patients is almost always higher than dis-

eased patients for a particular disease or group of diseases. Actually, there is a low number of researches that study the impact of class imbalance on DL. Anand et al [154] show that in shallow neural networks the majority class is dominating the net gradient that is responsible for updating the model weights. Training CNN with unbalanced data can induce models that are biased, giving more importance to the majority class [155].

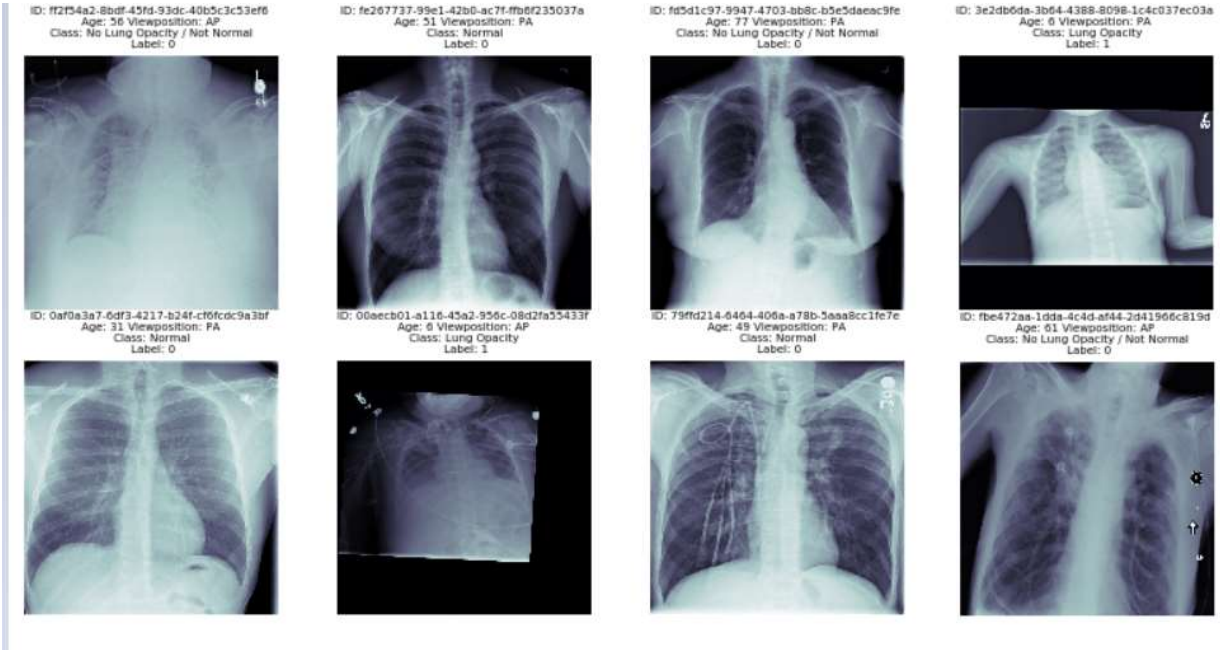


Figure 25: Examples of chest ray images present on dataset XP1'.

Either way, several advised methods have proven that can counteract the problem of unbalanced data in DL tasks. Data sampling methods such as augmentations, over-sampling, under-sampling, SMOTE (Synthetic Minority Over-sampling Technique) [156], which consists of creating artificial samples of the minority class, are widely used. However, in spite of these approaches being simple to apply, they may remove some important data or add redundant data to the training set. Other algorithmic methods such as new loss functions, cost-sensitive learning and threshold moving are an alternative to the data sampling methods. Instead of changing the training data distribution, these methods adjust the learning or decision process in a way that increases the importance of the positive class [155].

To address the unbalanced data issue and how it could affect the CNN operation, we decided to see the effect of some of these methods. Firstly, we apply augmentation techniques, such as rotations, translations and zooming, only on the minority class, obtaining an equal number of negative and positive samples. In another experiment, we use the SMOTE technique for gener-

ating artificial samples for the minority class. Unfortunately, in both experiments, no positive effects were seen in model behavior and performance.

As the data sampling methods shown no effect, we decided to apply an algorithmic change. Wang et al [157] and Lin et al [158] presented new loss functions that give more importance to the minority samples, contributing more to the model's loss. So, as stated before, we opted for developing a second model, only changing the loss function to a weighted cross-entropy loss function in relation to model 1. This was done to observe what effect the class weight changing has on the final performance of the model. We believe that this change allows increasing the sensitivity for the minority class, but as will be explained in the new data testing section, the majority class is widely affected.

Regarding the dataset XP2, this did not present any challenge since it was only used for the final test of the model's performance. Since the samples of this dataset were taken from children between 1 and 5 years of age, it would be expected a greater homogeneity of features at chest level, which can be proven by the visualization of the chest ray images.

5.2 MODELS HYPER OPTIMIZATION

The number of each class samples present in the training and validation set of dataset XP1' is shown in **Table 9**. As previously stated, this is an unbalanced classification task, which can be observed in the image number difference between class 0 and class 1, both on training and validation set of dataset XP1'.

Table 9: Number of each class samples on training and validation sets of dataset XP1'

	Training set	Validation set
Normal	5480	1324
Pneumonia	1020	238

Even though the input data is the same, the model experiments do not behave identical, given the different combinations of hyperparameters that define their composition. The joint action of hyperparameters can lead to different loss and accuracy curves. The assumption that small training and validation loss values, and high accuracy values increase the model's performance, was used in the evaluation of each experiment. The joint analysis of each of the values allows us to infer the quality of the model. High training accuracy allied with low training loss, means that the learning process is going on track. In its turn, high validation accuracy and low validation loss mean that the model is predicting well on new data.

As explained before, we opted for using a random search method presented by Talos package. This was done due to three principally reasons: high chances of finding a better configuration in

fewer evaluations, contrary to the traditionally exhaustive grid search methods; evaluating more different values for each of hyperparameters; each evaluation can be stopped at any time and the trials form a complete experiment [127].

The optimization results for model 1 are shown in **Table 10**. Looking at each experiment it's possible to take some individual observations.

Table 10: Hyper optimization results for model 1.

Exp	Val loss	Val acc	Train loss	Train acc	Batch size	Lr	Decay rate	Dense 1	Dense 2	Optimizer
0	0.427	0.848	0.435	0.843	64	0.01	1e-06	256	128	ADAM
1	0.334	0.909	0.071	0.975	128	0.01	1e-04	512	256	SGD
2	0.427	0.848	0.435	0.843	64	0.01	1e-06	256	256	ADAM
3	0.256	0.905	0.251	0.901	64	0.001	1e-06	512	256	SGD
4	0.319	0.917	0.081	0.971	128	0.01	1e-06	256	128	SGD
5	0.655	0.919	0.017	0.994	64	0.001	1e-04	512	256	ADAM

Exp, Experiment; Val, Validation; Train, Training; Acc, Accuracy; Lr, learning rate.

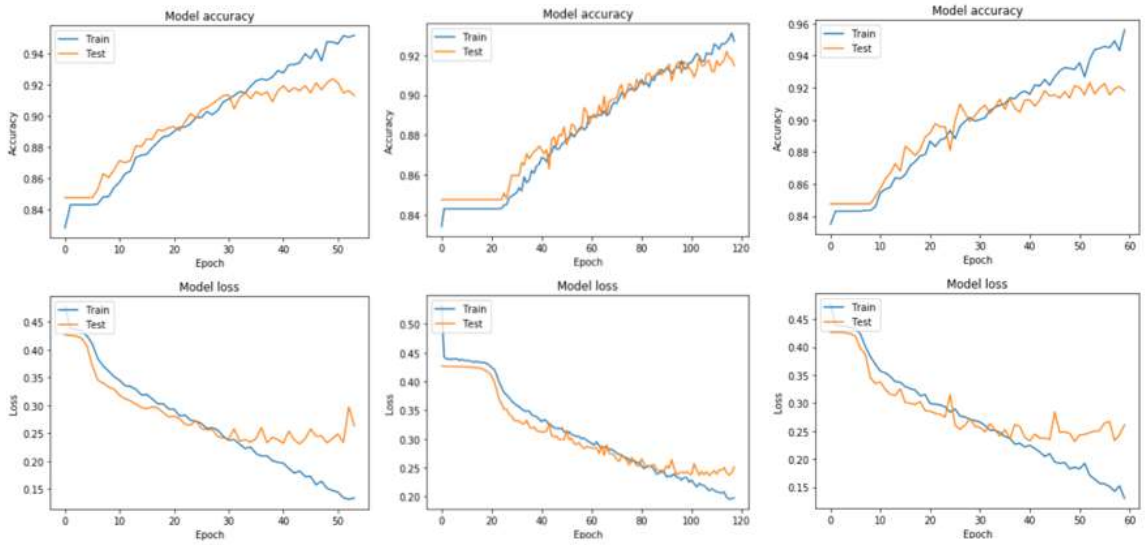
There are no verified signs of progress on the learning process at experiments 0 and 2. The training accuracy doesn't keep a high landing, and the training loss remains high. The lack of learning is visible in the validation results, where the value of validation accuracy is 84,8%. The percentage of healthy cases on the validation set is 84,8%, which means that this model is predicting all the validation data as belonging to "normal" class. It is notorious that the accuracy and loss curves for both training and validation do not change during training. At experiments 1,3 and 4, from the low training loss and high training accuracy, it's possible to conclude that the learning process went well. In its turn, the validation accuracy reached a high value in both cases, while the validation loss remains at a median-low level. Finally, in experiment 5, the low training loss and high training accuracy mean that the learning process went well. On the other hand, when analyzing the validation results, it's possible to see a strange case, where high values of validation loss and validation accuracy are obtained. This might be a case of overfitting.

