UNIVERSITÀ POLITECNICA DELLE MARCHE

Faculty of Engineering

Master's Degree in Biomedical Engineering

Interpretable classification of computed-tomography scans for identification of lung cancer

Classificazione interpretabile di scanner da tomografia computerizzata per l'identificazione di cancro polmonare

Relatore:

Prof. Laura Burattini

Correlatori:

Dott. **Selene Tomassini**

Dott. **Agnese Sbrollini**

Laureando:

Giulia Bruschi

A.A. 2021 / 2022

Abstract

Among all types of cancer, lung cancer is the most life-threatening disease all over the world. For this reason, its early detection is of crucial importance for a better prognosis. In the medical field, the best imaging technique for cancer diagnosis is computed-tomography. However, it is difficult even for expert doctors to interpret and identify lung cancer from computed-tomography scans or slices. Thus, a very promising help nowadays comes from computer-based diagnostic systems which rely on artificial intelligence, together with machine and deep learning. Artificial intelligence aims to mimic the human brain and so it can be exploited as "a second reader" to support clinicians in lung cancer diagnosis. The aim of this thesis is to propose a supervised deep learning model able to non-invasively discriminate between the main histotypes of non-small-cells lung cancer: adenocarcinoma and squamous cell carcinoma. In particular, this thesis focuses on the interpretability of this model. The interpretability is a necessary requirement in the medical field to let clinicians interpret how these models work. To this aim, this thesis is structured in two implementation parts: a preliminary experiment and a final experiment. In both experiments, the model as well as the procedures used are the same; the only difference relies on the data used. In the preliminary experiment a reduced number of slices (10) for each computed-tomography scan with a lower resolution have been used. In the final experiment, instead, the original data (augmented) have been exploited with 250 slices per scan and an higher resolution. In both experiments hyperparameter tuning has been accomplished to let the model perform as best as possible. Then, the gradient activation mapping has been exploited for generating heatmaps which are graphical representations of where the model "is focusing its attention the most". These heatmaps have been then superimposed to the original computed-tomography slices and a video for the dynamic visualization of the volumetric computed-tomography scan has been created. The performances achieved have then been evaluated by focusing on the area under the curve and the recall measurements which are the most significant metrics in the medical field. It turned out that the model reaches better performances in the final experiment. This was expected because more data with higher resolution allow the model to "learn better". The achieved area under the curve in the final experiment is 62% and the recalls relative to adenocarcinoma and squamous cell carcinoma are respectively 55% and 70%. These results mean that the model is able to better discriminate the squamous cell carcinoma. This is also reflected in the generated activation maps, which are more confined in the external edge of the medial part of right lung (where the offending lung nodule is). On the contrary, looking at the activation maps for the adenocarcinoma class, they are less focused. The achieved results in terms of area under the curve and accuracy are not so high, however, asking the modle to identify the tumour histotype prior to lung biopsy is a very demanding task. So, the work

presented in this thesis can be considered to be a very promising starting point for future improvements.

Index

Introduction

Cancer is defined as abnormal mass of tissue, the growth of which exceeds and is uncoordinated with that of the surrounding tissue. Among all types of cancer, lung cancer is the most life-threatening disease all over the world. According to the World Health Organization, lung cancer deaths have become more numerous than the deaths from prostate, breast, brain, and colorectal cancer combined. It has now become the most common cause of cancer deaths in men and the second most common in women.

There are different types of lung cancer and the most common are non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC is a term that includes a variety of different lung cancers, most notably adenocarcinoma (ADC) and squamous cell carcinoma (SCC) which are the most common ones.

The early detection of cancer plays crucial role in preventing cancerous cells from multiplying and spreading. However computed tomography (CT) scan imaging is best imaging technique in the medical field, it is difficult for doctors to interpret and diagnose the cancer from CT scan images or slices. For this reason many studies nowadays aim to develop computer aided diagnosis (CAD) systems. These help physicians to make diagnoses, acting as second readers. In current years a promising frontier in cancer detection from CT images and CT slices is represented by the artificial intelligence (AI). The AI gives a device some form of human-like intelligence which can be exploited for helping the clinicians in lung cancer diagnosis. The aim of this work is to propose a supervised ML model able to discriminate ADC and SCC histotypes from whole lungs CT scans. The work proposed here is subdivided in two parts; a preliminary experiment and an advanced experiment. In both experiments has been used the same model with different data. In the preliminary experiment a reduced number of slices per scan with a poor resolution have been considered. For the advanced experiment instead original scans with an higher number of slices and an higher resolution have been exploited. In both experiments firstly has been performed an hyperparameter tuning for improving as much as possible the network's performances evaluated in terms of area under the curve and some other metrics. However, the problem of AI-based models is that these are a sort of "black-box" thus the observer cannot understand how they work. For this reason in this work the attention is put also on model interpretability, a concept of fundamental importance in medical field. It makes the model "more transparent" and so the clinicians' confidence level in these model increases. The gradient activation mapping has been exploited for generating heatmaps which have been then superimposed to the original CT slices. With the obtained superimposed images have been created a video for the dynamic visualization of the areas of interest for the network.

1 Anatomy and physiology of respiratory system

The respiratory system is composed of a set of hollow, canal-shaped organs, the airways, and lungs, organ parenchymatous in which the function of haematosis occurs, i.e. the exchange of gas between air and blood. The airways are distinguished in:

- Superior: consisting of the external nose, nasal and paranasal cavities and from the nasopharynx.
- Inferior, consisting of the laryngeal tracheal duct and the bronchi [1].

The main function of the upper respiratory tract is to filter the air and thus protect the conduction and exchange surfaces of the lower tract, which are very delicate. The filtering mechanisms are the basis of the defence system of the respiratory apparatus [2].

The internal walls of these organs are covered by a mucous membrane which has various functions besides that of coating, such as that of heating (with its rich vascularization), of humidifying (with the secretion of its glands) and to filter (with the mucus and with the action of the eyelashes) the air that comes inhaled before it reaches the lungs [1].

Functionally, the respiratory system can be divided into a conducting zone and a respiratory zone. The conducting zone of the respiratory system includes those organs and structures which are not directly involved in gas exchange. The respiratory zone instead, is where the gas exchanges take place. The respiratory zone begins where the terminal bronchioles join a respiratory bronchiole, the smallest type of bronchiole, which then leads to an alveolar duct, opening into a cluster of alveoli.

The functions of the respiratory system are multiple and most of them are carried out in cooperation with the lymphatic system, the cardiovascular system and the nervous system, as well as certain skeletal muscles [3].

1.1 Structural organization and anatomy of lungs

The lungs are the major organs of the respiratory system; they are 2, right and left, contained in the pleuropulmonary lodges respectively right and left of the thoracic cavity, separated medially by an area called mediastinum. The pleuropulmonary lodges are delimited laterally by the ribs and intercostal muscles, medially by the mediastinum, inferiorly by the diaphragm and superiorly by the organs of the upper chest opening. Actually the lungs “flutter” in the thoracic cavity due to the fact that do not have an attachment point, but they are suspended from their hilum at the mediastinum. The 2 lungs have a conic shape and can be divided into two parts: the basis which is concave and adapts itself to the diaphragms and the apex which is facing up. The consistency of the lungs is spongy and elastic; this elastic structure collapses in absence of an elastic force that keeps it expanded,

causing the expulsion of air to the outside through the trachea. The colour of the lungs changes during the years: in children is rosy, in adults is swarthy due to the settling of dust inhaled with the air and phagocytosed by macrophages, in the septa on the surface of the lungs. Both the lungs are composed of smaller units called lobes. Fissures in the surface of the lungs separate these lobes from each other. The right lung consists of three lobes: the superior, middle, and inferior lobe. The left lung consists of two lobes: the superior and inferior lobe. The lobes represent a primary division of the lungs in functionally independent parts, with its own vascularization and ventilation. The lungs can be further subdivided in pulmonary segments and successively in pulmonary lobules. In turn every lobule is formed by pulmonary acini [2].

The two lungs present some differences: the right lung is shorter and wider than the left one and the left lung occupies a smaller volume than the right one. In the surface of the left lung is present an indentation, the cardiac notch, which allows space for the heart [3].

Each lung is wrapped in a membrane enveloped by a double-walled serous membrane, the pleura, which makes up the completely closed pleural sacs. In the thin space between the 2 pleural sheets of each sac (visceral and parietal), there is a negative pressure which allow the lung to expand and receive atmospheric air during inspiration. The lungs as can be seen in the Figure 1.1 have a quite complex structural organization. Upon entering the lungs, the main bronchi originating from the trachea branch out giving rise to the bronchial tree. Every principal bronchus begins to divide into collateral branches, in proximity of the ilium, the area in the mediastinal face through which principal bronchi pass. The right principal bronchus gives rise to three lobular bronchi, which insert on the three right lobes. The left principal bronchus instead give rise to two lobular bronchi. The lobular

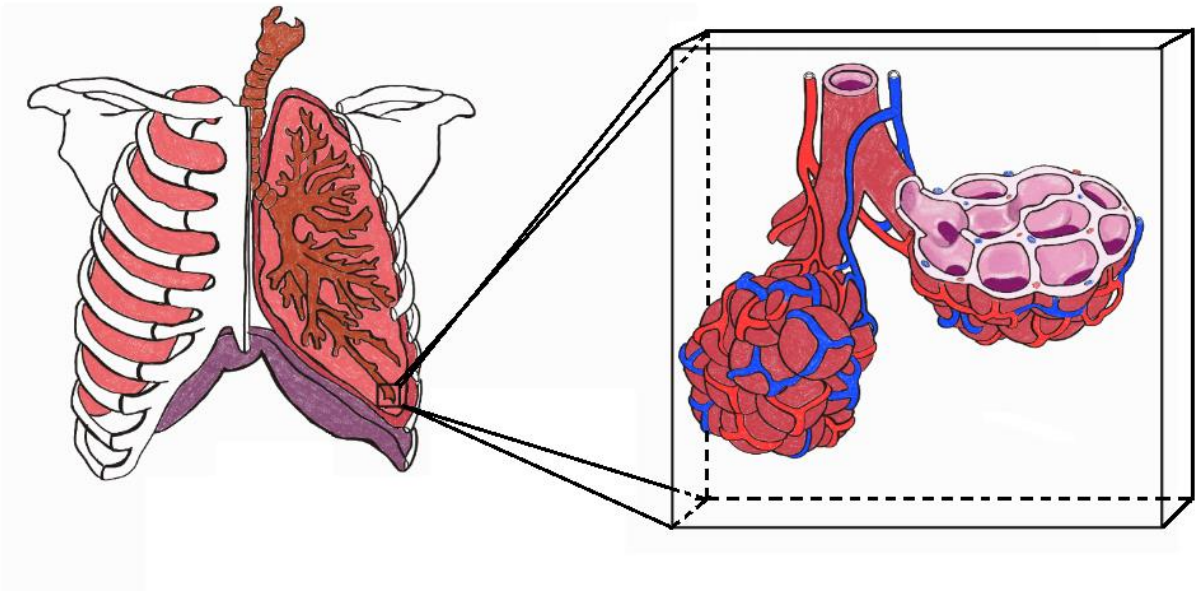

Figure 1.1 Representation of right lung, distribution of zonal bronchi to lobes of the left lung and pulmonary acinus structure.

bronchi branch are in turn divided into zonal bronchi, which then give rise to the interlobular bronchi. These interlobular bronchi branch further into terminal bronchioles, which bifurcates into two respiratory bronchioles. At this level the pulmonary alveoli can be found; here gaseous exchanges take place. These alveoli are closely contiguous one to each other in the pulmonary acinus which is the elementary unit of the lung parenchyma (the set of lung lobules). The pulmonary acinus is defined as the set of branches, provided with pulmonary alveoli, that originate from a terminal bronchiole. In every acinus are present five hundred-two thousand alveoli. Each pulmonary alveolus has the wall lined with a simple paving epithelium composed by two types of cells: small alveolar cells (first type pneumocytes) and large alveolar cells (second type pneumocytes). The large alveolar cells have the typical characteristics of secreting elements, with large cytoplasmic vesicles, containing systems of lamellae arranged in a vortex, the multilamellar bodies. These have the function of pouring their secretion into the alveolus, a surfactant substance that prevents the excessive extension of the alveolus during inspiration and its collapse on exhalation. The alveolar epithelium is supported by a fibroelastic reticular weave, in which can be found a dense capillary network deriving from the pulmonary artery. This fibroelastic texture is responsible for the elastic properties of the lung, which are important during both inhalation and exhalation. The endothelium of the capillaries that surround the alveoli and the alveolar epithelium establish an intimate relationship through the respective basement membranes and constitute the air-blood barrier. This, thanks to the different partial pressure of the oxygen (O_2) in the air and of the carbon dioxide (CO_2) in the blood, allow the gaseous exchanges with no energy expenditure. The inspired air yields O_2 to blood, which from venous becomes arterial; the oxygenated blood drains from the alveoli by way of multiple pulmonary veins, which exit the lungs through the hilum. The CO_2 instead exit from the blood and pass to the alveolus, then is expelled in the external environment through the expiration [4].

As described above, the blood supply of the lungs plays an important role in the haematosis processes and serves as a transport system for gases throughout the body. The most important vessel in lungs is the pulmonary artery that arises from the pulmonary trunk and carries deoxygenated, arterial blood to alveoli. The pulmonary artery branches multiple times as it follows the bronchi, and each branch becomes progressively smaller in diameter. One arteriole and an accompanying venule supply and drain each pulmonary lobule. At the level of the alveoli, the pulmonary arteries become smaller in size giving rise to a pulmonary capillary network which consists of tiny vessels with very thin walls. The capillaries branch and follow the bronchioles and the structure of the alveoli. It is at this point, as previously mentioned, that the capillary wall meets the alveolar wall, creating the respiratory membrane. In addition, innervation by the both the parasympathetic and sympathetic nervous systems provide an important level of control through dilation and constriction of the airway. The

parasympathetic system causes bronchoconstriction, whereas the sympathetic nervous system stimulates bronchodilation. Reflexes such as coughing, and the ability of the lungs to regulate oxygen and carbon dioxide levels, also result from this autonomic nervous system control. Sensory nerve fibres arise from the vagus nerve, and from the second to fifth thoracic ganglia. The pulmonary plexus is a region on the lung root formed by the entrance of the nerves at the hilum. The nerves then follow the bronchi in the lungs and branch to innervate muscle fibres, glands, and blood vessels [3].

1.2 Pulmonary ventilation

Pulmonary ventilation is the act of breathing, which can be described as the movement of air into and out of the lungs. The main function of the pulmonary ventilation is to maintain an adequate alveolar ventilation to prevent accumulation of carbon dioxide in the alveoli. Furthermore it ensures a continuous intake of oxygen thanks to its absorption from the blood flow [5].