Analyzing all the optimization experiments, it is possible to conclude that the biggest performance is obtained with the hyperparameters combination present in experiments 1, 3 and 4. For better analyzes of these three experiments, a re-training process was made with a final validation data prediction. The re-training process allows us to monitor the loss/accuracy curve throughout the training, procedure that Talos scans did not allow. In its turn, the validation data prediction lets concretely know how many cases of each of the classes each experiment can correctly predict, giving the value of accuracy, recall, precision and f1 score.

During each experiment re-training, the best model weights were saved, recurring to two important Keras functions: "early stopping" and "checkpoint". The "early stopping" allows

stopping the training process when it was apparent that the learning process had stalled, more properly, when the validation loss starts to increase. The ”checkpoint” allows saving the model weights for the highest validation accuracy obtained. So, these combined techniques allow to stop training and save the model weights that give the best validation accuracy until the point that the loss curve starts to move away from the minimum and starts to rise. This procedure helps to prevent overfitting.

From the analysis of the different loss-accuracy curves obtained (**Figure 26**) it is possible to observe that at epoch 45-50 on experiment 1, the validation loss starts to increase (**Figure 26a**). So, the training was stopped at this point, and the model weights for experiment 1 were saved. For the experiment 3, the validation loss starts to increase only after 100-110 epochs (**Figure 26b**). This late increase is due to a lower learning rate. The training was stopped at this point, and the model weights for experiment 3 were saved. The loss-accuracy curves presented by experiment 4 (**Figure 26c**) look similar to experiment 1, being the training stopped at epoch 40-45. The model weights for experiment 4 were saved.



(a) (b) (c)

Figure 26: Accuracy-loss curves for model 1 experiments 1, 3 and 4. For the 3 experiments the model weights for the highest validation accuracy obtained until the point that the validation loss starts to increase, were saved. (a) In experiment 1, both training and validation accuracy increase along the epochs. As well, both loss curves are decreasing. At epoch 45-50 the validation loss starts to increase. The best weights were saved before this point, in order to prevent the overfitting phenomenon; (b) at experiment 3, both accuracy curves increases but, given to a lower learning rate, the number of needed epochs was higher. At epoch 100-110 the validation loss starts to increase; (c) the behavior of accuracy-loss curves for experiment 4 were similar to those presented by experiment 1. The validation loss starts to increase at epoch 40-45.

For choosing the best of the three models a prediction analysis was made. The class of all the 1562 validation images was predicted from each experiment and respective loaded weights. The ”normal” class is defined as negative class (0), while the ”pneumonia” class is defined as positive class (1). The prediction results for model 1 experiments are shown on **Table** 11.

Table 11: Confusion matrix of validation data prediction and performance metrics for experiment 1, 3 and 4 (Model 1).

Exp	Real/Predicted	Normal	Pneumonia	Precision	Recall	F1 score	Acc
1	**Normal**	1292	32	0.94	0.98	0.96	0.92
	Pneumonia	87	151	0.83	0.63	0.72	
3	**Normal**	1297	27	0.93	0.98	0.95	0.92
	Pneumonia	103	135	0.83	0.57	0.68	
4	**Normal**	1305	19	0.93	0.99	0.96	0.92
	Pneumonia	101	137	0.88	0.58	0.7	

Exp, Experiment; Acc, Accuracy.

In medical cases, it’s important to verify that the hit rate is always high. However, in unbalanced datasets, the accuracy might not be a good metric to evaluate the model’s performance. The number of FN and FP must be treated differently, once is preferable to have healthy patients that are diagnosed as diseased as opposed to having diseased patients that are diagnosed as healthy. So, the important point at the evaluation of the three experiments is the lowest number of FN, which is expressed as the highest recall for pneumonia class. Having the highest recall means that the majority of the diseased patients are correctly classified, resulting in a low number of FN. However, there is special attention to the other metrics, since their values could not be good for the medical problem. Even though the recall obtained is high, the hypothetical low accuracy and low precision values make the model inefficient, leading to an incorrect prediction of healthy cases. In conclusion, it’s important that the model hit as many cases as possible, but, in a medical context, the higher the number of illness cases predicted correctly the better their performance.

For class 0, all the experiments predicted well, with values slightly above 90% for accuracy and precision and around 100% for recall. As for class 1, the results were more varied, with precision values between 83% and 88% and recall values between 57% and 63%. Analyzing the results and considering that in medical applications of this type recall is more valued than accuracy and precision, we conclude that experiment 1 outperformed experiments 3 and 4.

Concluding from model 1 hyperparameter optimization, the SGD optimizer proved to be better than Adam optimizer on these experiments. Inside this, batch size of 128, learning rate of 0.01 and decay rate of 1e-04 outperformed the other values. As for dense 1 and dense 2, as bigger the number of units, the better the performance of the model. So, looking at that, the

hyperparameters combination present on experiment 1 was the one that correctly predicted the highest number of cases of pneumonia and, as such, experiment 1 was chosen as the best model 1.

Similarly to the model 1, the optimization results for model 2 are presented in **Table 12**. Analyzing each experiment, it's possible to reach a few conclusions.

Table 12: Hyper optimization results for model 2.

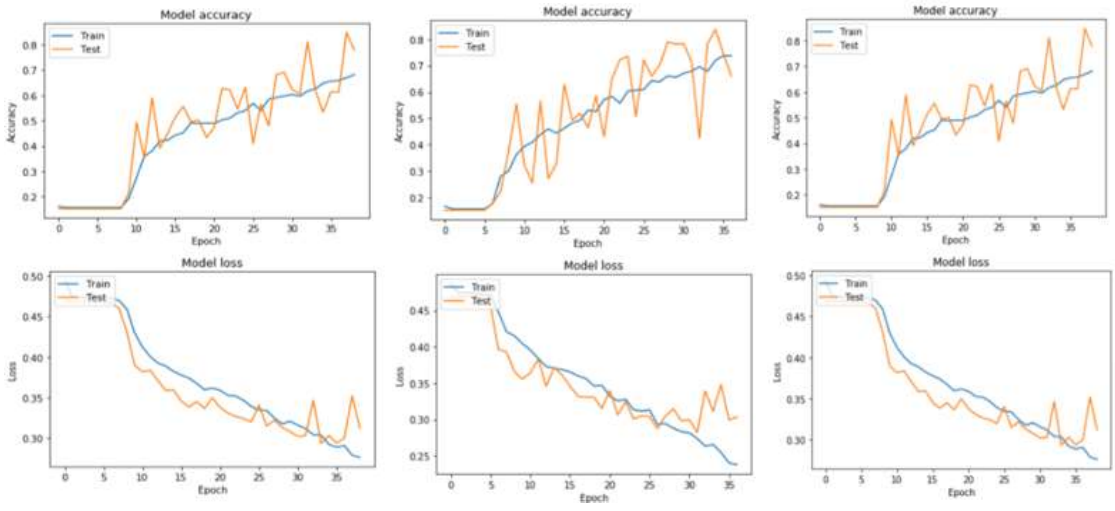
Exp	Val loss	Val acc	Train loss	Train acc	Batch size	Lr	Decay rate	Dense 1	Dense 2	Optimizer
0	0.988	0.881	0.053	0.957	64	0.01	1e-04	256	128	SGD
1	0.948	0.892	0.038	0.967	128	0.01	1e-06	512	256	SGD
2	1.511	0.889	0.028	0.979	128	0.001	1e-04	512	256	ADAM
3	0.888	0.894	0.063	0.945	128	0.01	1e-04	512	256	SGD
4	3.416	0.152	3.398	0.157	64	0.01	1e-06	512	128	ADAM
5	1.338	0.872	0.044	0.963	64	0.001	1e-04	256	128	ADAM

Exp, Experiment; Val, Validation; Train, Training; Acc, Accuracy; Lr, learning rate.

At experiments 0, 1 and 3, from the low training loss and high training accuracy, it is possible to conclude that the learning process went well. As for the validation values, the loss was median-high, and the accuracy was high. The training accuracy and training loss values in experiments 2 and 5 indicate that the learning ability reached high levels on both experiments. However, the high validation loss obtained indicates that the prediction process did not perform well. Poor results were observed in experiment 4, namely at the learning level and prediction level. The high validation loss and validation accuracy of 15,2% means that the model is predicting all the validation data as belonging to pneumonia class, once it is the proportion of positive cases on the training data. Observing the loss and accuracy values for training and validation it's possible to conclude that these do not change during the training.

Analyzing all the optimization experiments for model 2, it's possible to conclude that experiments 0, 1 and 3 had a better performance. For a better analysis of these experiments, a re-training process followed by a validation data prediction was made. The best model weights save was done in the same way as for model 1.

From the analysis of the different loss-accuracy curves (**Figure 27**) obtained for experiments 0, 1 and 3 during the optimization of model 2, it's visible that, for all the experiments, the validation loss begins to rise in the range of epoch 35-40, being all the trainings stopped inside this range (**Figure 27a,27b,27c**). Looking for the training and validation curves there are not many differences between the three experiments. All the model weights were separately saved.



(a) (b) (c)

Figure 27: Accuracy-loss curves for model 2 experiment 0, 1 and 3. (a-c): All the training behaved similarly, with both accuracy curves increasing and both loss curves decreasing until epoch 35-40. In the three experiments, some oscillations on validation accuracy were observed. These oscillations are an effect of the weighted loss function applied in this model, which attempts to counter the unbalanced data present on the dataset XP1'.

Next, for choosing the best model 2, the weights were loaded into the respective experiment, and a validation data prediction was made in the same way as done for model 1. The confusion matrices, precision, recall, f1 score and accuracy values are shown in **Table** 13.

Table 13: Confusion matrix of validation data prediction and performance metrics for experiment 0, 1 and 3 (Model 2).

Exp	Real/Predicted	Normal	Pneumonia	Precision	Recall	F1 score	Acc
0	**Normal**	1111	213	0.96	0.84	0.9	0.84
	Pneumonia	43	195	0.48	0.82	0.6	
1	**Normal**	1119	205	0.96	0.85	0.9	0.84
	Pneumonia	41	197	0.49	0.83	0.62	
3	**Normal**	1139	185	0.96	0.86	0.91	0.85
	Pneumonia	52	186	0.5	0.78	0.61	

Exp, Experiment; Acc, Accuracy.

As said before, it's important to pick up a model with high recall value, but with balanced precision and accuracy values. Observing the prediction results, it's possible to conclude that all the three predictions are very similar for both class 0 and class 1. It's possible to see that the experiments prediction were good for class 0, with high values of precision (96-97%) and recall (82-86%). As for class 1, the recall obtained values were good (78-84%), but the precision only approached the halfway percentage in all experiments. The choice of best model 2 relapsed on

experiment 1 since it exhibits the highest recall and the middle value of precision for pneumonia class when compared with the other two experiments.

From model 2 hyper optimization, it's possible to conclude that, as for model 1, the SGD optimizer revealed higher performance than Adam optimizer. Learning rate of 0.01, decay rate of 1e-06 and batch size of 128 outperformed the other options. Similarly to model 1, units of 512 and 256, for first and second dense layers respectively, proved to be better options for improving the model's performance in this specific problem.

Both grid search and random search methods are widely used in the literature. However, there are scarce studies on the application of hyper optimization methods in medical imaging. As for similar studies, i.e. using DL techniques on chest ray images in order to predict particular lung diseases, no hyper optimization methods were deepened as far as we are aware of. Thus, we can only rely on studies carried out with CNN on other types of images. Most researchers have been based on random search experiments from the Bergrtra and Bengio work [127], which presented good results. This study showed that random search methods are more efficient than exhaustive search methods, however, the final result could be not the optimum. Compared with grid experiments of Larochelle [159], the random search method found better models with less computational time in most cases. Grid search techniques encompass too many trials to the exploration of dimensions that do not make a difference and suffer from poor coverage in dimensions that are important. However, despite the advantages of random search over grid search, the random approach is limited because presents poor adaptability and inability to exploit the performance scores of each combination to direct the search [127].

From a global perspective, the obtained results for the hyper optimization procedure using random search were good, with both optimized models being able to predict at a high-level on validation data.

5.3 MODELS PERFORMANCE ON NEW DATA: TEST SET XP1'

To obtain the model's final performance, new unseen data was inferred by the two selected models: experiment 1 for model 1 and experiment 1 for model 2. In a first approach, a prediction was made in the XP1' test set and, in a second phase, another prediction was made in the XP2 test set. The prediction process for both test sets was conducted in the same way as the prediction of validation set during the choice of the best models in the previous section. So, as stated before, the "normal" class is defined as negative class (0), while the "pneumonia" class is defined as positive class (1).

The model weights of both models were loaded, and the Keras function "predict ()" were used on all the 500 test images of dataset XP1'. For further evaluation of the model's ability to

predict correctly on the test set, the previously saved test labels were used. So, for visualizing the results, confusion matrices were created and the values of precision, recall, f1 and accuracy were measured for each model **Table** 14.

Table 14: Confusion matrix of dataset XP1' test set prediction and performance metrics for model 1 and 2.

Model	Real/Predicted	Normal	Pneumonia	Precision	Recall	F1 score	Acc
1	**Normal**	395	15	0.91	0.96	0.93	0.89
	Pneumonia	41	49	0.77	0.54	0.64	
2	**Normal**	341	69	0.94	0.83	0.88	0.82
	Pneumonia	21	69	0.5	0.77	0.61	

Real pneumonia cases: 90; Real normal cases: 410; Acc, Accuracy.

Analyzing the model 1 results, it's possible to conclude that the model went well on "normal" class prediction. From a total of 410 Normal cases, the model predicted correctly 395, failing only 15 cases. The 91% value for precision, 96% for recall and 93% for f1 confirm the good results. Regarding class 1, which has greater relevance, the prediction results were not so good, reaching a medium level. From a total of 90 pneumonia cases, the model predicted correctly 49 cases, misclassifying 41 cases. As for class 1 precision, the obtained value was 77%, which means that from the total of predicted pneumonia cases (64) the model hits 77% of these cases (49). For class 1 recall, the value was 54%, which means that from the total of real pneumonia cases (90) the model predicted correctly 54% of them (49). Finally, the class 1 f1 score was 64%. As for global accuracy, the model obtained 89%, hitting 444 cases and misclassifying 56 cases.

From model 2 results, it's possible to say that the model is predicting well in class 0. From the 410 normal cases, the model predicted correctly 341 cases. For the precision, recall and f1 score, the obtained values were 94%, 83%, and 88%, respectively. For the total cases of class 1, the model predicted correctly 69 cases. For class 1 precision, the obtained value was 50%, which means that from the total of predicted pneumonia cases (138), the model hits 50% of the cases (69). For class 1 recall, the value of 77% indicates that from the total of real pneumonia cases (90), the model predicted correctly 77% of these cases (69). The class 1 f1 score was 61%. Finally, the model's 2 global accuracy was 82%, hitting 410 cases and misclassifying 90.

Comparing model 2 and model 1 results, is possible to observe some differences when testing on the test set XP1', principally at class 1 prediction level. Is possible to see some variance on precision and recall values, a kind of position exchange between the two models. The change in class weights in model 2 led to greater learning of class 1 image features, countering their numerical inferiority compared to class 0. At the prediction process, this greater importance given to class 1 images led to a higher number of class 1 predicted images (138) relating to model 1 (64), which is almost double. As the number of pneumonia cases predicted by the

model increases, the number of pneumonia hits also yields to increase. So, this process allowed a decrease in the FN cases, leading to an increase in recall percentage. On the other hand, as the number of predicted pneumonia cases increases, the number of FP also increases, leading to a precision reduction.