The respiratory function is carried out by two acts: inspiration (or inhalation) and expiration (or exhalation):

- Inspiration: the air carrying oxygen from the external environment, arrive to lungs through airways. Mechanically the inspiration consists of contracting the diaphragms (lowering) with a consequent bottom expansion of the lungs.
- Expiration: the air rich of carbon dioxide from lungs retraces in reverse the airways and is expelled into the external environment. In this phase the diaphragms are relaxed, the lungs retract the chest wall and abdominal structures thanks to the elasticity [2].

A respiratory cycle is one sequence of inspiration and expiration. Both acts are dependent on the differences in pressure between atmosphere and lungs, indeed the major mechanisms that drive pulmonary ventilation is the pressure gradient that originates by differences in pressures. This gradient let the air flow from an area of higher pressure to an area of lower pressure. The main pressures taken into account for the pulmonary ventilation are the atmospheric pressure (Patm); the air pressure within the alveoli, called intra-alveolar pressure (Palv); and the pressure within the pleural cavity, called intrapleural pressure (Pip). As can be seen in the Figure 1.2, the atmospheric pressure (considered to be 0 millimetres of mercury (mmHg)) is greater than the intra-alveolar pressure, which in turn is greater than the intra-alveolar pressure, which in turn is greater than the intrapleural pressure. This difference in pressure let the air flow into the lungs during inspiration. The flowing out of the air from the lungs during expiration is based on the same principle; pressure within the lungs becomes greater than the atmospheric pressure. The alveolar and intrapleural pressures are dependent on

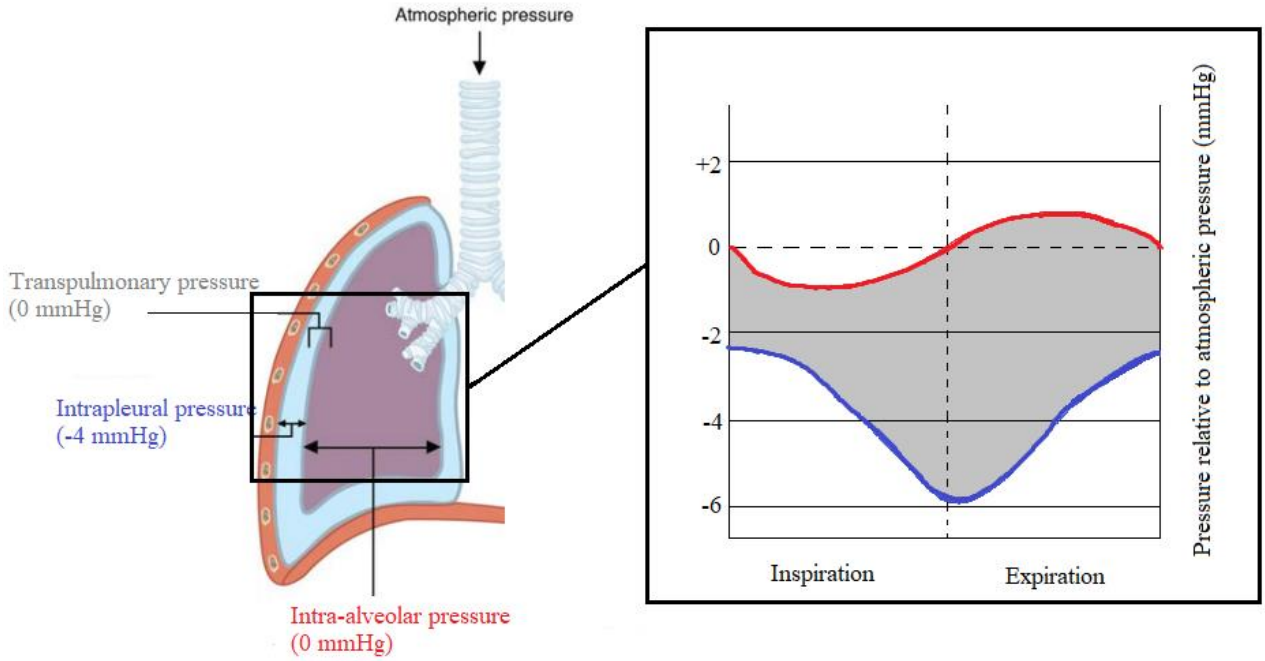

Figure 1.2 Representation of pulmonary pressures values during inspiration and expiration.

certain physical features of the lung. The atmospheric pressure is the amount of force that is exerted by gases in the air surrounding any given surface, such as the body. Atmospheric pressure can be expressed in terms of the unit atmosphere (atm) or mmHg. Typically, for respiration, pressure values are discussed in relation to atmospheric pressure. Intra-alveolar pressure is the pressure of the air within the alveoli, which changes during the different phases of breathing. The airways represent a connection between alveoli and atmosphere thus the interpulmonary pressure of the alveoli, always equalize the atmospheric pressure. The intrapleural pressure is the pressure of the air within the pleural cavity, between the visceral and parietal pleurae. This pressure changes during the different phases of respiration, however, due to certain characteristics of the lungs, the intrapleural pressure is always lower than the intra-alveolar pressure (and therefore also to atmospheric pressure). Although it fluctuates during inspiration and expiration, intrapleural pressure remains approximately -4 mm Hg throughout the breathing cycle. This negative pressure value is due to competing forces within the thorax. There is an inward pulling away from the thoracic wall due to the elastic tissue of the lungs and to the surface tension of the alveolar fluid (mostly water). This inward tension is countered by the pulling outward caused by the surface tension within the pleural cavity, generated by opposing forces of pleural fluid and thoracic wall. Ultimately, the outward pull is slightly greater than the inward pull, creating the -4 mmHg intrapleural pressure relative to the intra-alveolar pressure. The transpulmonary pressure is the difference between the intrapleural and intra-alveolar pressures, and

it determines the size of the lungs. A higher transpulmonary pressure corresponds to a larger lung. In addition to the differences in pressures, breathing is also dependent upon the contraction and relaxation of muscle fibres of both the diaphragm and thorax. The lungs themselves are passive during breathing, meaning that are not involved in creating the movement that helps inspiration and expiration. The main causes of pressure changes that result in inspiration and expiration, are the contraction and relaxation of diaphragm and intercostals muscles (found between the ribs). These muscle movements and subsequent pressure changes cause air to either rush in or be forced out of the lungs. It is clear that an effort must be expended to ventilate. The magnitude of this effort is dependent not only on the pressure gradient, but depends on some characteristics of the lung itself:

- Airway resistance: a force that slows down the flow of gases; mainly due to the reduced size of the airways. In particular there is a direct proportionality between pressure changes and airway resistance.
- Thoracic compliance: the ability of the thoracic wall to stretch while under pressure [3].

As described until here, the pulmonary ventilation consists of movements of air in and out of the lungs. Thus in order to quantify the total amount of air inhaled, exhaled and stored within the lungs at any given time, the respiratory volumes can be measured. The pulmonary ventilation is quantified with the volume of air which enters end exits from the lungs in the unit of time, through the equation (1):

$$V = VC * FR \tag{1}$$

Where V is the total volume of air which enter and exit from the body, VC is the current volume, the air entered and expelled during a single respiratory cycle at rest and FR is the respiratory frequency [5].

Apart from this there are 4 major types of respiratory volumes which are usually evaluated. The tidal volume (TV) which represents the amount of air that normally enters the lungs during quiet breathing (about 500 millilitres (ml)). The expiratory reserve volume (ERV) is the amount of air which can be forcefully exhale past a normal tidal expiration, up to 1200 ml for men. The inspiratory reserve volume (IRV) is produced by a deep inhalation, past a tidal inspiration. This is the extra volume that can be brought into the lungs during a forced inspiration. The residual volume (RV) is the air left in the lungs if you exhale as much air as possible. This volume makes breathing easier by preventing the alveoli from collapsing. The respiratory volumes are dependent on a variety of factors, and measuring them can provide important clues about a person's respiratory health. Combining together two or more selected volume, allow to also retrieve the respiratory capacity, like the functional residual capacity (FRC) which is the amount of air that remain in the lungs after a normal tidal

expiration. It is the sum of expiratory reserve volume and residual volume. In addition to air that create respiratory volumes, the respiratory system also contains anatomical dead space, which is the air present in the airway but that won't never reach the alveoli. For instance the total dead space is the anatomical dead space, together with the alveolar dead space (air within alveoli that are not able to function); it represents the total amount of air in the respiratory system which is not used in the gas exchange process [3].

2 Computed tomography

X-ray CT is an imaging method aiming for 3-dimensional (3D) visualization of human body.
It exploits x-rays thus it is a ionizing imaging technique. CT was introduced in the clinical practice in 1972 and has revolutionized the medical radiology. For the first time clinicians were able to obtain high-quality tomographic (cross-sectional) images of internal structures of the body. When a traditional radiography is performed, the obtained image onto the photographic film is representative of the differential absorption of x-rays. Its absorption depends on the linear absorption coefficient μ which causes the radiation to exit attenuated from the body. These residual rays hit the photographic film generating a blackening; the lower is the attenuation and the greater is the blackening. Looking at the differential absorptions it is possible to retrieve information about the morphology of the investigated body part. In addition it is also possible to get information about the tissue's type. For example the bones can be recognized because are highly attenuating and so appear to be white in the photographic film [6].
The conventional radiography presents 2 main limitations in examining internal body structures. Firstly, the super-imposition of the 3D information onto a single plane which make diagnosis confusing and often difficult. Secondly, the photographic film usually used for making radiographs, has a limited dynamic range. This means that only objects which have large variations in x-rays absorption relative to their surroundings, will cause sufficient contrast differences on the film to be distinguished by eyes. Thus, whilst details of bony structures can be clearly seen, it is difficult to discern the shape and the composition of soft tissue organs accurately. In such situations, growths and abnormalities within tissues only show a very small contrast difference on the film and consequently, it is extremely difficult to detect them, even after using injected contrast media [7].
Another problem is that the obtained information does not take into account the relative quote of each crossed tissue. The image in the photographic film, represents the sum of all the crossed layers, with no information about their position with respect to the direction of propagation along which the radiation is travelling. For example a chest radiography in frontal position allow to identify the presence of an anomalous body in a lung, but do not give sufficient information for identifying its exact position. With a unique radiography it is not possible to establish the quote of the object under examination. In order to have its exact position, it is necessary to make a second radiography, looking at the lateral side of the patience. To overcome all these limitations, it is better to have a photographic image corresponding to what can be obtained by "cutting" the body. From here the introduction of CT, together with the concept of slice, defined as the quantity of matter contained between 2 parallel planes. The distance between these 2 planes is defined as slice thickness. The slice should be

approximately uniform and with a quite small thickness, ranging from 1 to 10 millimetres (mm). It is always necessary to define the position of the cutting plane with respect to the body patience. Hence with CT examinations can be retrieved tomographic images which are pictures of slices of the patient's anatomy [6].

The first type of tomography is the most ancient one; it is realized with mechanical tomographs which furnish the radiogram of a slice directly on the photographic film. This tomograph consists of the x-ray tube and the radiographic cassette (that contains the radiographic film) connected by a rod hinged in a pin. The x-ray tube is positioned at a certain height (quote) from the table where the patience lies and moves back and forth during the examination with a certain angle. The thickness of the image reproduced on the film depends on the angle of oscillation of the rod to which the x-ray tube is connected. There is an inverse proportionality; the larger is the angle and the smaller is the thickness of the slice. But the mechanical tomography has some limits, identified essentially in the type of image realized. It is well defined only in the fulcrum, the point where all the x-rays always pass. All the other slices are not well defined, but blurred with high noise. To overcome these problems have been introduce the use of reconstructed images by using a computer, thus the computed tomography [8].

The pictures displayed are not photographs but are reconstructed from a large number of absorption profiles (x-ray transmission measurements), called projection data, taken at regular angular intervals around a slice. Each profile is made up from a parallel set of absorption values [7].

Thus the resulting images are tomographic maps of the x-ray linear attenuation coefficient. With CT are generated 3D images, whose fundamental elements are called voxels [6].

All these images normally belong to a tomographic plane which is orthogonal to the feet-head axis (z-axis). The tomograph which allows to create such type of images uses x-rays and it is called computed axial tomography (TAC or CT). The configuration of this machinery with its main components is reported on the Figure 2.1. The fundamental element is the x-ray tube which is fixed on a rotating frame. This spins around an axis, generally coincident with the z axis. Such configuration allows to illuminate the subject from all directions. The x-ray tube is similar in design to the one used for conventional radiography, but it is used the rotating anode for a much higher heat loading. However this is not enough because the heat generated during the functioning of the machine is very high, so it is also used a cooling system. This consists of a forced circulation around the tube. The x-ray tube generates a beam with a well-defined shape and a width along the z direction which can vary (1-5 mm). The thickness of the beam is the thickness of the slice as can be seen in Figure 2.2, therefore it is aimed to get a "line of x-rays", the so called fan beam. Consequently, in order to generate a beam as thin as possible and to define precisely the area that need to be examined, it is necessary to

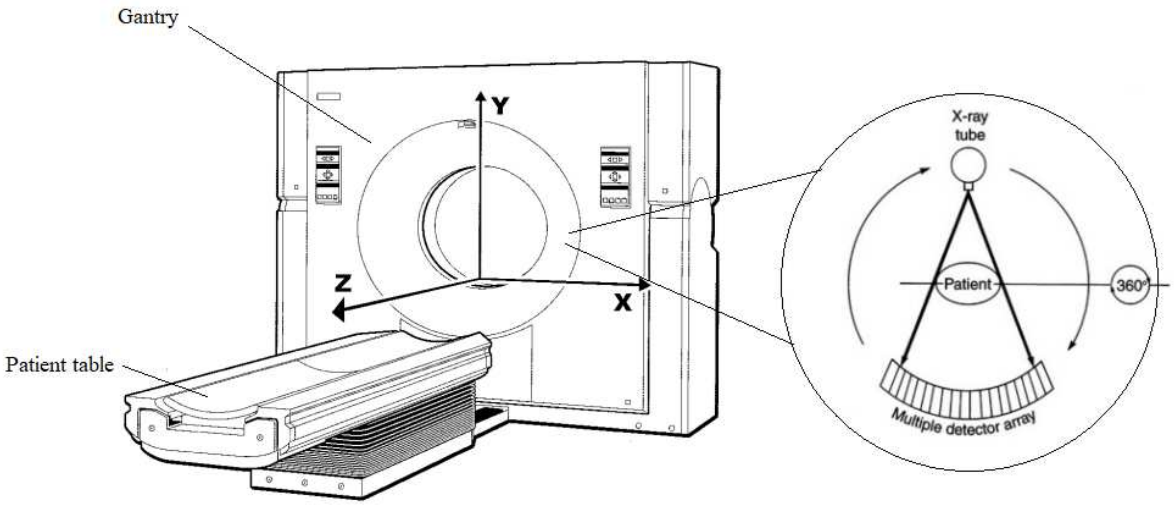

Figure 2.1 Representation of CT structure, with main reference axes.