It is often difficult to compare DL models in literature due to differences in the task, dataset, images, and architecture. The performance of several CNN on diverse abnormalities based on publicly available OpenI dataset shown that the same architecture does not perform well across all abnormalities [160]. Performances for model 1 and model 2 are hard to discuss since the chosen dataset has been scarcely used in research and studies beyond the RSNA 2019 challenge. The fact that the purpose of this work is using the dataset into a classification task, contrary to the original pneumonia detection task, leads to a more difficult term of comparison. From the RSNA Pneumonia Challenge 2019, many submissions were done, but the highest score was low. The winning solution was based on CoupleNet, a fully convolutional network that couples the global structure of images with local parts for object detection [161]. It is known that ensemble models significantly improve the classification performance when compared with a single model [162]. Even though the winning solution consisted of a deeper model based on ResNet [163] pre-trained on the ImageNet dataset [164], it was still centered on a classifier that resembles the one used in this work. The chosen optimizer was the SGD, similarly to the work here presented. However, the authors opted for a lower learning rate of 0.001 and only 14 training epochs. As their used metric was defined as the mean of the intersection over union of matching prediction and ground-truth boxes at several matching thresholds, no comparison can be done with our work, once it is only based on classification metrics such as precision, recall, and f1 score.

5.4 MODELS PERFORMANCE ON NEW DATA: TEST SET XP2

To see both model's performance on another dataset, the weights of both models were loaded, and the Keras function "predict ()" were used on all the 624 dataset XP2 test images. The evaluation of the models was done in the same way, being the saved test labels used in this process. The prediction results, confusion matrix and performance for both models are shown on **Table** 15.

Table 15: Confusion matrix of dataset XP2 test set prediction and performance metrics for model 1 and 2.

Model	Real/Predicted	Normal	Pneumonia	Precision	Recall	F1 score	Acc
1	**Normal**	168	66	0.87	0.72	0.79	0.85
	Pneumonia	26	364	0.85	0.93	0.89	
2	**Normal**	60	174	0.97	0.26	0.41	0.72
	Pneumonia	2	388	0.69	0.99	0.82	

Real pneumonia cases: 390; Real normal cases: 234; Acc, Accuracy.

Observing the results, it can quickly be concluded that the results were very good for model 1. Starting from class 0, the precision of 87%, recall of 72% and f1 of 79% reveals that the model was able to correctly predict a high number of "normal" images. From a total of 234 "normal" cases, the model predicted correctly 168, failing 66. Regarding class 1, the prediction results were very appreciable, reaching a high level. From a total of 390 cases, the model predicted correctly an impressive number of 364 cases, misclassifying only 26 cases. The class 1 precision of 85%, indicates that from the total of predicted pneumonia cases (430) the model hits correctly 364. The class 1 recall of 93% indicates that from the total of real pneumonia cases (390) the model correctly predicted 364 cases. Additionally, the class 1 f1 score was also high, 89%. Globally, the model managed to hit 532 out of 624 cases, obtaining an accuracy of 85%.

In a brief analysis, model 1 has better performance on predicting in the dataset XP2 than in the dataset XP1' There is no plausible explanation, nor is there reference to such a similar event in the literature. We can only infer that the high homogeneity present in the dataset XP2 comparing with the dataset XP1', could be the most likely justification. The high recall percentage obtained for class 1 indicates that, in the new dataset, model 1 can correctly predict most of the "pneumonia" cases.

For model 2, it's possible to see the effect of the higher weight attributed to the pneumonia class. There is a high discrepancy between the number of predicted cases for class 0 and class 1, which results in only 2 FN and 174 FP. In this dataset, the model is predicting a high number of Pneumonia cases (562) and a low number of Normal cases (62). As such, the recall, precision, and f1 score values are affected.

From class 0 total cases (234), the model predicted correctly 60 cases and incorrectly 174 cases. As for class 0 precision, the results were good, 97%, which indicates that from the total of predicted Normal cases (62), the model hits correctly 60. On the other hand, the class 0 recall is affected, being obtained a low value of 26%. This indicates that from the total number of Normal cases (234), the model just predicted correctly 60 cases. As expected, the f1 score obtained was also low, 41%. Regarding class 1 predictions, the results were good, since most cases were predicted to belong to this class. From a total of 390 pneumonia cases, the model predicted correctly 388 cases, misclassifying only 2 cases. As such, the class 1 recall was 99%. The class 1 precision was 69%, which indicates that from the total of predicted "pneumonia" cases (562) the model hits correctly 388 cases, misclassifying 174 cases. The f1 score for class 1 was 82%. Globally, the model correctly predicted 448 out of 624 cases, obtaining an accuracy of 72%.

In conclusion, the performance of model 2 for class 1 was better in the dataset XP2 test data than in the dataset XP1' test data. In contrast, the results for class 0 were better in the dataset XP1'. The better performance of model 2 for class 1 in the new test data is explained by the

high number of ”pneumonia” class predictions. It is known that the more cases of pneumonia are predicted, the greater the number of cases of pneumonia hit. If the model predicted all the cases as belonging to class 1, the recall value would be 100%. The fact that only 2 patients were incorrectly classified as healthy while ill is impressive. However, these results are misleading, once in the medical context does not matter if the model can correctly predict the most diseased patient while failing for most normal patients (174 out of 234). There needs to be a balance and of course, the higher the number of correct predictions for both classes, the better the performance of the model.

There are not many studies involving the application of DL methods on detecting pneumonia with this dataset. Within the studies already published, there are many differences in the level of data preparation, data division, metrics used in the evaluation, and adopted CNN architecture. The models proposed in this work and some other on the literature that uses the dataset XP2 provided by Kermany et al [149] are resumed in **Tables** 16. In addition to building the dataset, Kermany et al [149] also proposed a CNN model with transfer learning for chest ray classification. In the ”normal” versus ”pneumonia” classification task, this work obtained an accuracy of 92,8%, sensitivity of 93,2% and specificity of 90,1%.

Table 16: Comparison of the proposed models and literature models applied on dataset XP2.

	Train data	**Test data**	**Acc**	**Specificity**	**Sensitivity**	**Method**
Kermany [149]	XP2	XP2	92,8%	90,1%	93,2%	CNN + transfer learning
Stephen [165]	XP2	XP2	93,7%	-	-	CNN
Saraiva [166]	XP2	XP2	95%**	84,5%**	99,2%**	CNN
Rajaraman [167]	XP2	XP2	95,7%	91,5%	98,3%	Customized VGG16
Rajaraman [167]	XP2	XP2	94,3%	85,5%	98%	CNN
Proposed 1	**XP1’**	XP2	85%	72%	93,3%	CNN
Proposed 2	**XP1’**	XP2	72%	25,6%	99%	CNN

Train, Training; Acc, Accuracy; ** average values; - information not provided.

In other research, Stephen et al [165] proposed a CNN model trained from scratch with the same purpose as our work, i.e. classification of CXR images into ”normal” or ”pneumonia” class. To achieve such a feat, the authors used all the images present on the dataset XP2, dividing into only two sets: training and validation. The proposed CNN architecture was similar to the presented in this work, with an input size of 200x200 and composed by 4 convolutional layers interleaved with 4 pooling layers and 3 dense layers. Contrary to our final output layer, they decided to apply a sigmoid layer, considering the problem as a binary classification task. Relating to the model performance of the work, they obtained a validation accuracy of 93,73%. No other

metrics, such as recall, and precision were measured, and the model was not tested on other new data. The fact that the validation data used on this work is composed of 2134 images and our test data is composed of 624 images, leads to an incomparable point between our highest test accuracy (85%), obtained for model 1, and their highest validation accuracy (93,73%). The usage of accuracy as a single metric on unbalanced datasets is often uninformative and their comparison between different works is hard to achieve.

Saraiva et al [166] proposed a CNN model of input size of 300x300, composed of 7 convolutional layers, 3 pooling layers, and 3 dense layers. As in our work, a softmax output layer was used. A cross-validation procedure of k fold was used, dividing the dataset XP2 into a train set and test set on 5 repetitions. Regarding the 5 experiments, Saravia et al obtained an average accuracy of 95%, average sensitivity of 98,3% and an average specificity of 83,3%.

The research made by Rajaraman et al [167] consisted on evaluate, visualize and explain the performance of customized CNN's to detect pneumonia from the pediatric chest rays provided by Kermany et al [149] dataset. The first stage of the work consisted on the lungs segmentation to avoid irrelevant features provided by surrounding chest ray regions. The proposed CNN model is based on the input size of 1024x1024, composed of 6 convolutional layers interleaved with 6 pooling layers followed by 1 dense layer. In terms of performance, CNN had an accuracy of 94.3%, specificity of 85.5% and sensitivity of 98%. Rajaraman et al [167] also proposed a customized VGG16 which obtained high results. The model is based on a deepest CNN designed for object recognition. The customized VGG16 presented by the authors outperformed similar state-of-the-art articles in all performance metrics, being the segmentation process and localization of regions of interest important approaches that allow achieving better results. However, this is a more complex DL approach that covers both classification and localization tasks.

As we stated before, the difference between the performance metrics of the referred papers can be a subjective term of comparison. Inside the papers that provide all the performance metrics, only Rajaraman et al [167] and Kermany et al [149] used the original XP2 test set of 624 images.

Even though our proposed model 1 does not outperformed the models proposed in the other researches, it yields good results for an approach not previously used in similar cases. Unlike the scientific research papers aforementioned, where the training data and test data belong to the same dataset, our work consisted on training a CNN from the dataset provided by the RSNA Pneumonia Challenge 2019 and testing on another dataset, the childhood pneumonia dataset provided by Kermany et al [149]. It is important to refer that in the literature no similar process was done, as we aware of. For technological advances in radiology to occur, several kinds of research at a testing level on independent and new data must be conducted. Not many conclusions can be drawn if an algorithm is only tested on the same data distribution that it has been trained on. In-

side the same data type, for example, CXR, a particular algorithm can reach a high-performance level when tested on the same distribution data but reach poor results when tested on new data.

6

CONCLUSIONS AND FUTURE WORK

6.1 GENERAL CONCLUSIONS

Looking at the medical imaging field, there are three groups of x-ray-based techniques very successfully used in diagnosis and detection of lung abnormalities, namely CXR, CT scans, and HRCT scans. At the first line of diagnosis, the CXR are still the standard method in most clinical settings. However, taking full advantage of the information in CXR is challenging and not always successful in the desired timeframe. One of the limitations in this process is its dependency on the analysis by experts. Even with medical training it might be difficult for a physician to identify some lesions or to perform differential diagnostic of diseases and abnormalities during the visual examination of the image [6]. Other x-ray imaging methods, such as CT and HRCT, improve assertive confirmation of exclusion of a certain diagnosis by giving more detailed and high definition images. However, these approaches require a more costly and an even larger technical differentiation for its use and analysis. The largest amount of available data is in the form of CXR, since it is historically the most frequently used method, and its exploitation combined with automated detection and classification holds large promise in the field of medical imaging for being a resource-efficient and value-creating option. This kind of approach is also of interest because it can improve the accuracy of diagnosis even in clinical settings not equipped with state-of-the-art equipment.

Over the last few years, the development of AI techniques combined with the accumulation of large sets of medical imaging data has opened up new technologies useful for medical applications extending from diagnosis to treatment and even disease prediction and prevention. These rising technological advances have been investigated in the most diverse areas of the body, namely the brain, liver, breast, and intestines [168, 169]. Recently, the Google Health company developed and published an AI system that is able to surpass human experts in breast cancer prediction [86]. In an independent study of six radiologists, the AI system outperformed all.

Furthermore, the authors showed that the system was capable of generalizing on two different datasets, one from the United Kingdom and one from the United States of America [86].

Inside AI branches, DL is emerging in several areas, being its application on healthcare one of the most promising for automatic lesion detection and differential diagnoses [8]. The DL can be defined as a modern field of ML based on ANNs assumption, however, DL methods are deepest and with larger ability to model and extract important features from large-scale data, such images, videos, and sounds. DL appears as a high potential and efficient field that can overcome previous limitations with ANN, solving some big data issues, particularly at storing, analyzing and interpretation levels.

The increase of computation tools, such as the advent of powerfull graphics processing units (GPU), allied with the availability of different software, such as Tensorflow and Keras, has much-improved the possibilities and accessibly in the application of DL methods towards the medical field. Some well-rated CNN, such as AlexNet [121], VGGNet [163], ResNet[170] and GoogleNet [171], have been widely used as a basis for many research projects. Recently, all of these CNN and other DL methods have contributed to more promising diagnostic performance, across medical specialties, such as ophthalmology, radiology, and dermatology [172].

Relating to the availability of the medical data and how it can be applied, even though the number of medical images has increased exponentially over the past few years, these are under enormous amount of restrictions that create barriers to the development of new studies and technologies. Regarding the availability of medical datasets, the privacy of the image owners must be guaranteed, to prevent disclosure of personal information or even malicious practice. So, the importance of having bigger medical data for research and development needs to be always in line with the mandate of protecting human rights by imposing limits and restrictions on study design quality. However, large efforts should be made to promote the correct of medical datasets for research, eliminating all barriers regarding to bureaucracy, political will or lack of efficient and interdisciplinary communication.

It is very challenging to differentiate among a large number of lung diseases, being most of them very similar with very identical symptoms and radiological patterns. ILDs are the most common chest diseases, being grouped into major categories that present similar abnormalities and symptoms [2]. Inside ILD, pneumonia groups are the major categories, and their categorization is still hard to achieve, even for experts. Many of the available medical imaging datasets present a mixture of many lung diseases that hindering the performance of AI application for disease classification or differential diagnosis.

During the construction of medical datasets, namely with imaging data, some large drawbacks are difficult to avoid. The quality of the images is the hardest challenge, because the quality of each image depends, beyond the x-ray machine, on the patient's body. The hetero-

geneity of individuals from which the samples are collected, including age, ethnicity, anatomical particularities, add to the noise presented in the final dataset influencing the capacity of the models to extract important and differential features. Furthermore, the ground truth labels for each sample sometimes diverge from subject to subject, being the dataset construction vulnerable to the inter and intra-observer variation. So, the scientific and medical communities must ensure robust training datasets with reliable ground truth labels in order to successfully implement DL models in clinical practice.

One of the biggest problems present in similar literature is the non-extrapolation approach across datasets. Not many conclusions can be drawn if a method is only tested on the same data distribution that it has been trained on. For technological advances to be made, a less reserved approach is essential. It's critical that more in-depth studies are done to allow a relevant role of DL in the future of medicine, especially at the level of diseases that are difficult to diagnose and differentiate, such as ILD.

The main objective of this work was to create a DL pipeline based on CNN, to predict "normal" or "pneumonia" classes from CXR images. This was accomplished by building two different models based on two different loss functions. This work implemented an innovative approach that, contrary to other relevant works, consisted of training a model with a dataset and test it on another dataset. It's possible to conclude that model 1 outperformed model 2 in a global analysis. Even though our proposed model 1 does not outperform the models proposed in similar literature, it yields good results while using a generalization approach not previously used in similar studies.

Model 1 obtained higher performance when tested on new data from dataset XP2, scoring accuracy of 85%, recall of 93% and precision of 85% for positive class, than in the dataset XP1', which scored an accuracy of 89%, recall of 77% and precision of 54%. No conclusive facts can be drawn to explain the difference of performance of the model between datasets, it can be only inferred that the higher homogeneity present on dataset XP2 compared with dataset XP1' could be a plausible justification.

Model 2 behavior was different. Using XP1', the extracted features from the training set allow the model to predict well, being that the change on class weights allows the model to predict and correctly identify more pneumonia cases. At this point, model 2 was considered better than model 1, once it can correctly more pneumonia cases (49 vs. 69). However, when tested on the test set XP2, the poor results for negative class (only 10% of total cases predicted as normal) proven that model 2 only predicts well on the same data distribution, being not able to distinguish well between normal and pneumonia cases on the test set XP2.

Comparing both models, it's possible to verify that model 1 outperformed model 2 by attaining higher accuracy, recall and precision levels on both unrelated datasets, thus demonstrating that it is able to generalize well on different data distributions.

6.2 FUTURE WORK

The work developed during this thesis consisted of the development of CNN for the classification of CXR images into the ones obtained from individuals with or without pneumonia. Altogether, the developed models showed a considerable capacity to perform the intended task not only in the dataset used for training but also in unseen data. However, there is still room for improvement in the continuation of the work herein performed. The priorities for future work are presented in this section.

A relevant limitation to the development of CNN for the purpose of pneumonia classification is the quantity of available CXR images and more importantly the correct classification of these images. Furthermore, the term pneumonia is still used to classify a broad range of lung diseases that may have differences in disease-related presentations. Overall, an increase in the quantity and quality of available CXR images will likely have a relevant impact in the performance of CNN developed for this task. There is also large variability in the CXR in what regards to instance in the area of the images dedicated to the lungs.

One of the possible ways to improve the performance of the prediction of the CNN models will be the implementation of a lung segmentation procedure or algorithm. This approach will focus the feature extraction process only on the regions of interest, that are in this case the lungs. By removing the background present in CXR the possibility of have undesired inputs influencing the output will be reduced. In addition to image segmentation a deep optimization could also positively influence the performance of the CNNs.