collimate. In modern systems there is a pre-patience collimation to regulate the divergence angle of aperture of the beam (α). This allows to precisely regulate the area of illumination on the patience's body. Defining and illuminating only the area of interest is a fundamental aspect for what concern the dose, so the amount of radiations delivered to the patience. As previously said, the x-rays are ionizing rays, thus dangerous for people. So it is aimed to use only the necessary amount of radiations. For example the area of irradiation of an obese patient for chest examination, is much wider than that of a children. Therefore the technician reduces the angle of aperture of the beam for the children's examination and consequently the delivered dose is reduced. Then there is also a second collimation stage, called post-patience collimation, at the sensors. This is needed to eliminate the penumbra region, an area in which radiations are not uniform, hence the image generated here is not good. The requirement is to allow detectors to be illuminated only by the uniform part of the beam. All the collimators are lead sheets with a well-defined thickness. In CT it is also needed to have a filter, positioned between the x-ray tube and the pre-patience collimation stage as can be seen in Figure 2.2. This is called Bowtie filter, characterized by a special geometry. It has two main functions; the first one is to stop the soft radiations, which are not useful for the generation of the image but are harmful for the patience. The second function is to let the x-ray beam intensity being as much uniform as

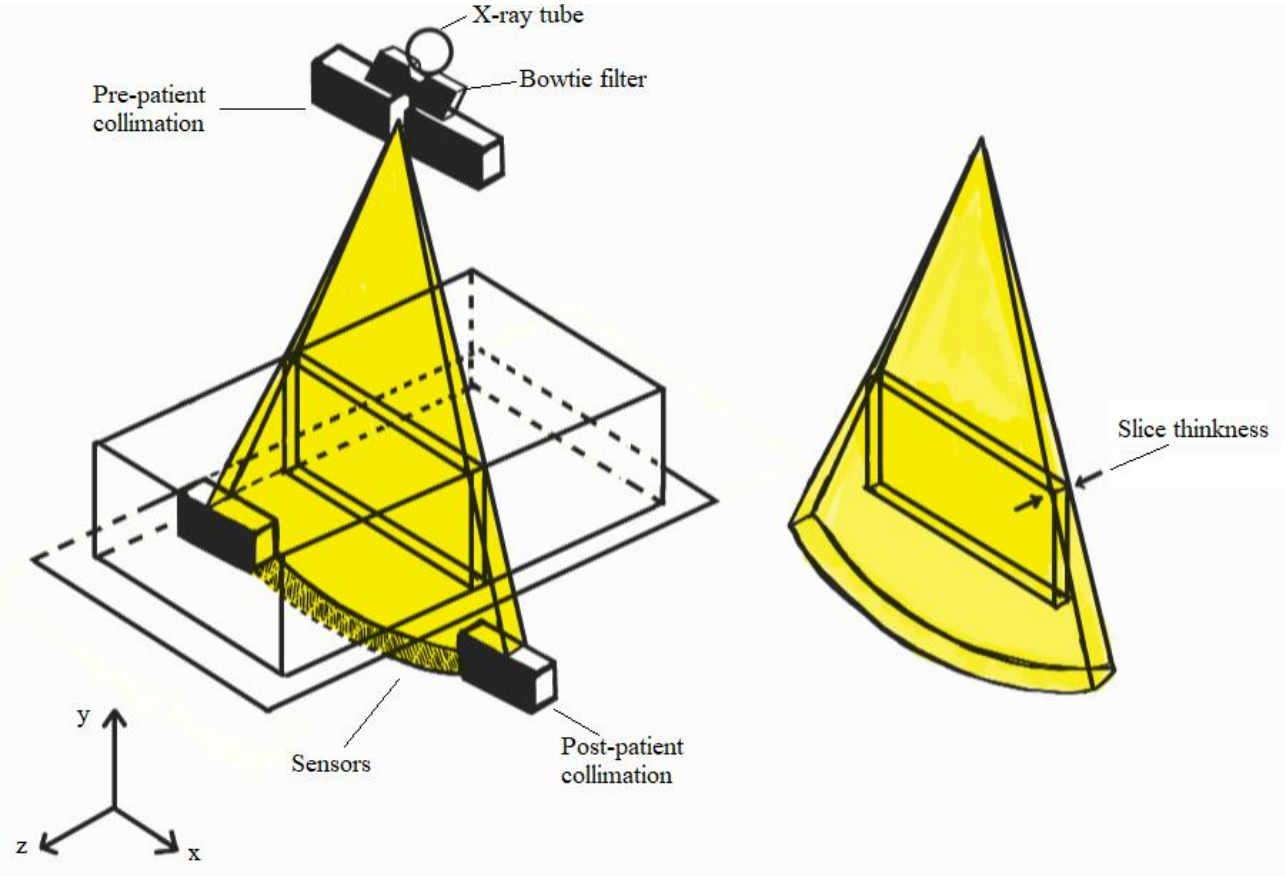

Figure 2.2 Schematization of body as parallelepiped hit by x-ray beam, filtered and collimated.

possible. On the opposite side of the x-ray tube, there need to be a receiver, analogous to photographic film in

- Classical radiography, because x-rays are based on transmission concept. Thus on the same frame, in the opposite side of the x-ray tube, there is a series of x-ray solid state sensors (typically seven hundred-one thousand). Sensors are transducers that convert the received radiation intensity into an electrical signal. The received intensity is proportional to the amplitude of the electric signal which is generated. At the same time, the received intensity is proportional to the absorption, thus the amplitude of the electrical signal is proportional to the absorption of x-rays in the body. There are two basic detector types used: Scintillation detectors and photodiode: are crystals of calcium tungstate or rare earth oxides (the most modern materials) that produce light when hit by x-rays. Then the light is converted into electrical signal through photodiodes. These are the most used.
- Ionization detectors: they are chambers with xenon gas inside, which is ionized when x-rays pass. These ions generate a quite high voltage that can be measured. These are not used anymore.

The frame with x-ray tube and sensors is inside a container called gantry, represented in Figure 2.1. Inside this gantry the x-ray tube and the sensors rotate around the patience table of 360° and always guarantee the alignment between these two elements. There are seven generations of CTs, but the sixth generation is the one most used nowadays. It has been introduced in 1990 as upgrade of the fourth generation. The main innovation is the acquisition modality which is helical. This is obtained by letting the x-ray tube move around a circle in the gantry and simultaneously move the table. When the x-ray tube completes one turn and the patience is moved of 1 mm (rotation and translation simultaneous), a slice is imaged. Each turn requires 1 second, thus the acquisition speed is 1 second/slice. The image is generated with a continuous scanning and the helical acquisition reduces drastically the time required for a complete examination. For example the entire abdomen or chest is imaged in only 30 seconds. This is a huge advantage in terms of reduction of the dose delivered to the patience. This CT generation produces one continuous volume set of data for the entire region of interest [8].

This generation of CT presents 3 important advancements in diagnosis. First of all there is an improved detection; now it is allowed to find out small lesions thanks to continuous slice acquisition. Secondly there is an improved contrast thanks to the fact that the region is imaged in short period of time, so the contrast can be timed. And finally an improved reconstruction and manipulation allow to reconstruct transverse data in any plane ("strip away skin, muscles etc…) [9].

The developers of this CT generation have introduced this one to solve the problem of lung nodules identification. To detect them with standard thick-slice techniques and to repeat the examination with thin slices for a morphometric analysis or to repeat the study after given time intervals to monitor their growth was a pending task. With the standard tomographic technique, it was very hard even to find the nodule again, and if found, it was difficult also to isolate the image where the nodule is visible. Continuous scanning along the patient's longitudinal axis, which is the z-axis in the CT coordinate system, appeared to offer the solution [10].

The examination of a patience with a CT scanner consists mainly of 5 phases:

1. Scanning of the patience: it consists on generating x-rays in the x-ray tube and launch the generated beam towards the patience. The x-ray fan beam pass through the patience and it is received at the sensors attenuated following the equation (2):

$$I = I_0 * e^{-\mu x} \tag{2}$$

where I is the intensity of the beam after attenuation. I_0 is the initial intensity that hits the patience, μ is the linear attenuation coefficient, dependent on the type of material and x is the physical thickness of the absorber.

2. Data collection: it consists of collecting projections, thus multiple attenuation profiles taken all around the patience. For this purpose it is used the data acquisition system (DAS).

At this point the patience can leave the room.

3. Image reconstruction: it consists of the reconstruction of the intensity profile.
4. Image display.
5. Image archival.

As can be seen in the Figure 2.3, the frame with x-ray tube and detectors, rotate around the z axis of the patience with continuity. At each position an x-ray beam is generated and an attenuation profile is collected, which corresponds to the electrical signal coming from each of the receiver sensors. All the attenuation profiles together, taken after a complete round of 360°, give the scan from which can be extracted the slice. In particular, by performing successive scans along the z axis, it is possible to reconstruct sequences of layers (tomograms) to rebuild the 3D structure. The purpose of CT hardware is to acquire a large number of transmission measurements, called projections, through the patience

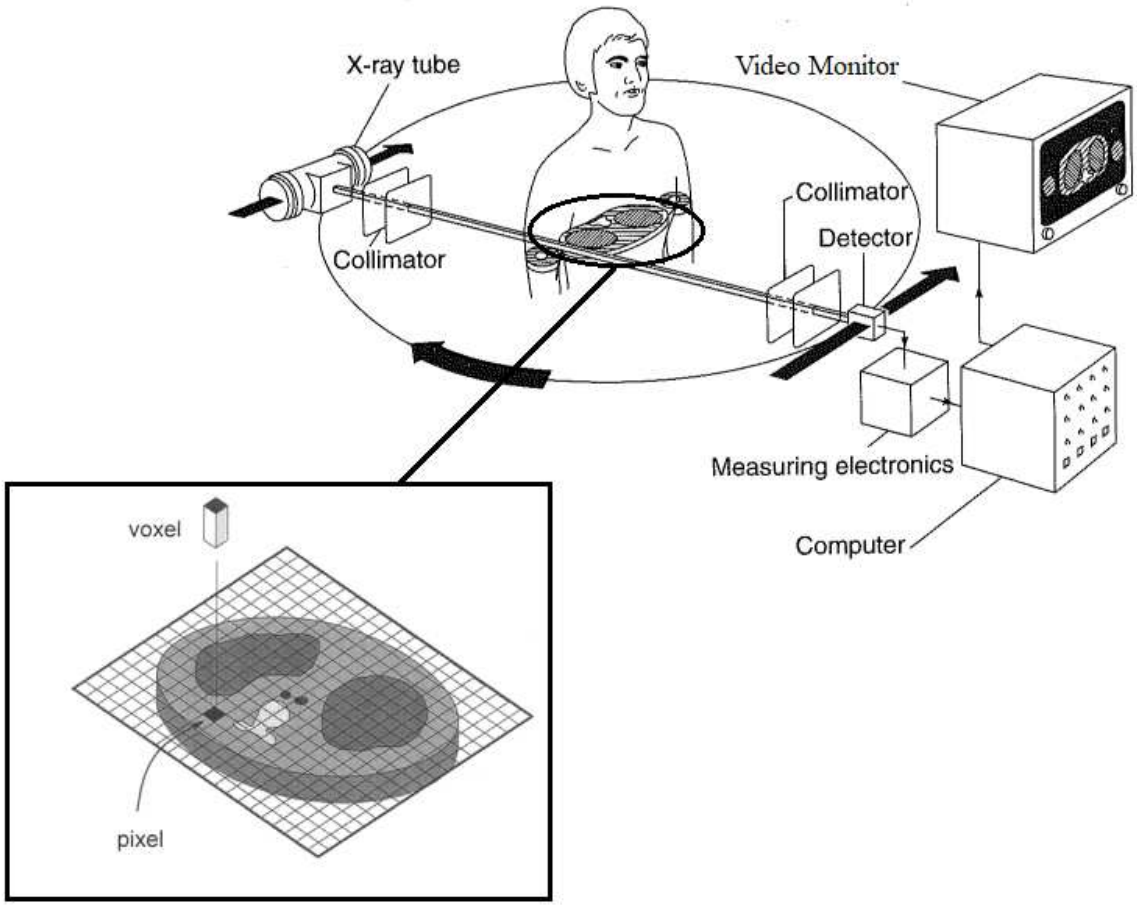

Figure 2.3 Representation of slice acquisition and corresponding pixels and voxels.

at different positions. A single CT image, involves approximately eight hundred rays, taken at one thousand different projection angles. Every single ray impinging on one single sensor creates one value of absorption μ. Each slice is created with a huge amount of data, in this case 800*1000; thus if one hundred rounds are preformed around the patience, there are 800*1000*100 single data. From this slice, with a thickness equal to the thickness of the x-ray beam, can be retrieved the 3D CT images and from this one the 2-dimensional (2D) CT images. The slice is defined as a series of voxels (volumetric elements) on the same plane. The voxels are 3D cubes or parallelepipeds (piles) of absorption. The smallest slice is obtained by considering only a plane of just 1 voxel. From the voxel can be retrieved the pixel, which is the base surface of the voxel as reported in Figure 2.3. The 2D array of pixels in CT image corresponds to an

equal number of 3D voxels. Each pixel represents the average attenuation inside each corresponding voxel. It is like a grid is superimposed to the surface of the slice as reported in Figure 2.3. This grid is generally composed by m rows and r columns. Every element is called pixel and has a dimension of $m\ x\ r$. This is obviously a matrix, usually composed by 256x256 or 512x512 pixels, chosen according to the necessary image resolution and to the availability of calculation capacity of microprocessor. Each pixel is usually a square and it is supposed to be associated to a value equal to the mean linear absorption coefficient in its corresponding voxel. The smaller is the pixel and the more constant is the linear attenuation coefficient. The values of μ for each pixel, can be retrieved by knowing the absorption map and consequently the pixel is associated to a corresponding level of grey. The knowledge of the grey level distribution of each pixel allow to construct the image. Actually in CT in order to represent the linear attenuation coefficient, is not used the physical quantity μ associated to it, but the CT number. This is an a-dimensional quantity linked to absorption as expressed in the equation (3):

$$CT\ = 1000 * \frac{\mu_t - \mu_w}{\mu_w} \tag{3}$$

Where μ_t is the linear absorption coefficient measured at the sensors and μ_w is the linear absorption coefficient of the water, taken as reference. The main and most important quantity in an imaging technique is the CT image resolution, which determines the goodness of the image. Looking at the voxels, along the z axis the resolution is defined by the thickness of the x-ray beam and cannot be changed. It depends only on the collimation capacity as previously said. Instead the resolution in x-y direction can be regulated acting on the displayed field of view (DFOV). This refers to how much scan field of view (SFOV), so scanned anatomy, is reconstructed in terms of voxels. With this parameter can be defined the pixel size according to the formula (4):

$$Pixel\ size = \frac{DFOV}{matrix\ side} \tag{4}$$

Where the matrix side is the dimension of the side of each pixel (usually 512x512). The lower is the number and the better is the resolution, but the drawback is that the examined area is small. For example the best possible resolution is 0.20 mm, but this corresponds to a DFOV of only 10 cm. So in order to have this high resolution for displaying a large area it is needed to perform multiple scanning of 10 cm each. Then the image needs to be reconstructed from all the information received at the sensors (the absorption profiles related to each projection). For this purpose there are multiple reconstruction algorithms. The most widely used is the filtered back projection. This is an analytic reconstruction algorithm that uses all the acquired projection signals to project back the value and all together recreate the slice absorption image. This is done by back projecting the absorption profiles onto the image matrix as reported in the Figure 2.4. It is composed by 2 steps: filtering the data and performing the back projection operation where the data is "painted" back in the image along the direction of measurement. Basically it over-poses the signal from different projections on the same area and sum up all of them. All the projections are added one up to the other through the sinogram, which is the reconstructive algorithm, and in addition it is used also a topographic reconstruction. The drawbacks of this reconstruction algorithm are the star shape and streak patterns present near to the object that need to be displayed. These are called 'Star' and 'streak' artifacts. In order to remove or at least limit as much as possible these artifacts a filter function is applied to each point along the attenuation profile. Different filter functions are used to create sharper (higher resolution) or smoother

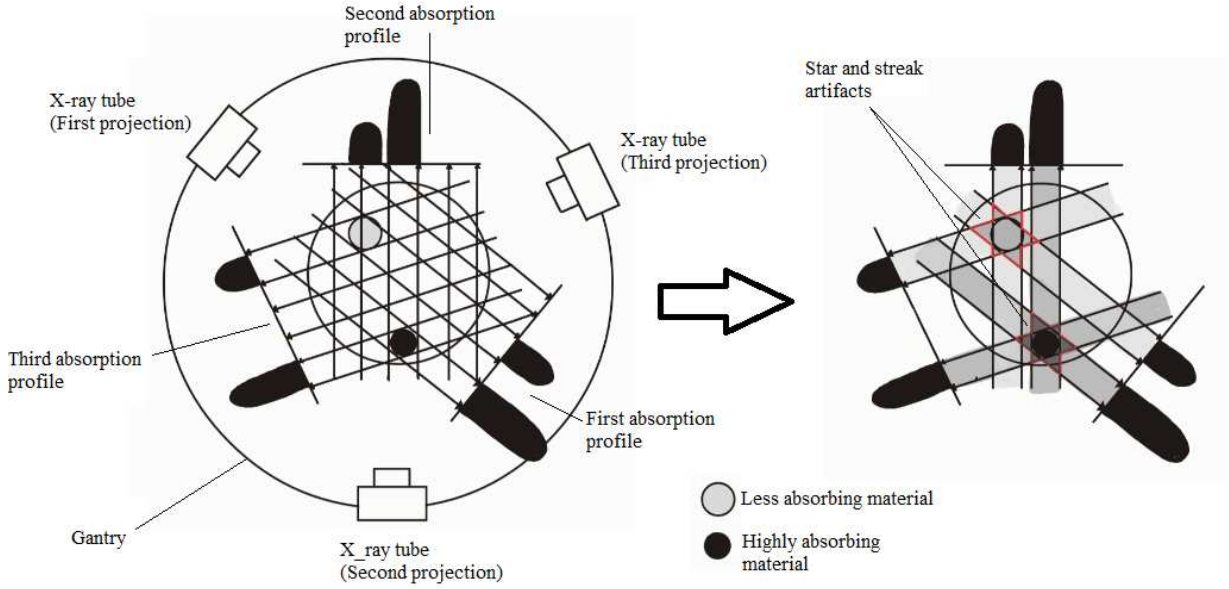

Figure 2.4 Acquisition of absorption profiles and reconstruction of the image with back projection algorithm.

(lower noise) images. Then the image has to be displayed but at this point need to be taken into account some additional aspects concerning the visualization. First of all there is a limitation in displaying CT numbers, linked to the capacity of the system. In theory can be displayed a very wide range of attenuation, from -1000 to +1000, for a total of 3000 possible levels of grey. Every organ, characterized by a different tissue has some specific ranges of CT number values. For example lungs have very low values (from -200 to -1000), due to the fact thar are filled of air. An 8-bit system, can display only 2^8=256 levels and in addition the human eyes can distinguish only about twenty/thirty levels of grey (from here the necessity to have computer-based diagnosis). This problem is solved by regulating the window width and the window level. The window width is the range of CT levels displayed using multiple levels of grey scale. It controls the image contrast. Small windows allow to separate CT numbers one from each other letting possible to appreciate many more details, like differences of concentration and density in a certain tissue. The centre of this CT window is called CT level, which controls the image density. This moves the visible grey scale in post processing allowing to be focusing on whatever the clinician wants [8].

3 Lung cancer

Lung cancer is a type of cancer that starts in the lungs. Lung cancer deaths have become more numerous than the deaths from prostate, breast, brain, and colorectal cancer combined. It has now become the most common cause of cancer deaths in men and the second most common in women. Epidemiology highlights that the use of tobacco is the cause of approximately 90% of all lung cancer. Thus the statistic is now largely declining due to anti-smoking campaigns and decreased tobacco use in the United States [11].

3.1 Carcinogenesis

Cancer is defined as abnormal mass of tissue, the growth of which exceeds and is uncoordinated with that of the surrounding tissue, and that continues to grow in the same excessive manner after cessation of the stimulus that caused it [12].

In order to prevent and cure tumours, a lot of studies aim to discover the processes involved in the genesis of tumoral cells. In particular these studies are mainly focused on the carcinogenesis, defined as the initiation of cancer. Tumours develop in those tissues in which cellular homeostasis has been disturbed by:

- Hyperplastic changes: an increase in the number of cells in tissue or organ that appear normal under a microscope.
- Dysplastic changes: the cells look abnormal under a microscope.
- Regenerative changes: regrowth of a damaged or missing organ part from the remaining tissue [13].

In particular this growth of cells is caused by a somatic cell clone's "loss of rules", caused by genetic and epigenetic changes. A typical malignant cell develops over years in a step-by-step microevolution process, accumulating from 5 up to 10 critical mutations in genes important for apoptosis (cellular division), deoxyribonucleic acid (DNA), repair and other aspects of cell behaviour [14].

Clinical and experimental data have proved that during the division process the cell is more susceptible to carcinogenic factors than at rest [13].

To maintain a long-living multicellular organism, like the human one, is a complex task. Many of its cells have a limited life span. Lost cells have to be replaced. In addition, tissues need plasticity, enabling them to adapt to changing physiological demands. An example of this is the homeostatic mechanisms of red blood cells. An example of this is what happen to erythrocytes when a person performs a physical task and requires more oxygen. Thanks to a feedback mechanism, when there is a reduction in kidney partial oxygen pressure, the production of net erythrocytes is stimulated in the

bone marrow. The kidney cells constantly produce a transcription factor which is broken down fast and efficiently at normal values of partial oxygen pressure at kidneys. As soon as this pressure sinks below a certain threshold, the transcription factor starts to accumulate. Consequently in the nucleus activate its target genes which produce erythropoietin, responsible of an increasing rate of proliferation in erythroid progenitors in bone marrow. These are cells able to proliferate, but lack the ability of doing specialized work. The net effect is an increase in erythrocytes output. This example illustrates that the proliferation is regulated stringently and occurs only in situations where new cells are required. The cells that are able to proliferate are called stem cells and represent only a small part of the cells of a tissue. It is assumed that for most, but not all, types of cancers, mutations are accumulated in a stem cell lineage. In fact some studies have highlighted that in many cases cancerous tissue maintains the division into cancer stem cell and are more or less differentiated. Several safeguards are in place to minimize accumulation of mutations in stem cells:

- Stem cells divide as rarely as possible, only when it is strictly necessary. This is to minimize the number of rounds of DNA replication, which is inherently error-prone. Many stem cells spend most of their lives in a state called dormancy or quiescence.
- Stem cells are in niches which offer protection. For example stem cells of the epidermis have to be protected primarily from UV radiation, so are positioned at the bulge of hair follicles.
- Some stem cells have a lowered threshold for apoptosis. DNA damages in one of these cells, cause the cell to enter apoptosis. The logic suggests that it is better to lose a stem cell than to risk of replication of a damaged genome. Instead this is a problem during radiotherapy treatments, which can lead to necrosis.
- Stem cells in order to protect themselves express high levels of a specific transmembrane protein, encoded by a gene responsible for the multi drug-resistance.

To summarize, in spite of all these protection mechanisms, some mutations can still occur in the stem cell resulting in cancer stem cells [14].

The multistage carcinogenesis theory is generally accepted to describe the processes at the basis of the cancer genesis. This is said to be multistage because between the initial carcinogenic stimulus and the final manifestation of cancer several stages can be identified: initiation, promotion and progression [13].

As can be seen in the Figure 3.1, at the initiation stage mutations occurs in one or more critical genes involved in the control of cell proliferation. These genes in their physiological form are called proto-oncogenes and become oncogene after mutations [15].

These are involved in processing information that reach the cell and result in a structured cell division. In normal conditions these genes respond to extracellular growth factors by switching “off” and “on” some proteins. Specific mutations may block these proteins in the ”on” state. Switching off the signal is physically impossible. For the cell it is like there is a lot of growth factor around. Hence the cell reacts by proliferating [14].

Initiation starts with the action of the carcinogen on chromosomal DNA, inducing a lesion which can be repaired or reproduced. If the specific endonucleases responsible for the lesion identification and deletion fails or act with a certain delay, the lesion will be replicated and transmitted to new cells. Thus the initiation stage of a cell starts with the impossibility of repairing DNA lesions. The initiated cell represents an irreversible alteration of the genetic material and can potentially develop a neoplastic cell clone. Initiation is a rapid process of the order of minutes or hours and confers the cell proliferative capacity which can also remain latent without leading to the promotion stage [13].

Initiation requires one or more rounds of cell division for the fixation of the process. This is irreversible although the cell may eventually die during the development of the neoplasm. Mutations that give rise to cancerous cells can be of many types:

- Point mutations: the replacement of a single base of DNA with another base (as represented in Figure 3.1).

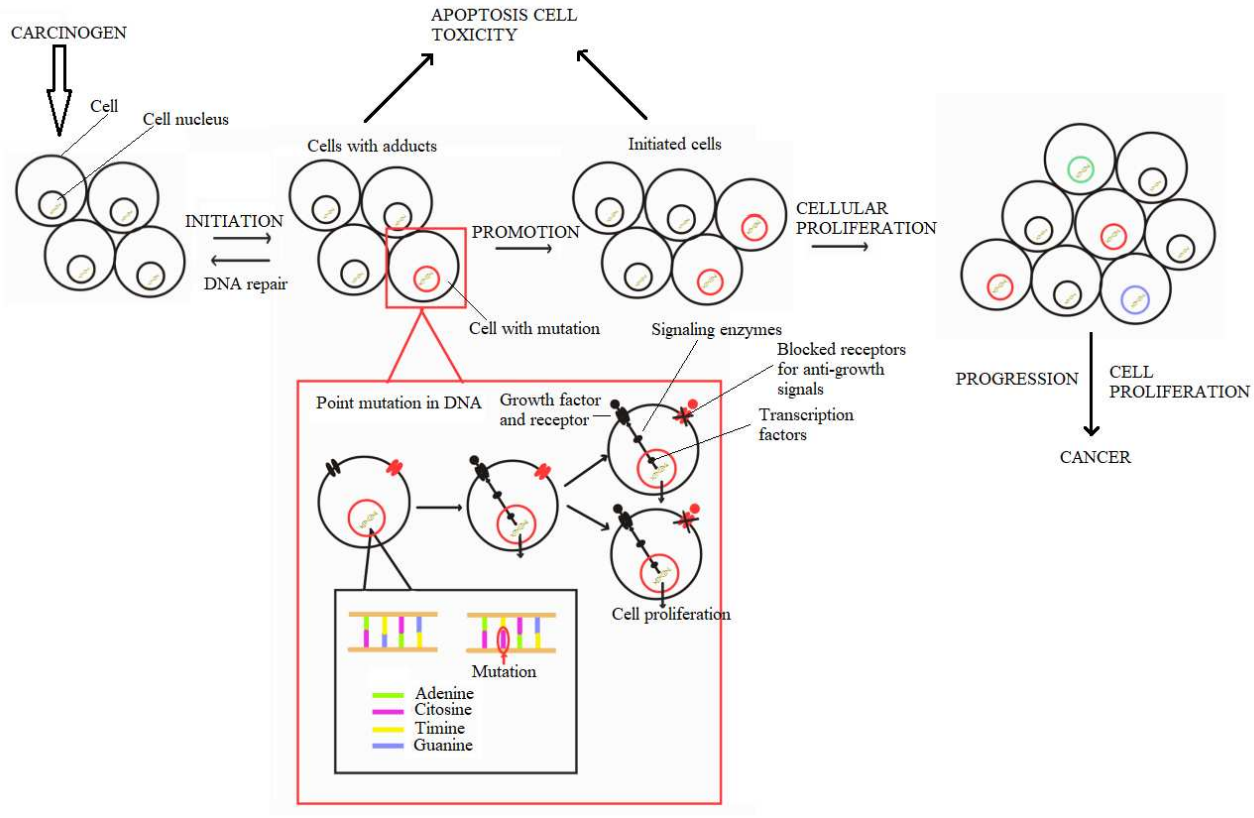

Figure 3.1 Representation of stages of carcinogenesis.

- Frameshift mutations: addition or deletion of a nucleotide such that the protein sequence is altered.
- Chromosomal aberrations: any change at the level of chromosomes, involving their structures or numbers.
- Aneuploidy: the chromosome number is not a multiple of 23, which is the normal haploid.
- Polyploid: more than twice the haploid number of chromosomes.

The substances that somehow cause these mutations and thus cancer, are called carcinogens [15].