The process of testing hyperparameters is very resource and time-consuming. However, finding the best combinations of hyperparameter values is essential to find the optimal model performance. A deeper hyper optimization procedure could be implemented in the future by a exhaustive search method that can drawn more detailed information about the effect of hyperparameters on model behavior and performance.

Nowadays, in addition to classification tasks, object detection tasks are gaining relevance in applications in the field of Health. Object detection tasks can be applied on the detection and localization of disease-related organ abnormalities. It would be interesting to evolve this work to an object detection task in order to detect specific lung patterns associated with pneumonia. In order to analyze those regions, the development of a graphical interface to heat map the lung affected areas would be considered essential.

The CXR are the most available data in medical imaging with a large unexplored potential related to automated analysis and being explored in many research and development projects at radiology and technology field. However, the level of information related with lung disease that can be extracted for CXR is potentially less that what could be obtained by other X-ray based diagnostic methods. Working with more detailed images, such 3D images from CT scans, could be more accurate during the feature extraction of pneumonia cases. Of course, the use of these images or videos will also place more stress on the required computational resources that could still be a limitation for the development and optimization of DL models. Paradoxically, the high information present on these data could also become a barrier on many studies of this kind. Applying transfer learning methods would help to use more deepest and published CNN, such AlexNet and VGGnet. These methods allow to storing knowledge acquired while solving one problem and applying it on a different but related problem.

Collectively, the work presented in this thesis represents advances for the application of DL for pneumonia diagnostic opening new lines of investigation for the future in order to achieve this complex task. This is of large relevance due to the global burden of pneumonia and other ILD. Furthermore, the imminent health concerns provoked for instance by outbreaks of pneumonia-causing virus also raise the demand for automated and efficient methods for the study, characterization and detection of these diseases.

BIBLIOGRAPHY

[1] William J. DePaseo and Richard H. Winterbauer. Interstitial lung disease. *Disease-a-Month*, 37(2):67–133, 1991.

[2] G Gibson, Robert Loddenkemper, Yves Sibille, and Bo Lundbäck. The European Lung White Book. *European Respiratory Society*, pages 256–269, 2013.

[3] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Transactions on Medical Imaging*, 2016.

[4] B. T. SOCIETY and S. O. C. COMMITTEE. The Diagnosis, Assessment and Treatment of Diffuse Parenchymal Lung Disease in Adults. *Thorax*, 1999.

[5] Luna Gargani and Eugenio Picano. The risk of cumulative radiation exposure in chest imaging and the advantage of bedside ultrasound, 2015.

[6] Guk Kim, Kyu-Hwan Jung, Yeha Lee, Hyun-Jun Kim, Namkug Kim, Sanghoon Jun, Joon Seo, and David Lynch. Comparison of shallow and deep learning methods on classifying the regional pattern of diffuse lung disease. *Journal of Digital Imaging*, 31, 10 2017.

[7] Tomas Plankis, Algimantas Juozapavičius, Egl Stašien, and Vytautas Usonis. Computer-aided detection of interstitial lung diseases: A texture approach. *Nonlinear Analysis: Modelling and Control*, 22(3):404–411, 2017.

[8] Chunli Qin, Demin Yao, Yonghong Shi, and Zhijian Song. Computer-aided detection in chest radiography based on artificial intelligence: A survey. *BioMedical Engineering Online*, 17(1):1–23, 2018.

[9] A. Marblestone, G Wayne, and K Kording. Toward an Integration of Deep Learning and Neuroscience. *Frontiers in computational neuroscience*, 10, 2016.

[10] June Lee, Sanghoon Jun, Young Cho, Hyunna Lee, Guk Kim, Joon Seo, and Namkug Kim. Deep learning in medical imaging: General overview, 2017.

[11] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: Past, present and future, 2017.

[12] Mingchen Gao, Ulas Bagci, Le Lu, Aaron Wu, Mario Buty, Hoo Chang Shin, Holger Roth, Georgios Z. Papadakis, Adrien Depeursinge, Ronald M. Summers, Ziyue Xu, and Daniel J. Mollura. Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 2018.

[13] Yoshikazu Inoue. Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias: Summarized points of the new classification. *Japanese Journal of Chest Diseases*, 73(11):1280–1287, 2014.

[14] Rajendra Prasad, Nikhil Gupta, Abhijeet Singh, and Pawan Gupta. Diagnosis of idiopathic pulmonary fibrosis: Current issues. *Intractable and Rare Diseases Research*, 4(2):65–69, 2015.

[15] Praveen Akhutota and Peter Weller. Eosinophilic Pneumonias. *Clinical Microbiology Reviews*, 25(4):649–660, 2012.

[16] Ting Wen and Marc Rothenberg. The Regulatory Function of Eosinophils. *Microbiol Spectr.*, 4(5), 2016.

[17] Gillian Bain and Christopher Flower. Pulmonary eosinophilia. *Radiology*, 23:3–8, 1996.

[18] Jang Won. Acute eosinophilic pneumonia. *Tuberculosis and respiratory diseases*, 74(2):51–55, 2013.

[19] Eric Marchand and Jean Cordier. Idiopathic chronic eosinophilic pneumonia, 2006.

[20] Hilário Nunes, Diane Bouvry, Paul Soler, and Dominique Valeyre. Sarcoidosis Review. *Orphanet Journal of Rare Diseases*, 2:46, 2007.

[21] Ganesh Raghu, Marianna Sockrider, Hrishikesh Kulkarni, Ginger Spitzer, and Robert Baughman. What is sarcoidosis?, 2018.

[22] Robert Baughman, Elyse Lower, and Roland Bois. Sarcoidosis. pages 1111–1118, 2003.

[23] Natalia Soto-Gomez, Jay Peters, and Anoop Nambiar. Diagnosis and Management of Sarcoidosis. *American family physician*, 93(10):840–8, 2016.

[24] Jie Shen and Shicheng Feng. Bone Langerhans cell histiocytosis with pulmonary involvement in an adult non-smoker: A case report and brief review of the literature. *Molecular and Clinical Oncology*, 6(1):67–70, 2017.

[25] Chalinee Monsereenusorn and Carlos Rodriguez-Galindo. Clinical Characteristics and Treatment of Langerhans Cell Histiocytosis, 2015.

[26] Francis McCormack, William Travis, Thomas Colby, Elizabeth Henske, and Joel Moss. Lymphangioleiomyomatosis - Calling it what it is: A low-grade, destructive, metastasizing neoplasm. *American Journal of Respiratory and Critical Care Medicine*, 186(12):1210–1212, 2012.

[27] Connie Glasgow, Souheil El-Chemaly, and Joel Moss. Lymphatics in lymphangioleiomyomatosis and idiopathic pulmonary fibrosis. *European Respiratory Review*, 21(125):196–206, 2012.

[28] Nilo Avila, Andrew Dwyer, Antoinette Rabel, and Joel Moss. Sporadic lymphangioleiomyomatosis and tuberous sclerosis complex with lymphangioleiomyomatosis: comparison of CT features. *Radiology*, 242(1):277–85, 2007.

[29] Joshua Solomon and Aryeh Fischer. Connective tissue disease-associated interstitial lung disease : A focused review, 2015.

[30] So Koo and Soo Uh. Treatment of connective tissue disease-associated interstitial lung disease: The pulmonologist's point of view, 2017.

[31] Deborah Assayag and Christopher Ryerson. Determining respiratory impairment in connective tissue disease-associated interstitial lung disease, 2015.

[32] Gian Sforza and Androula Marinou. Hypersensitivity pneumonitis: A complex lung disease. *Clinical and Molecular Allergy*, 15(1):1–8, 2017.

[33] Jordan Fink, Hector Ortega, Herbert Reynolds, Yvon Cormier, Leland Fan, Terl Franks, Kathleen Kreiss, Steven Kunkel, David Lynch, Santiago Quirce, Cecile Rose, Robert Schleimer, Mark Schuyler, Moises Selman, Douglas Trout, and Yasuyuki Yoshizawa. Needs and opportunities for research in hypersensitivity pneumonitis, 2005.

[34] Anja Roden and Philippe Camus. Iatrogenic pulmonary lesions, 2018.

[35] Martin Schwaiblmair. Drug Induced Interstitial Lung Disease. *The Open Respiratory Medicine Journal*, 2012.

[36] Douglas Flieder and William Travis. Pathologic characteristics of drug-induced lung disease, 2004.

[37] Kun Kim, Chang Kim, Min Lee, Kyung Lee, Choong Park, Seok Choi, and Jong Kim. Imaging of Occupational Lung Disease. *Radiographics*, 21(6):1371–1391, 2001.