The main carcinogenic factors can be grouped into primary determining factors, secondary determining factors and favouring factors. All of them act at molecular level, especially on nucleic acids of RNA and DNA. According to their nature, these factors can be classified as exogenous and endogenous factors. In particular scientific studies have demonstrated that exogenous factors are the main causes of neoplasms. The primary determining factors include chemical substances, the carcinogenic action of physical agents and the action of carcinogenic transformation of viruses. The chemical substances can act either directly causing mutations or indirectly reactivating repressed carcinogens. These substances are distinguished in 3 classes according to their mechanisms of action. Direct action or ultimate carcinogens, whose structure confers them the capacity to induce cancer without a previous metabolic activation in the host organism. Procarcinogens, group that includes the majority of chemical carcinogens which become active after a previous metabolic activation to ultimate carcinogens. Co-carcinogens, chemical substances that cannot induce cancer when they are administered alone, but can enhance the carcinogenic effect of other substances. In general, co-carcinogens act as promoters in tissues in which the initiation stage has appeared. Physical factors instead consist of non-ionizing and ionizing radiations. Non-ionizing radiations are electromagnetic, with low penetration power, ultraviolet and infrared radiations. All these have the sun as main source. What let the radiations being more carcinogenic is the penetrability, in fact ultraviolet B are not so dangerous because can penetrate only for 10% in the skin. The indirectly ionizing radiations instead are particles with no electric charges like photons, gamma rays and x-rays. The latter ones are at the basis of radiography and CT thus it is very important to reduce as much as possible the dose provided to the patience. Regarding the biological action the major impact is at the level of cell nucleus, chromosomes (with rupture, deletion or translocation) and DNA (rupture of a bond causing the breakage of one or two filaments). Also the cytoplasm is a target site. The secondary determining causes instead are represented by hereditary determinism, like the unilateral or bilateral retinoblastoma that affects children. This is determined by a pathological gene. Favouring causes are

risk factors whose intervention (occasional or systematic) can be observed in the incidence of tumours. An example of these are geographic factors, nutrition factors, sex, age etc [13].
The targets of initiation are:

- Mutational activation of oncogenic (proliferative) pathways, like growth factor receptors and signalling proteins, involved in the cell cycle checkpoints.
- Mutational inactivation of apoptotic pathways for inducing cell death, like tumours suppressors (p53, a transcriptional factor that controls cell cycle, apoptosis and DNA repair mechanisms).
- Mutational inactivation of DNA repair mechanism.
- Mutational inactivation of antioxidant response [15].

After the initiation there is the second step called promotion. Chronic genetic alterations of the initiated cell determine the neoplastic transformation and the appearance of cells capable of autonomous growth. At this stage on these modified cells, act the so-called promoters, substances like polycyclic aromatic compounds, that alter the normal growth process by mechanisms similar to hormonal or growth factor. Promoters can be defined as non-carcinogenic or weakly carcinogenic if used alone and need to be applied several times after the initiating carcinogen. Some other promoters instead are so "powerful" that do not require the initiation stage [13].
This is a long-term process (fifteen-thirty years) that starts with one damage cell that has a growth advantage (initiation). A benign mass of abnormal cells, the tumour, is developing. There are 2 key aspects about promotion. It is a stage of cellular growth, with an excessive division and multiplication. And it is characterized by progressive genomic damages that make these cells gnomically unstable. Consequently these cells can potentially develop the skills needed to become cancer cells. This stage as the others, appears to be driven by chronic inflammation, thus controlling it can help to prevent the development of tumoral cells. During this period cells must learn how to survive by avoiding the apoptosis (programmed death) [16].
Finally there is the third stage which is called progression, which is considered to be irreversible. This requires further mutations, leading to chromosomal instability and recruitment of inflammatory immune cells to the tumour. These cells are characterized by gene alterations and rearrangements, even karyotype alterations. At this stage, cells acquire "wound-healing" characteristics, like secretion of chemo-attractants to attract inflammatory immune cells, angiogenesis etc [13].

3.2 Lung cancer types

The World Health Organization (WHO) has proposed a classification of lung tumours. This classification system relies on immunohistochemistry and light microscopy in order to better guide treatments and determine a prognostic course [11].

There are different types of lung cancer, like lung nodules, non-small cell lung cancer, small cell lung cancer and mesothelioma (rare). The most common are NSCLC and SCLC [17].

3.2.1 Small cell lung cancer

Small-cell lung cancer represents about 15% of all lung cancers and it is characterized by an exceptionally high proliferative rate, strong predilection for early metastasis and poor prognosis. SCLC is a high-grade neuroendocrine carcinoma. It is among all the cancer types the one with the strongest epidemiological link to tobacco, and its prevalence tends to mirror the prevalence of smoking, with a lag time of about 30 years. Only 2% of SCLC cases arise in never-smokers (defined as lifetime smoking of fewer than 100 cigarettes). Studies suggest possible links also with exposure to air pollution. Inherited genetic factors are thought to play a minor role in the susceptibility to developing SCLC. Most patients have metastatic disease at diagnosis, with only one-third having earlier-stage disease, thus the importance of an early detection. Genomic profiling of SCLC reveals extensive chromosomal rearrangements and a high mutation burden, almost always including functional inactivation of the tumour suppressor genes TP53 and RB1. Changes in the lung stroma and immune microenvironment also presumably contribute to SCLC tumorigenesis. Analyses highlighted a substantial intra-tumoral heterogeneity. Patients typically present respiratory symptoms, including cough, dyspnoea (laboured breathing) or haemoptysis (coughing up blood), with imaging revealing a centrally located lung mass and often bulky thoracic lymph node involvement. Small cells lung cancer has an exceptionally high mortality rate relative to other common solid tumour. The main problem in the SCLC field is the small amounts of material available for histological diagnosis and subsequent research. The ability to isolate CTCs from the blood of patients with SCLC can alleviate the lack of tumour material. However, there is a still great need for clinical trials that include the collection of tumour material to identify key genetic drivers of SCLC and accelerate both clinical and basic research. Given the aggressive nature of SCLC, diagnostic and staging work-up should be performed as quickly as possible after presentation. This assessment includes imaging (typically contrast-enhanced CT or F-FDG PET/CT of the chest, abdomen and pelvis, and brain MRI with contrast) to define the extent of disease, blood tests, including cell counts, liver and kidney function. Owing to the usual central location of the tumour, biopsies are often

obtained by bronchoscopy. Depending on accessibility, a preferred option can be the biopsy of a distal metastatic site. The diagnosis is only confirmed by histopathological examination. The radiological findings in SCLC are similar to those of other lung cancers, with a tendency for tumours to be larger, centrally located and at a more advanced stage at presentation. Bulky mediastinal lymph nodes are common. Metastatic spread is often radiologically evident and may include pleural and pericardial effusions. Rare cases (about 5% of patients with SCLC) present with isolated peripheral nodules without lymph node involvement and may be amenable to surgery [17].

3.2.2 Non-small cell lung cancer

Non-small cell lung cancer (NSCLC) is a term that includes a variety of different lung cancers, most notably adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. ADC is the most common type of lung cancer, and it includes one-half of all lung cancer cases [11].

It usually begins in the outer areas of the lung, in mucus-producing cells that line the small airways, the bronchioles. Adenocarcinoma tends to grow more slowly than other types of lung cancer, which can help lead to a better prognosis [18].

Squamous cell carcinoma is another type of NSCLC which is the second most common type. The different types of tumour are characterized also by a different site of origin. For example SCC usually originates at the origin of the tracheobronchial tree, but more cases are now diagnosed in the periphery of the lung [11].

Squamous cells, thin flat cells lining the surfaces of organs, are found in the lining of the bronchi. These cancers are more likely to spread to other areas of the body, making them more difficult to treat. Squamous cell carcinoma is more closely associated with smoking than any other type of NSCLC cancer [18].

Large cell carcinoma instead is a subset of NSCLC which is diagnosed by exclusion. This is a rare form, accounting for only 10 to 15 percent of all diagnoses. It is poorly differentiated and cannot be further classified by immunohistochemistry (IHC) or with electron microscopy. NSCLC also includes other subsets of lung cancer, with both heterogeneous categories and broad terminology, like adenosquamous carcinoma, sarcomatoid carcinoma, and non-small cell neuroendocrine tumours. The aetiology, thus the cause, of NSCLC can be categorized into avoidable and unavoidable risk factors. The most well-known avoidable risk factor is the use of tobacco, like for the SCLC. Other causes of lung cancer include alcohol use, environmental exposure to second-hand smoke, arsenic etc. Also the exposure to ionizing radiation has been identifies as a cause of this type of cancer. For example radiation therapy utilized for the treatment of other malignancies such as the breast cancer. In

addition, has been noticed that some pathologies can increase the risk of developing lung cancer, like human immunodeficiency virus (HIV) and pulmonary fibrosis. Clinical manifestations of non-small cell lung cancer can be divided into intrathoracic effects and extra thoracic effects. Intrathoracic effects can include cough, haemoptysis, chest pain, dyspnoea, or hoarseness. For what concern the diagnosis, it is needed to have some specific blood analysis, together with imaging. It can begin with a chest radiograph as the presentation of NSCLC can be nonspecific. If lung cancer is suspected, further evaluation with CT imaging would likely be warranted to further characterize the pathology noted on the chest radiograph. After this tissue biopsy will be needed for histopathologic and immunohistochemistry evaluation to form the diagnosis of NSCLC. Once a diagnosis is made, a CT chest and upper abdomen to include the adrenals should be ordered to assess for metastatic disease, and a positron emission tomography (PET) scan can be utilized to further stage the patient for the extent of disease and treatment [11].

3.3 Lung cancer diagnosis from computed tomography images

The lung cancer, due to the lack of symptoms at initial stage is generally detected during evaluation for an unrelated health issues or on a chest radiograph performed for preoperative evaluation [19].
Chest radiograph is the first investigation which is performed when there is a suspect case of lung cancer. Though it is a very good tool in providing preliminary information about the disease, it is inadequate for optimal characterization and staging. The CT scan of the chest instead is the cornerstone of lung cancer imaging [21].
For example the low-dose spiral CT scanning is an effective screening tool that gives detailed information. It may allow for improvement in the potential to diagnose this disease at an earlier stage [19].
The primary tumour shows a wide spectrum of imaging appearances. NSCLCs can be centrally located masses, invading the mediastinal structures as reported in Figure 3.2A, or peripherally situated lesions (Figure 3.2B) that invade the chest wall. Tumours can have margins which are smooth, lobulated (Figure 3.2C), or irregular and spiculated (Figure 3.2D). They can be uniformly solid or can have central necrosis and cavitation as reported in Figure 3.2E. Centrally situated and cavitating tumours are more likely to be of squamous histology. Sometimes the neoplasia resembles an infective pathology and is seen as an area of consolidation (Figure 3.2F), a ground-glass opacity (Figure 3.2G), or a combination of both (Figure 3.2H). Such an appearance is more commonly seen with adenocarcinoma and its subtypes. Mixed density or pure ground-glass nodules and consolidation with air bronchogram are seen in bronchoalveolar carcinomas, which are now referred to as adenocarcinoma in situ (Figure 3.2D).

Whatever the imaging appearance of the suspected lung cancer, obtaining tissue diagnosis by performing a bronchoscopy or an image-guided biopsy is necessary. When lung cancer is incidentally detected in an asymptomatic patient, it is often seen as a solitary pulmonary nodule (SPN) which can have varied imaging appearances [21].

The early detection of cancer plays crucial role in preventing cancerous cells from multiplying and spreading. In recent years the image processing algorithms are widely used in various medical areas for enhancing the earlier detection and treatment stages. However CT scan imaging is best imaging technique in the medical field, as already said, it is difficult for doctors to interpret and diagnose the cancer from CT scan images or slices, thus many studies nowadays aim to develop CAD systems [19].

CAD is helping physicians to make diagnoses, acting as second readers. This means that a physician will make his/her first attempt at diagnosing a disease in the patient and then the computer will come in as a backup to confirm that diagnosis. The main objective of CAD is to decrease the rate of false diagnosis by assisting physicians with a second opinion [20].

In the biomedical field, the examination of CT images, done by experts for lung cancer diagnosis, is a sensitive process that need time and high qualification. The subjective examination leads to variability among the observers. and this is a problem for what concern the correctness of the diagnosis. Moreover the variance of intensity in CT scan images and anatomical structure misjudgement by doctors and radiologists might cause difficulty in marking the cancerous cell. For these reasons, computer-based systems are required. From here the necessity of CAD systems which

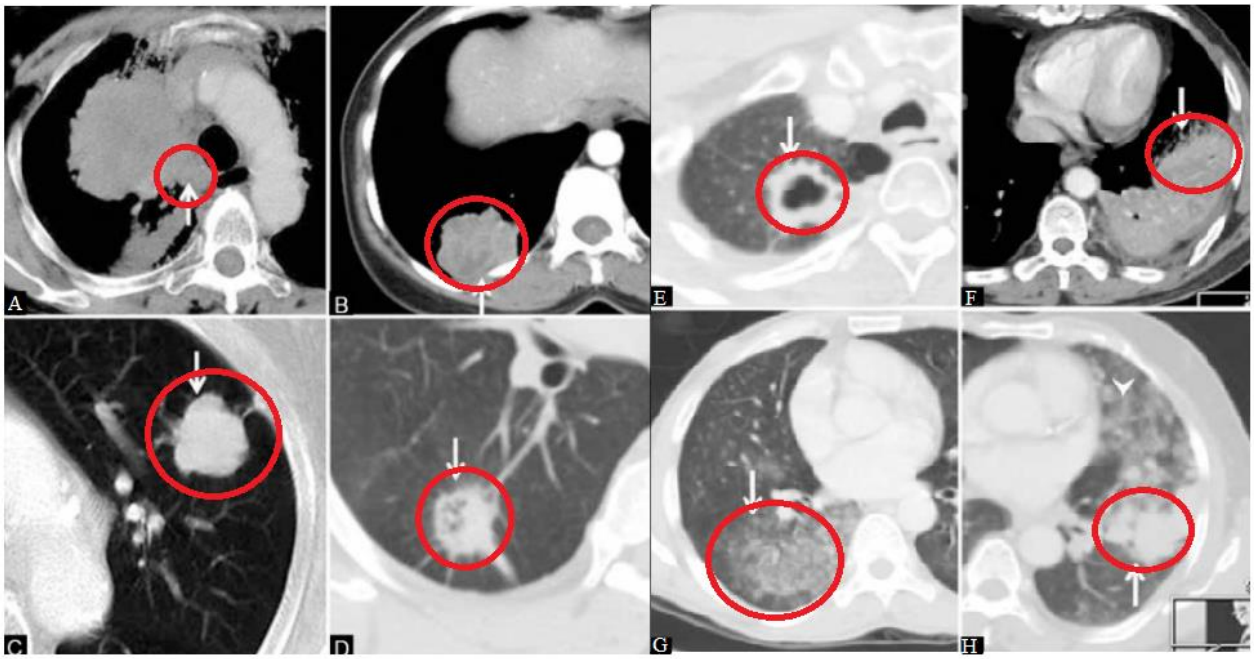

Figure 3.2. Representation of common radiological appearance of lung cancer (A,B,C,D) and atypical radiological pattern of lung cancer (E,F,G,H).

help the diagnostic process by using existing technological means and software. The main advantage is the cost and diagnosis effort reduction. In particular, image processing techniques have been proven to be efficient in detection of tumour cells. Nowadays a promising frontier in cancer detection from CT images and CT slices is represented by the artificial intelligence (AI). Several researchers has proposed and implemented detection of lung cancer using different approaches of image processing and machine learning (ML) techniques. The AI gives a device some form of human-like intelligence thus it can be exploited for helping the clinicians in diagnostic procedures. Of particular importance for this purpose is the deep learning (DL) which is a sub-field of the ML which in turn is defined as a sub-field of AI. In literature a lor of AI-based algorithms for cancer detection can be found. These are considered to be CAD systems. All of them are based on 3 main processing steps, which are essential for lung cancer detection. The first one is called pre-processing step for images preparation. The pre-processing consists of filtering, noise removal, image smoothing etc. In some cases it is also needed to apply several image enhancement techniques to improve the visual quality of an image. The aim of this step is to have an image as clear as possible for being further processed. A second stage is the image segmentation, based on some image segmentation algorithms which play an effective role in image processing. The Image segmentation is the partitioning of an image into relevant regions according to some criteria. The regions are meaningful and disjoint. Image segmentation is generally considered an intermediate step of some pattern-recognition applications. The third stage is the feature extraction which is a form of data reduction. This allows to retrieve some important 'information' on which the algorithm need to focus for classifying a suspicious area and determine if it is a tumour or not [19].