[38] Thomas Sporn and Victor Roggli. Occupational lung disease. In *Spencer's Pathology of the Lung, Sixth Edition*. 2012.

[39] J. Balmes. Occupational respiratory diseases, 2000.

[40] R. Borie, C. Kannengiesser, N. Nathan, L. Tabèze, P. Pradère, and B. Crestani. Familial pulmonary fibrosis. *Revue des Maladies Respiratoires*, 2015.

[41] Janet Talbert and David Schwartz. Familial Pulmonary Fibrosis. *GeneReviews*, 2015.

[42] Jangsuk Oh, Lingling Ho, Sirpa Ala-Mello, Dominick Amato, Linda Armstrong, Sylvia Bellucci, Gerson Carakushansky, Julia P. Ellis, Chin-To Fong, Jane S. Green, Elise Heon, Eric Legius, Alex V. Levin, H. Karel Nieuwenhuis, A. Pinckers, Naoaki Tamura, Margo L. Whiteford, Hisato Yamasaki, and Richard A. Spritz. Mutation Analysis of Patients with Hermansky-Pudlak Syndrome: A Frameshift Hot Spot in the HPS Gene and Apparent Locus Heterogeneity. *The American Journal of Human Genetics*, 1998.

[43] Souheil El-Chemaly and Lisa Young. Hermansky-Pudlak Syndrome, 2016.

[44] Joel Howell and Ann Arbor. Early Clinical Use of the X-Ray. *Transactions of the American Clinical and Climatological Association*, 2016.

[45] Andrew Maidmen. X-rays. In *Introduction to the Science of Medical Imaging*. 2009.

[46] Jim Lucas. What are X-rays? *Livescience*, 2018.

[47] William Herring. *Learning Radiology*. 2012.

[48] Abi Berger. Bone mineral density scans. *BMJ*, 2002.

[49] C. Staren. The soft tissues. *In: Sutton D, editor. A Textbook of Radiology. 3rd edn*, pages 1025–1048, 1980.

[50] Hongyu Chen, Melissa Rogalski, and Jeffrey Anker. Advances in functional X-ray imaging techniques and contrast agents, 2012.

[51] M. Yaffe and J. Rowlands. X-ray detectors for digital radiography, 1997.

[52] Tim Newman. Are X-rays really safe? *Medicalnewstoday*, 2018.

[53] Barry Kelly. The chest radiograph. *Ulster Medical Journal*, 2012.

[54] Abraham Ittyachen, Anuroopa Vijayan, and Megha Isac. The forgotten view: Chest X-ray - Lateral view. *Respiratory Medicine Case Reports*, 2017.

[55] Harold Goerne and Prabhakar Rajiah. Computed tomography. In *Right Heart Pathology: From Mechanism to Management*. 2018.

[56] Stephen Power, Fiachra Moloney, Maria Twomey, Karl James, Owen O'Connor, and Michael Maher. Computed tomography and patient risk: Facts, perceptions and uncertainties. *World Journal of Radiology*, 2016.

[57] Carlo Liguori, Giulia Frauenfelder, Carlo Massaroni, Paola Saccomandi, Francesco Giurazza, Francesca Pitocco, Riccardo Marano, and Emiliano Schena. Emerging clinical applications of computed tomography. *Medical Devices: Evidence and Research*, 2015.

[58] A. Cormack. Reconstruction of densities from their projections, with applications in radiological physics. *Physics in Medicine and Biology*, 1973.

[59] G. Hounsfield. Computerized transverse axial scanning (tomography): Description of system. *British Journal of Radiology*, 1973.

[60] David Esses, Adrienne Birnbaum, Polly Bijur, Sachin Shah, Aleksandr Gleyzer, and E. Gallagher. Ability of CT to alter decision making in elderly patients with acute abdominal pain. *American Journal of Emergency Medicine*, 2004.

[61] ASTM. Standard Guide for Computed Tomography (CT) Imaging. *ASTM International*, 2005.

[62] Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation. *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2*. 2006.

[63] Katrina Beckett, Andrew Moriarity, and Jessica Langer. Safe Use of Contrast Media: What the Radiologist Needs to Know. *RadioGraphics*, 2015.

[64] Noriyasu Homma. *Theory and applications of CT imaging and analysis*. 2011.

[65] Alan Mclean, Michael Sproule, Michael Cowman, and Neil Tomson. High resolution computed tomography in asthma. *Thorax*, 53:308–314, 1998.

[66] B Sundaram, A Chughtai, and E Kazerooni. Multidetector high-resolution computed tomography of the lungs: protocols and applications. *Journal of Thorax Imaging*, 25:125–141, 2010.

[67] Simon Walsh and David Hansell. High-resolution CT of interstitial lung disease: A continuous evolution. *Seminars in Respiratory and Critical Care Medicine*, 2014.

[68] AbdulazizH Alzeer. HRCT score in bronchiectasis: Correlation with pulmonary function tests and pulmonary artery pressure. *Annals of Thoracic Medicine*, 2008.

[69] Herbert Fred. Drawbacks and limitations of computed tomography: views from a medical educator. *Texas Heart Institute journal*, 2004.

[70] Gwilym Lodwick, Theodore Keats, and John Dorst. The Coding of Roentgen Images for Computer Analysis as Applied to Lung Cancer. *Radiology*, 1963.

[71] Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 2007.

[72] Macedo Firmino, Giovani Angelo, Higor Morais, Marcel Dantas, and Ricardo Valentim. Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. *BioMedical Engineering Online*, 2016.

[73] Sang Park, Jun Tan, Xingwei Wang, Dror Lederman, Joseph Leader, Hyung Kim, and Bin Zheng. Computer-aided detection of early interstitial lung diseases using low-dose CT images. *Physics in Medicine & Biology*, 56(4):1139–1153, 2011.

[74] Stuart Shapiro. Artificial Intelligence (AI). In *Encyclopedia of Computer Science*. 2003.

[75] Alan Turing. Computing machinery and intelligence. *Mind*, 1950.

[76] Jack Clark. Why 2015 was a breakthrough year in Artificial Intelligence. *Bloomberg*, 2015.

[77] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. 2002.

[78] Ajit Nazre and Rahul Garg. A deep dive in the venture landscape of ai. 2015.

[79] AN Ramesh, C Kambhampati, JR Monson, and PJ Drew. Artificial intelligence in medicine. *Annals of The Royal College of Surgeons of England*, 86(0):334–338, 2004.

[80] Travis Murdoch and Allan Detsky. The inevitable application of big data to health care, 2013.

[81] Alison Darcy, Alan Louie, and Laura Weiss Roberts. Machine learning and the profession of medicine, 2016.

[82] Harvey Murff, Fern FitzHenry, Michael Matheny, Nancy Gentry, Kristen Kotter, Kimberly Crimin, Robert Dittus, Amy Rosen, Peter Elkin, Steven Brown, and Theodore Speroff. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA - Journal of the American Medical Association*, 2011.

[83] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An Overview of the Deep QA Project. *AI Magazine*, 2010.

[84] Steve Lohr. IBM Is Counting on Its Bet on Watson, and Paying Big Money for It - The New York Times, 2016.

[85] Fei Wang and Anita Preininger. AI in Health: State of the Art, Challenges, and Future Directions. *Yearbook of medical informatics*, 2019.

[86] Scott Mayer Mckinney, Marcin Sienik, Godbole Varun, Godwin Jonathan, and Natasha Antropova. International evaluation of an AI system for breast cancer screening. *Nature*, 2020.

[87] Donald Michie. Learning concepts from data. *Expert Systems with Applications*, 1998.

[88] Shijun Wang and Ronald Summers. Machine learning and radiology, 2012.

[89] Bradley Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy Kline. Machine Learning for Medical Imaging. *RadioGraphics*, 2017.

[90] Yohannes Kassahun, Bingbin Yu, Abraham Temesgen Tibebu, Danail Stoyanov, Stamatia Giannarou, Jan Hendrik Metzen, and Emmanuel Vander Poorten. Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions, 2016.

[91] Rahul Deo. Machine learning in medicine. *Circulation*, 2015.

[92] Miguel Caixinha and Sandrina Nunes. Machine Learning Techniques in Clinical Vision Sciences, 2017.

[93] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 1995.

[94] William S Noble. What is a support vector machine? *Nature Biotechnology*, 2006.

[95] Christopher Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 2007.

[96] Shujun Huang, C. Nianguang, Pedro Pacheco, Shavira Narandes, Yang Wang, and X. Wayne. Applications of support vector machine (SVM) learning in cancer genomics, 2018.

[97] Marek Wesolowski and Bogdan Suchacz. Artificial neural networks: Theoretical background and pharmaceutical applications: A review, 2012.

[98] Zhongheng Zhang. A gentle introduction to artificial neural networks. *Annals of Translational Medicine*, 2016.

[99] Daniel Shiffman. *The Nature of Code*. 2012.

[100] Frank Rosenblatt. Perceptron Simulation Experiments. *Proceedings of the IRE*, 1960.

[101] Amelia J. Averitt and Karthik Natarajan. Going Deep: The Role of Neural Networks for Renal Survival and Beyond. *Kidney International Reports*, 2018.

[102] J. Dheeba, N. Albert Singh, and S. Tamil Selvi. Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. *Journal of Biomedical Informatics*, 2014.

[103] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. 2017.

[104] Jonathan Guo and Bin Li. The Application of Medical Artificial Intelligence Technology in Rural Areas of Developing Countries. *Health Equity*, 2018.

[105] Quentin Chometon. Assessment of lung damages from CT images using machine learning methods . 2018.

[106] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 1980.

[107] Chensi Cao, Feng Liu, Hai Tan, Deshou Song, Wenjie Shu, Weizhong Li, Yiming Zhou, Xiaochen Bo, and Zhi Xie. Deep Learning and Its Applications in Biomedicine, 2018.

[108] Krzysztof Cios, Hiroshi Mamitsuka, Tomomasa Nagashima, and Ryszard Tadeusiewicz. Computational intelligence in solving bioinformatics problems. *Artificial Intelligence in Medicine*, 2005.

[109] Dong Yu, Li Deng, Inseon Jang, Panos Kudumakis, Mark Sandler, and Kyeongok Kang. Deep learning and its applications to signal and information processing. *IEEE Signal Processing Magazine*, 2011.

[110] Geoffrey Hinton, Simon Osindero, and Yee Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006.

[111] Vinod Nair and Geoffrey Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[112] Martin Längkvist, Lars Karlsson, and Amy Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 2014.

[113] Y Bengio, P Lamblin, and D Popovici. Greedy Layer-Wise Training of Deep Networks. *Advances in Neural Information Processing Systems*, 2007.

[114] Dan Ciregan, Ueli Meier, and Jurgen Schmidhuber. Multi-column deep neural networks for image classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[115] K. K. Lai. An Integrated Data Preparation Scheme for Neural Network Data Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2006.

[116] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[117] Wei Zhang, Chuanhao Li, Gaoliang Peng, Yuanhang Chen, and Zhujun Zhang. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mechanical Systems and Signal Processing*, 2018.

[118] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology, 2018.

[119] Wenfeng Gong, Hui Chen, Zehui Zhang, Meiling Zhang, Ruihan Wang, Cong Guan, and Qin Wang. A novel deep learning method for intelligent fault diagnosis of rotating machinery based on improved CNN-SVM and multichannel data fusion. *Sensors (Switzerland)*, 2019.

[120] Min Xia, Teng Li, Lin Xu, Lizhi Liu, and Clarence W. De Silva. Fault Diagnosis for Rotating Machinery Using Multiple Sensors and Convolutional Neural Networks. *IEEE/ASME Transactions on Mechatronics*, 2018.

[121] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2017.

[122] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting Chun Wang, Andrew Tao, and Bryan Catanzaro. Image Inpainting for Irregular Holes Using Partial Convolutions. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[123] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei Ling Shyu, Shu Ching Chen, and S. S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications, 2018.

[124] Jürgen Schmidhuber. Deep Learning in neural networks: An overview, 2015.

[125] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

[126] Mingyang Jiang, Yanchun Liang, Xiaoyue Feng, Xiaojing Fan, Zhili Pei, Yu Xue, and Renchu Guan. Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, 2018.

[127] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.

[128] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[129] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 1999.

[130] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT 2010 - 19th International Conference on Computational Statistics, Keynote, Invited and Contributed Papers*, 2010.

[131] Leslie N Smith. Disciplined approach to neural network hyper-parameters. *arXiv:1803.09820v2 [cs.LG]*, 2018.

[132] Justine Boulent, Samuel Foucher, Jérôme Théau, and Pierre Luc St-Charles. Convolutional Neural Networks for the Automatic Identification of Plant Diseases, 2019.

[133] B. S. Everitt and A. Skrondal. *The Cambridge Dictionary of Statistics*. 2010.

[134] Andrei Dmitri Gavrilov, Alex Jordache, Maya Vasdani, and Jack Deng. Preventing Model Overfitting and Underfitting in Convolutional Neural Networks. *International Journal of Software Science and Computational Intelligence*, 2019.

[135] Hossin M and Sulaiman M.N. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 2015.

[136] Aditya Mishra. Metrics to Evaluate your Machine Learning Algorithm. *Towards Data Science*, 2018.

[137] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[138] Ulas Bagci, Jianhua Yao, Albert Wu, Jesus Caban, Tara N. Palmore, Anthony F. Suffredini, Omer Aras, and Daniel J. Mollura. Automatic detection and quantification of tree-in-bud (TIB) opacities from CT Scans. *IEEE Transactions on Biomedical Engineering*, 2012.

[139] Hiroyuki Abe, Heber MacMahon, Junji Shiraishi, Qiang Li, Roger Engelmann, and Kunio Doi. Computer-aided diagnosis in chest radiology, 2004.

[140] Ivan Idris. *NumPy Beginner' s Guide*. 2015.

[141] Aurelien Geron. *Hands-On Machine Learing With Scikit-Learn & Tensor Flow*. 2017.

[142] Fabio Nelli. *Python data analytics: With Pandas, NumPy, and Matplotlib: Second edition*. 2018.

[143] John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 2007.

[144] Autonomio. Talos [Computer software]., 2019.

[145] Spyder. SPYDER IDE, 2018.

[146] Tiago Carneiro, Raul Victor Medeiros Da Nobrega, Thiago Nepomuceno, Gui Bin Bian, Victor Hugo C. De Albuquerque, and Pedro Pedrosa Reboucas Filho. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, 2018.

[147] Street Rosslyn. Digital Imaging and Communications in Medicine (DICOM) Part 1 : Introduction and Overview. *Access*, 2006.

[148] S. Sheard. The Chest X-ray: a Survival Guide. *Clinical Radiology*, 2009.

[149] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalena Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 2018.

[150] Joshua Broder. *Diagnostic Imaging for the Emergency Physician*. 2011.

[151] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using Convolutional Networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[152] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis, 2017.

[153] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI, 2019.

[154] Rangachari Anand, Kishan G. Mehrotra, Chilukuri K. Mohan, and Sanjay Ranka. An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets. *IEEE Transactions on Neural Networks*, 1993.

[155] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 2019.

[156] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002.

[157] Shoujin Wang, Wei Liu, Jia Wu, Longbing Cao, Qinxue Meng, and Paul J. Kennedy. Training deep neural networks on imbalanced data sets. In *Proceedings of the International Joint Conference on Neural Networks*, 2016.

[158] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[159] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *ACM International Conference Proceeding Series*, 2007.

[160] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 2016.

[161] DeepRadiology. Pneumonia Detection in Chest Radiographs. *arXiv.org*, 2018.

[162] Vijay Kotu and Bala Deshpande. Data Mining Process. In *Predictive Analytics and Data Mining*. 2015.

[163] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[164] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2010.

[165] Okeke Stephen, Mangal Sain, Uchenna Joseph Maduh, and Do Un Jeong. An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare. *Journal of Healthcare Engineering*, 2019.

[166] A. A. Saraiva, N. M. Fonseca Ferreira, Luciano Lopes De Sousa, Nator C. Costa, José Vigno Moura Sousa, D. B.S. Santos, Antonio Valente, and Salviano Soares. Classification of images of childhood pneumonia using convolutional neural networks. In *BIOIMAGING 2019 - 6th International Conference on Bioimaging, Proceedings; Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019*, 2019.

[167] Sivaramakrishnan Rajaraman, Sema Candemir, Incheol Kim, George Thoma, and Sameer Antani. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Applied Sciences (Switzerland)*, 2018.

[168] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. Deep learning for medical image processing: Overview, challenges and the future. In *Lecture Notes in Computational Vision and Biomechanics*. 2018.

[169] Rui Zhao, Ruqiang Yan, Zhenghua Chen, Kezhi Mao, Peng Wang, and Robert X. Gao. Deep learning and its applications to machine health monitoring, 2019.

[170] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[171] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[172] Xiaoxuan Liu, Livia Faes, Aditya U. Kale, Siegfried K. Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R. Ledsam, Martin K. Schmid, Konstantinos Balaskas, Eric J. Topol, Lucas M. Bachmann, Pearse A. Keane, and Alastair K. Denniston. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 2019.