From here on will be reported some examples of CAD systems to detect the lung cancer, based on AI techniques. In the following some classifiers will be presented. These aim to identify to which category a new observation belongs to. In particular have to determine if in a CT image, given as input to the model, is present or not a lung neoplasia. All the inputs are CT images, thus it will be implemented an analysis on the 2D space, looking at the pixels. Aggarwal, Furquan and Kalra [22] proposed a model that provides classification between nodules and normal lung anatomy structure. The method extracts geometrical, statistical and grey level characteristics. The linear discriminant analysis (LDA) is used as classifier and optimal thresholding for segmentation. LDA is a method that aims to maximize the separation between 2 or more groups, in this case the CT images of nodules and CT images of normal lung anatomy. This is a supervised method because it is already known a priori to which group a data point belongs. The system has 84% accuracy, 97.14% sensitivity and 53.33% specificity. Although the system detects the cancer nodule, its accuracy is still unacceptable.

Jin, Zhang et al. used convolution neural network (CNN) as classifier in his CAD system to detect the lung cancer [22].

Neural networks are a subset of ML, and they are at the heart of deep learning algorithms. There are various types of neural networks, which are used for different purposes and data types. For example convolutional neural networks (CNNs) are more often utilized for classification and computer vision tasks [23].

The system proposed by Jin, Zhang et al. [22] has 84.6% of accuracy, 82.5% of sensitivity and 86.7% of specificity. The advantage of this model is that it uses circular filter in the region of interest (ROI) during the extraction phase. This reduces the cost of training and recognition steps. Although, implementation cost is reduced, it has still unsatisfactory accuracy. Sangamithraa and Govindaraju [22] uses K mean unsupervised learning algorithm for clustering or segmentation. Unsupervised means that the classification is not known a priori. It groups the pixel dataset according to certain characteristics. For classification this model implements back propagation network. Features like entropy, correlation and homogeneity are extracted using grey-level co-occurrence matrix (GLCM) method. The system has accuracy of about 90.7%. Image pre-processing median filter is used for noise removal in order to improve the accuracy. Roy, Sirohi, and Patle [22] developed a system to detect lung cancer nodule using fuzzy interference system and active contour model. This system uses grey transformation for image contrast enhancement. Image binarization is performed before segmentation and resulted image is segmented using active contour model. Cancer classification is performed using fuzzy inference method. Some features like area, mean, entropy, correlation, major axis length, minor axis length is extracted to train the classifier. The overall accuracy of the system is 94.12%. However it is a good method its limitation is that does not classify the cancer as benign or malignant. Ignatious and Joseph [22] developed a system using watershed segmentation. In pre-processing a Gabor filter is used to enhance the image quality. It compares the accuracy with neural fuzzy model and region growing method. Accuracy of the proposed method is 90.1% which is comparatively higher than the model with segmentation using neural fuzzy model and region growing method. The advantage of this model is that it uses marker-controlled watershed segmentation that solves over segmentation problem. As the previous model, the main limitation is that it does not classify the cancer as benign or malignant. The accuracy is high but still not satisfactory. Despite of this the last method is considered to be the current best solution thus will be presented in detail in the following [22].

This system includes 5 main processing steps: pre-processing, segmentation, feature extraction, tumour detection and tumour stage identification. In particular there are mainly 6 stages:

- Step 1: Read in the CT image single slice of the patient.

- Step 2: Pre-process the CT image slice.
- Step3: Segment the pre-processed image using Marker Controlled Watershed segmentation.
- Step 4: Convert the segmented image into a binary image.
- Step 5: Extract the features from the binary image.
- Step 6: With the extracted features, identify the stage of the cancer.

Pre-processing is an essential step for many of the image processing applications. The technique used here is sharpening of the CT image. This enhances the finer details within the image [6].

The sharpening method uses a convolution method for contrast enhancement. The image obtained after the pre-processing step is suitable for further processing. The subsequent step, which is the image segmentation process, partitions the image into multiple segments and this is of fundamental importance in the medical field. It is used to locate objects and boundaries within the image. There are different segmentation methods available. In this work has been used the Watershed segmentation, also called Marker Controlled Segmentation. This segmentation is performed following multiples steps: Step 1: Read the colour image and convert it to grey scale image. Step 2: Compute the Gradient Magnitude as the segmentation function. Step 3: Mark the foreground objects within the image. Step 4: Find out the background marker points within the image. Step 5: Find out the watershed transform of the segmented function of the image. Step 6: Resultant segmented binary image is obtained. The step which then follows the segmentation is the feature extraction. This is an essential step for image analysis. This step determines the relevant information taken into account for the processing. Thus all that information that can be useful for identifying any abnormality within the lung. In fact the extracted features are used for detection and staging of the tumour. The different features that are extracting here are Area, Perimeter, Eccentricity, Convex Area and Mean Intensity. All these features are scalar values. The area is the summation of all the pixels within the tumour portion of lung. Perimeter is the summation of the outline of interconnected tumour portion within the lung. This is used for determining the roundness of the tumour. The value of eccentricity will be equal to 1 for a regular object and it is greater than or less than 1 for an irregular shape. The convex hull specifies the smallest convex polygon which encloses the tumour portion within the lung. The number of pixels within this convex polygon will give the convex area. The mean intensity indicates the average intensity of pixels within a particular region of interest, which is the tumour nodule. Finally the proposed model is able to identify all the probable tumours regions within the lungs with the features obtained from the experimental analysis. Finally the classifier, is able to accurately detect the stage of the tumour. The classifiers used for this study are Support Vector Machine (SVM), Naive Bayes

Multinominal classifier (NBM), Naive Bayes Tree (NB tree) and Random tree. SVM is a supervised learning algorithm for classification [23].

The objective of this algorithm is to find a hyperplane, also called decision boundary, in an N-dimensional space (indicating with N the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The aim is to find a plane that has the maximum margin, thus the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Data points falling on either side of the hyperplane can be attributed to different classes, so can be classified. The dimension of the hyperplane depends on the number of features considered. In this method a margin has been set in such a way that include the maximum support vector. The maximum support vectors are data points that are closer to the hyperplane and influence position and orientation of the hyperplane itself [24].

NBM and NB tree are the variations of Naive Bayes classifier. This classifier uses a probabilistic model for classification. During training, the probability of occurrence of each data value in a particular class is calculated. When a new data set is given for testing, the probability is calculated. The minimum probability difference class is selected and thus the class is predicted. The Random tree is another classifier used for a stochastic process [23].

A decision tree in general makes a statement which is the starting point and then makes a decision based on if the statement is true or not. If the model classify things into categories is named decision tree otherwise it is called is called regression tree when predict numeric values . Each tree is composed by a top node called root node. Then there are internal nodes (also called branches) which have entering arrows and exiting arrows. Finally there are the leaves which are nodes with no arrows exiting from them. These classification trees are used to predict if a CT image display the presence of a tumour or not. A decision tree splits the dataset recursively using the decision nodes, it finds the best fit by maximizing the entropy gain. If a data sample satisfies the condition at the decision node than it goes to a child node, otherwise reaches a leaf node where a class label is assigned to it. With this classifier there is a problem of overfitting thus to solve this issue, the root node for the random tree is selected randomly. During training, random instances are selected from the datasets. With the selected instances the decision trees are created. The randomness is achieved by using 2 random processes bootstrapping and random feature selection. Bootstrapping ensure that the data used for every tree are not the same; this helps the model to be less sensitive to the original training data. The random feature selection instead helps to reduce the correlation between the trees. If the features used are the same, the decision trees will have the same decision nodes and they will act very similarly. This causes an increase in the variance. Another problem is that the decision trees are highly sensitive

to training data that can result in high variance. So the model may fail to generalize. From here the need of using the random forest algorithm. It is a collection of multiple random decision trees and it's much less sensitive to the training data. It is a random sampling with replacement. Then train each decision tree on each dataset independently. For training the trees only some features are used thus a subset of features for each tree is randomly selected and used for training. During testing phase, classifier takes the input feature vector. This vector is taken by every tree in the forest and thus the classification process is done. The output will be class label of majority votes. This computer aided lung cancer detection system has been trained with 200 slices of CT image. The proposed system can detect and stage the tumour. For the staging of the tumour all the classifiers reported above have been tried. The best one is the random tree. The accuracy rate of random tree classifier is 94.4%. The drawback of this CAD is that the algorithm is not able to detect lung on the borders of the lungs and the accuracy can be enhanced by using many other features [23].

3.4 Lung cancer diagnosis from computed tomography scans

In the previous section have been reported and described some CAD systems to classify lung cancer from 2D CT images. The accuracy of these systems is quite good but one of the main limitation is that the cancer can be seen and analysed only looking at the bidimensional space, so some important information is lost. The cancer has a certain depth which cannot be "captured" by analysing 2D images. Thus a possible way to overcome this problem is by using the CT slices which allow to analyse the cancer in the 3D space. Working with CT scans means analysing voxels and not anymore pixels. From a practical point of view, this means that the model takes as input a series of CT slices. A lung cancer classifier as much accurate as possible could speed up and reduce costs of lung cancer screening, allowing for more widespread early detection and improved survival. The goal is to construct a CAD system that takes as input patient chest CT scans and outputs whether or not the patient has lung cancer. In literature many proposals can be found. There is a lot of research in the area of medical imaging using deep learning and in particular CNNs with very promising results. Xu et al. presented the effectiveness of using deep neural networks (DNNs) for feature extraction. Kumar et al. proposed a CAD system which uses deep features extracted from an autoencoder to classify lung nodules as either malignant or benign. Suna W. et al. [25], implemented three different deep learning algorithms, CNN, Deep Belief Networks (DBNs) and Stacked Denoising Autoencoder (SDAE), and compared them with the traditional image feature-based CAD system. The CNN architecture contains eight layers of convolutional and pooling layers, interchangeably [25].

The pooling layers usually are placed after the convolutional layers. Their function is to reduce the spatial size of the representation by reducing the amount of parameters and computations in the network. It operates on each feature map (channel) independently [26].

About 35 features have been extracted, in particular texture and morphological features. These features were fed to the kernel-based support vector machine (SVM) for training and classification. The resulted accuracy for the CNN approach reached 0.7976 which was little higher than the traditional SVM, with 0.7940. J. Tan et al. designed a framework that detected lung nodules by reducing the false positive. It is based on Deep neural network (DNN) and Convolutional Neural Network. The CNN has four convolutional layers and four pooling layers. The resulted sensitivity was of 0.82. The False positive reduction using DNN was of about 32.9%. R. Golan proposed a framework that train the weights of the CNN by a back propagation to detect lung nodules in the CT image sub-volumes. This system achieved sensitivity of 78.9% with 20 false positives, while 71.2% with 10 false positives per scan, on lung nodules that have been annotated by four radiologists. Convolutional neural networks have achieved better than Deep Belief Networks in current studies on benchmark computer vision datasets. The CNNs have attracted considerable interest in recent years since they have a strong ability in learning useful features from input data. Another possible method has been proposed by Wafaa Alakwaa, Mohammad Nassef and Amr Badr [25]. For this binary classification problem have been used a 3D convolutional neural networks, to build an accurate classifier. In order to determine whether or not a patient has early-stage cancer, the CAD system here presented, would have to detect the presence of a tiny nodule (< 10 mm in diameter for early-stage cancers) from a large 3D lung CT scan (typically around 200 mm × 400 mm × 400 mm). An extensive pre-processing technique is applied to enhance the accuracy of detection. Moreover, it is necessary an end-to-end training of convolutional neural networks from scratch to realize the full potential of the neural network i.e. to learn discriminative features. Extensive experimental evaluations have been performed on a dataset comprising lung nodules from more than 1390 low dose CT scans. These scans have been divided into 2 parts, the training set and the test set. This CAD system for lung cancer detection follow the pipeline reported in the Figure 3.3. it consists of image pre-processing, detection of cancerous nodule candidates, nodule candidate false positive reduction, malignancy prediction for each nodule candidate and malignancy prediction for overall CT scan. There are many phases, each of which is computationally expensive and requires well-labelled data during training. For example, the false positive reduction phase requires a dataset of labelled nodules. The proposed CAD system starts with pre-processing of the 3D CT scans using segmentation, normalization, down-sampling, and zero-cantering. The initial approach was to simply input the pre-processed 3D CT scans into 3D CNNs, but the results were poor. So an additional pre-processing was performed to input only regions

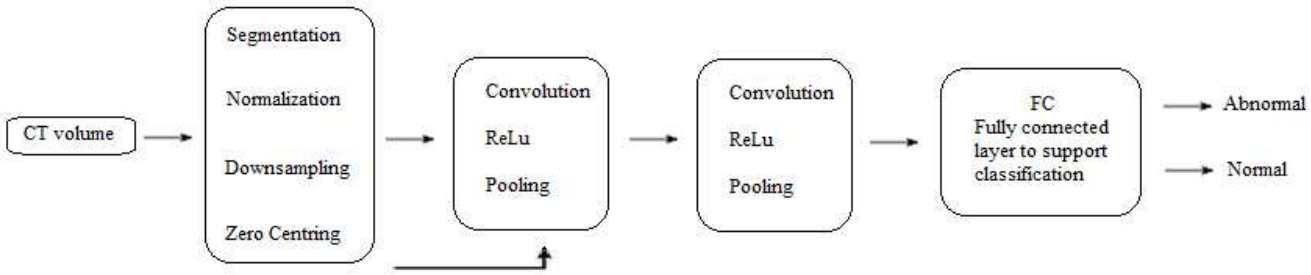

Figure 3.3 Architecture of convolutional neural network.

of interests into the 3D CNNs. To identify regions of interest, a U-Net (CNN developed for biomedical image segmentation), was trained for nodule candidate detection. Then input regions around nodule candidates were fed into 3D CNNs to ultimately classify the CT scans as positive or negative for lung cancer. For each patient, pixel values of each CT were first converted to Hounsfield units, a measurement of radiodensity, and 2D slices are stacked into a single 3D image. The segmentation step is then used to mask out the bone, air, and other substances that would make data noisy, and leave only lung tissue information for the classifier. A number of segmentation approaches were tried, including thresholding, clustering (k-means and Meanshift), and Watershed. Watershed produced the best qualitative results, but took too long to run [25].

The Watershed algorithm for segmentation extracts sure background and foreground and then using markers allow to detect the exact boundaries. This algorithm generally helps in detecting touching and overlapping objects in image. The markers can be defined manually by the users or can be retrieved with some algorithms such as thresholding or any other morphological operation [27].

After segmentation, the 3D image is normalized by applying the linear scaling to squeeze all pixels of the original unsegmented image to values between 0 and 1. Spline interpolation down-samples each 3D image by a scale of 0.5 in each of the three dimensions. Finally, zero-centring is performed on data by subtracting the mean of all the images from the training set. In the Figure 3.4 are reported the 3D images reconstructed by applying a thresholding on the pixels based on the radiodensity of various parts of the CT, thus the HU. The air is typically around -1000 HU, lung tissue -500 HU, water, blood and other tissues instead are around 0 HU. Bones have a value around 700 HU. Thus all the pixels close to -1000 or above -320 are masked out to leave lung tissue as the only segment. The segmentation obtained from thresholding has a lot of noise. The classifier is not able to correctly classify images in which cancerous nodules are located at the edge of the lungs. This is because many voxels that were part of lung tissue, especially the ones at the edge of the lung, tended to fall outside the range of lung tissue radiodensity due to CT scan noise. Thus to filter noise and include voxels from the edges, a Marker-driven watershed segmentation has been used. Feeding the entire segmented

lungs into malignancy classifiers made results very poor. Thus the first solution proposed, which consists of feeding the entire segmented 3D image to the model, is not producing good result. Thus it has been tried to feed the model with smaller region of interests. This was achieved by selecting small boxes containing top cancerous nodule candidates. To find these top nodule candidates, a modified version of the U-Net was trained. Here has been used a new version of the U-Net designed to limit memory expense. During training, the modified U-Net takes as input 256 × 256 2D CT slices, and labels are provided (256 × 256 mask where nodule pixels are 1, rest are 0). The model is trained to output images of shape 256 × 256 were each pixels of the output has a value between 0 and 1 indicating the probability the pixel belongs to a nodule. Then the trained U-net is applied to the segmented CT scan slices to find nodule candidates. Ideally the output of U-Net would give the exact locations of all the nodules, and it would be able to declare if a cancer is present or not in a image. However it produces a lot of false positives, so is needed an additional classifier that determines the malignancy. For this reason it is used a 3D CNN as linear classifier. A thousand low-dose CT images from high-risk patients in Digital Imaging and Communications in Medicine (DICOM) format have been used. The database used consists of 1397 CT scans and 248580 slices. Each scan contains a series with multiple axial slices of the chest cavity. The number of slices in each scan varies based on the machine taking the scan and patient. For simplicity in training and testing we selected the ratings

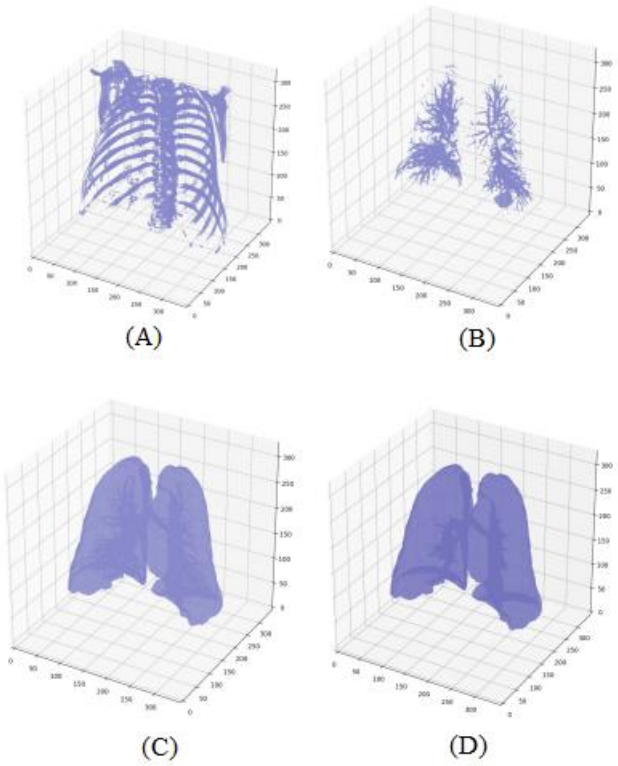

Figure 3.4: (A) 3D image representing the bone segment. (B) Bronchioles within lung. (C) Initial mask with no air. (D) Final mask with bronchioles included.

of a single radiologist. All experiments were done using 50% training set, 20% validation set and 30% testing set. The accuracy of this proposed model is 86.6%, mis-classification rate is 13.4%, false positive rate is 11.9% and false Negative is 14.7%. Almost all patients are classified correctly. The results demonstrate the efficiency of this 3D CNN to detect cancerous nodules by working on the suspicious areas retrieved with the modified U-Net architecture [25].

4 Computer-aided diagnostic systems

In the simplest terms, AI refers to systems or machines that mimic human intelligence to perform tasks. These machines, just like humans, can learn from the experience by iteratively improving themselves with the information collected. Machine learning is a subset of AI and in turn deep learning is a subset of ML. AI and machine learning, especially deep learning applications, have become very popular in everyday life. Between all fields of application of AI, the most promising one is the application in medicine [28].

AI in medicine is the use of machine learning models to search medical data and uncover insights to help improve health outcomes and patient experiences. Thanks to recent advances in computer science and informatics, AI is quickly becoming an integral part of modern healthcare. AI algorithms and other applications powered by AI are being used to support medical professionals in clinical settings and in ongoing research. Currently, the most common roles for AI in medical settings are clinical decision support and imaging analysis. Clinical decision support tools help providers make decisions about treatments, medications, mental health and other patient needs by providing them with quick access to information or research relevant to their patient. In medical imaging, AI tools are being used to analyse CT scans, x-rays, MRIs and other images for lesions or other findings that a human radiologist might miss [29].

Despite the promising nature of the AI in medical imaging, there are still a lot of regulatory and ethical issues. A proof of this is the fact that many encouraging ML and DL methods present in literature have not been approved by the food and drug administration (FDA) yet. Thus the translation to the clinic is not possible. Especially in medical imaging field, there are a lot of stakes since a wrong suggestion by AI application could cause serious problems. Despite all these obstacle, the conventional computer-aided detection systems, which is routinary used in clinics from decades, has helped pave the way for the AI tools. The main difference between CAD systems and AI tools is that the latter one exploits features extracted automatically by a deep neural network avoiding the need to have features extracted “by hand”. For example the deep learning models can automatically learn meaningful representation of the data, thereby eliminating the need of modelled imaging models and hand-crafted features. In spite of all these advantages, there are some drawbacks too. The large amount of data required to train AI systems, a much higher computational cost and a reduced interpretability. The interpretability is a concept of fundamental importance in medical field. It represents the degree to which an observer can understand the cause of a decision. The interpretability in AI models is almost lost thus these are usually discussed in terms of justification which refers as to the accuracy of the model itself. Justification and interpretability are inversely related factors. The

more accurate and advanced the model is and the less interpretable it is. It is like a black box in which the input and the output are known but what is in between is not known [30].

ML is a subset of AI and refers to all that algorithms that can "think" and provide an answer or take a decision. ML algorithms apply statistical methodologies to identify patterns and take a decisions on their own. The ML algorithms can be divided in 3 classes: supervised, unsupervised and reinforcement learning. A supervised ML algorithm has in input data that have to be analysed but also the correct results that are expected as output. The algorithm will train itself by its own, trying to give as output what has been previously given in input. It is like the algorithm create some internal rules in an autonomic way. The unsupervised algorithm instead is characterized by the fact that no information is given a priori. The algorithm will analyse the data trying to find some recurrent patterns in the data, giving as output some useful information. The reinforcement learning instead is based on algorithms able to learn something from its own errors. In this case the algorithm do not get any data as input. The strength of the ML is that the more "they work" and the better are the obtained results [31].

ML has many different tasks like data pre-processing, feature engineering, training machine learning models of regression, classification and clustering, dimensionality reduction, testing and matching and many others. These tasks can be resolved by many different machine learning methods. The data processing as already explained, is a fundamental step for preparing the data before starting the training of the model. For solving this tasks, 2 methods are available: data cleaning and missing data imputation. Data cleaning is a part of data pre-processing. It consists of detecting and correcting incomplete, inaccurate, incorrect or irrelevant data inside a dataset and replacing, modifying or deleting them. Missing data imputation instead aims to handle missing data using data imputation techniques which consist of replacing missing data with mean, median or mode. Feature engineering is another critical task, used when building machine learning models. Selecting the right features helps in reaching an high accuracy and allow also to reduce the complexity of the constructed model, limiting the overfitting problem [32].

This, together with the underfitting, is one of the main problems that affect the ML models. This is a consequence of an inefficient data cleaning. The impurities of the dataset may affect the accuracy and the performances of the model. The overfitting problem is present when the model performs well on the seen dataset but performs badly on unseen data. This can be detected once test the data [33].

Feature engineering includes some other tasks such as deriving features from raw features, identifying important features, feature extraction and feature selection. There are many techniques which could be used for feature selection: wrapper method, regularization techniques and filter method. The wrapper methods help in feature selection by using a subset of features and determining

the model accuracy. Some of the algorithms used by the wrapper method are forward selection, backward elimination and recursive feature elimination. Regularization techniques instead penalize one or more features appropriately to let the most important features come up. The algorithm used for this technique are LASSO (L1) regularization, Ridge (L2) regularization, elastic net regularization and regularization with classification algorithms such as logistic regression, SVM etc. The filter method helps in selecting features based on the outcomes of statistical tests like Pearson's correlation, linear discriminant analysis (LDA), analysis of variance (ANOVA) and chi-square test. As reported in the section 3.3, LDA is used as classifier and optimal thresholding for segmentation. It is a supervised method that aims to maximize the separation between 2 or more groups. It consists of combining variables by using optimal values of the weights. Starts from a dataset containing for example CT of patience with and without lung cancer. Take two features like dimension of a suspicious nodulus and mean shade of the area. Plot these two as can be seen in Figure 4.1. Firstly by putting on x axis CT with tumour and CT without tumour and on y axis the first variable, size of the nodulus (Figure 4.1(A)). Then in another plot use the same x axis but change the y axis, putting on it the second variable, mean shade of the area (Figure 4.1(B)). The aim is to find a line which best separates these 2 groups in both plots. In this case there is no line to separate the 2 groups. Then in another plot put on the x axis the first feature and on the y axis the second feature (Figure 4.1(C)). Now there is a line that can separate the 2 groups. Rotate the data into a new dimension to look at the discriminant functions (Figure 4.1(D)). Here is discriminant function 1 (D1) and 2 (D2). One of the two discriminant function can be used to separate well the two groups. Then delete the discriminant function which is useless. Using LDA it is possible to plot the data in function of the discriminant scores and place a line that divide the two groups. The discriminant score for this example is calculated through the formula (5):

$$LB = a_1 * x_1 + a_2 * x_2 \tag{5}$$

Where x1 and x2 are the variables associated to the features and a_1 and a_2 are the weights. The weights must give the optimal separation between the 2 groups. These can be retrieved by a statistical software considering the standardized data with a mean of 0 and a deviation of 1. To get a good

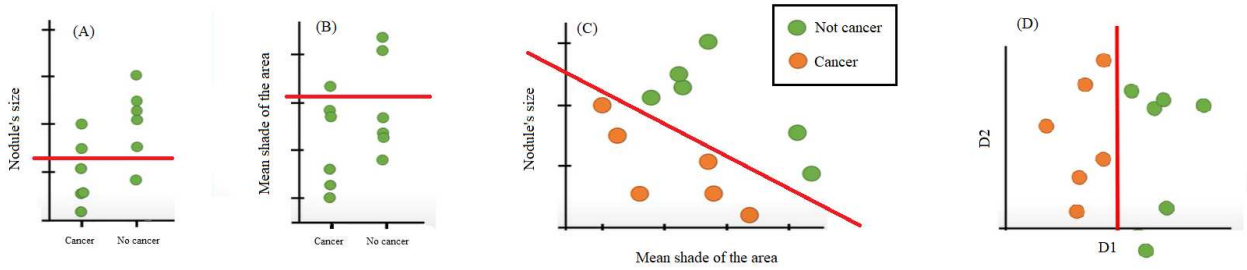

Figure 4.1 Steps of linear discriminant analysis.

separation between the 2 groups it is needed that the 2 group means are far away one from each other so the between group variance should be large. In addition the within group variance, thus the variance of all the data inside each group, should be small. The ratio between the between and within group variance, should be as large as possible to have a clear separation. Again the importance of the data processing phase; the aim is to transform data so that the between group variance is increased and the withing group variance decreased. Another important task of ML is training models, with data related to previously extracted features. Thus is an important step for improving the performances of the model in solving a determined problem. There are many different types of machine learning problems and related algorithms to solve them. A first problem is called regression; it deals with the estimation of numerical values (continuous variables). For solving the regression problem there are many ML methods available like linear regression, regression trees, random forests and Kernel regression (characterized by an higher accuracy). Then there is the classification problem, related to predicting a category of data (discrete variables). A typical example of this in medicine is predicting whether or not in a CT image is present a cancer. To solve the classification tasks may be applied some ML methods such as Kernel discriminant analysis (high accuracy), artificial neural networks (ANN), support vector machine, random forests (high accuracy) etc. Another problem is called clustering which consists of finding natural grouping of data and label associated with each of these identified clusters. To solve this can be used the K-means, topic models, hierarchical clustering and mean-shift (higher accuracy). Another task defined as feature extraction is the dimensionality reduction. It is the process of reducing the number of random variables under consideration and can be subdivided into feature selection and feature extraction. Some ML methods for dimensionality reduction are: independent component analysis, non-negative matrix factorization, gaussian graphical model, principal component analysis etc [32].

So after having defined the problem, thus the task of the ML algorithm, it is necessary to construct the ML model. As reported above, for each task there are many possible ML algorithms to implement. In order to define which is the best solution among all the possibilities, it is used the cross validation. This compare all the results obtained as output from all the available algorithms and identify which is the best one. After this the model must be trained and tested. Generally the 75% of the data available are used for training and only the 25% for testing. After the training, it is necessary to see how well the model is performing through the accuracy function. The problem linked to the accuracy function is that it does not provide insights on how to improve the performances of the model. Thus it is needed a correctional function that allows to compute when the model is the most accurate. From here the necessity of introducing the cost function. This is used to measure how wrong the model is in finding a relation between the input and the output. As previously said the strength of the AI and thus of ML

is that it is able to learn from experience. For this reason a wrong output of the model act as cost function which helps to improve its performances. For example look at how to retrieve the cost function for linear regression. A linear regression model uses a straight line to fit the model. This is done using the simple equation of a straight line (6):

$$y = mx + q \tag{6}$$

Where y is the output, x is the input, m is the slope of the line and q is the intercept. In this equation m and q are 2 changeable variables. If these variables are not properly optimized, the obtained line will not fit properly the model. The error associated to the prediction of the data will be high. Thus it is necessary to optimize the values of the variables in order to get the best fit. The cost function (Eq. (7)) for this linear regression model is the minimum of the root mean squared error of the model, obtained by subtracting the predictive values from the actual values.

$$J = \frac{1}{n}\sum_{i=0}^{n}(h_\theta(x_i) - y_i)^2 \tag{7}$$

Where J is the cost function, n is the number of data, $h_\theta(x_i)$ are the predicted values (calculated through the function h that has as input also the labels associated to the data θ) and y the actual values. Then the aim is to find those values for which the cost function is minimum. There are a lot of optimization algorithms for this purpose and one of this is called gradient descendent algorithm. It aims to compute the errors and minimize it. The gradient descendent function is represented by the formula (8):

$$\theta_J = \theta_J - \eta\frac{\partial J}{\partial \theta} = \theta_J - \eta\nabla J \tag{8}$$

Where θ_J is the gradient descendent value, η is the learning rate which defines how fast move down the slope and ∇J is the previous gradient descendent. It is necessary to choose an appropriate learning rate, not too high otherwise the least error could be missed and not too small otherwise it would take a very long time, wasting computational power. At the end, after the training, a testing phase is used to see the ability of the model to perform on unobserved inputs (test set). This property is called generalization. The problems that can arise from a lack of generalization is the underfitting and the overfitting. Underfitting refers to a model that can neither model the training data nor generalized the new data. This is caused by the fact that the epochs (number of iterations of the algorithm) or the number of neurons (in the case of deep learning), are not enough. On the contrary the overfitting refers to a model that models the training data too well, so when new data (testing data) are presented it is not able to handle them. It is like the model has focused only on the training data and not on the problem. The problem of overfitting is linked to a number of epochs and neurons too high. Consequently both the generalization error and the training error, must be limited as much as possible.

But this is not easy because as can be seen in the Figure 4.2, the training error always decreases, instead the generalization error first decreases but then increases again. Consequently to avoid as much as possible the generalization error it can be helpful the validation of the model. This allows to find the optimal capacity, thus the optimal number of epochs or neurons needed to have the minimum generalization and training error.

During the years there has been a shift from systems completely designed by humans to systems trained by computers using data from which feature vectors are extracted. As already said, with machine learning the extraction of discriminant features from the images (in the case of AI models for image analysis), has to be done by humans. Thus these systems are said to work with hand crafted features. A logical next step is to let the computers learn the features that optimally represents the data for the problem by its own. This is the fundamental concept at the basis of the deep learning. The DL consists of models (networks) composed of many layers that transform input data, like images, into an output (labels, disease present or not) [34] .

The simple machine learning algorithms work very well on a wide variety of important problems. However, they have not succeeded in solving the central problems in AI, such as recognizing speech or recognizing objects (especially in image analysis). The development of deep learning was motivated in part by the failure of traditional algorithms to generalize well on such AI tasks. In addition the challenge of generalizing to new examples becomes exponentially more difficult when working with high-dimensional data such as CT images and scans. The mechanisms used to achieve generalization in traditional machine learning are insufficient to learn complicated functions in such high-dimensional spaces. Moreover this also often impose high computational costs. Thus the deep

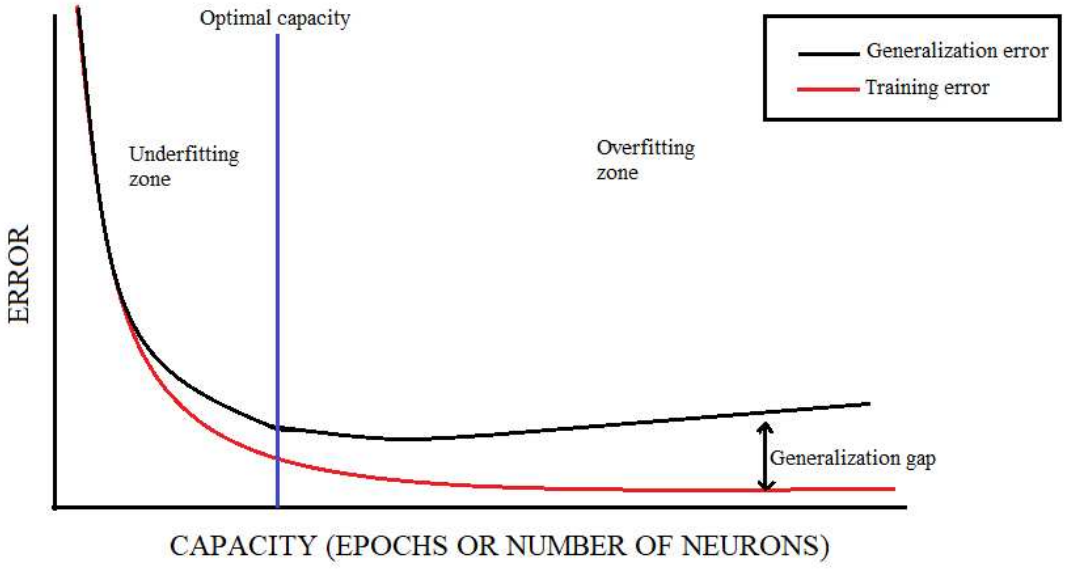

Figure 4.2 Trend of generalization error and training error.

learning was designed to overcome these and other obstacles. The DL can be defined as a technique of the ML. The neural networks, which mimic the human brain, are the basis of the deep learning. There is a parallel between the basic components of these networks (artificial neuron) and the ones of the human brain (neuron) as can be seen in the Figure 4.3. The human neuron is composed by a cell body, called soma, in which all the information is elaborated. The information enters the neuron through the dendrite and exit from there passing to another neuron through the axon. The connections between all these neurons are called synapsis. All the information flow in a unique direction. The deep feedforward networks, also called feedforward neural networks, are the quintessential deep learning models. The multilayer perceptron (MLP) is the most well-known of the traditional neural networks, characterized by several layers. The neural network is a concept, not a machine. The architecture of these networks resembles the brain structure and in particular its biologic neural network. The biological neurons are grouped in various layers and transmit signals. These signals contain information of various nature which allow a person, for example, to identify an image. These biological neural networks store the previous experiences, learn and update constantly the knowledge and the comprehension of the environment. The basic concept for the AI neural networks is the same of the biological networks but it is composed by artificial neurons containing algorithms. In the neural network multiple algorithms work together to process the given input and generate an output.

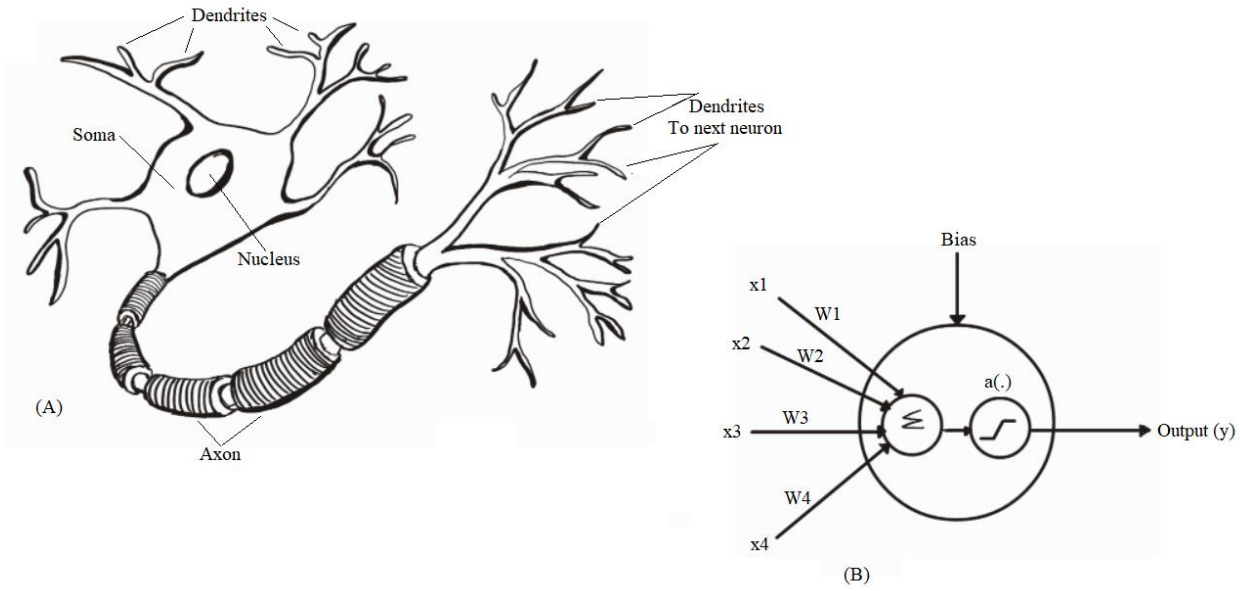

Figure 4.3 Parallel representation of human neuron (A) and artificial neuron (B).

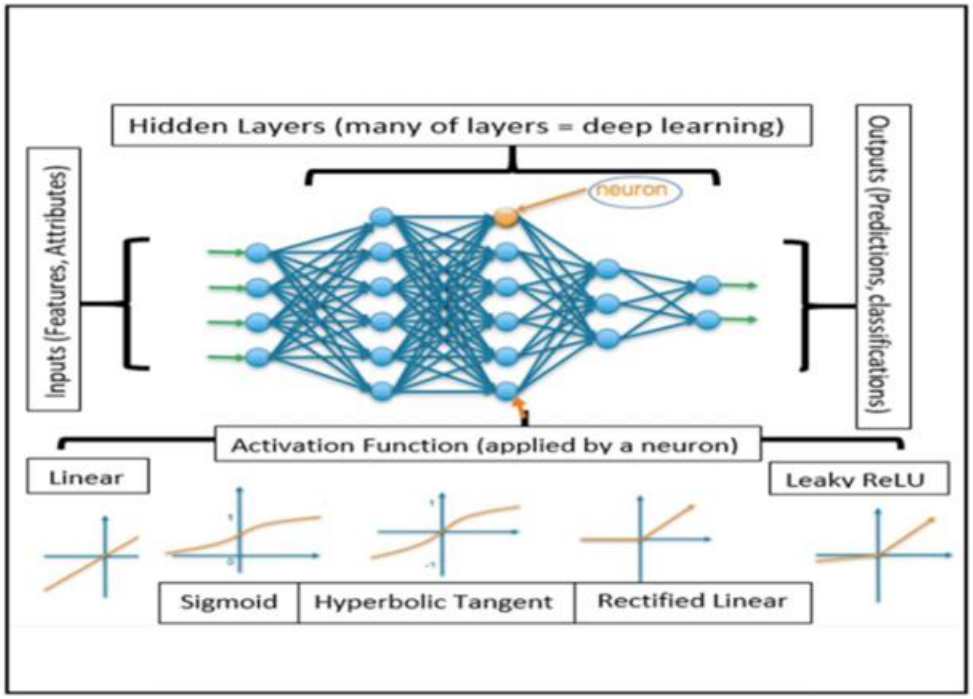

Figure 4.4 Basic structure of Feedforward neural network.

The algorithms allow to relate the given input data to training data with no prior programming of these algorithm. This is the main improvement of the DL with respect to the ML; it is able to perform its tasks without human help. These outputs just like the experience for the humans, can help the neural network to learn and improve its performance. Neural networks, as can be seen in the Figure 4.4, are composed by multiple "neurons", organized in multiple layers as reported in the Figure 4.3. Each progressive layer utilizes the output from the past layer as information. Feedforward network consist of the following layers:

- Input layer: it contains the neurons which receive the input. Then they pass the input to the next layer. The total number of neurons in the input layer is equal to the attributes in the dataset.
- Hidden layer: this is the middle layer. It is hidden between the input and the output layers. The number of neurons here is huge and aim to transform the input.
- Output layer: it is the last layer and is dependent upon the built of the model.

When a neural network contains multiple hidden layers, it is considered to be a deep neural network. Between each 2 neurons, there is a connection like the synapsis in the human brain. Each of this connection has a certain weight which ranges from 0 to 1. The goal of a feedforward network is to approximate some function f^*. For example, a classifier maps an input x to a category y through the function f^* as reported in the equation (9).

$$y = f^*(x) \tag{9}$$

A feedforward network is able to define a mapping (equation 10) and from here learn the values of the parameters θ which provides the best function approximation.

$$y = f^*(x, \theta) \tag{10}$$

As ML models, the algorithm can be supervised, unsupervised or even sei-unsupervised. Whit the deep learning it is the machine itself that choose and define the classifiers to be used. These classifiers are not chosen a priori from the human but they seem to have much better results. These networks are called feedforward because, as in the brain, the information can travel only in the forward direction. If feedforward neural networks are extended to include feedback connections, these takes the name of recurrent neural networks (RNN). They are the basis for object recognition in images, like suspicious nodules in CT images or scans [35].

Each artificial neuron need an activation function, which transform the input and the weight (from which is calculated the weighted sum of the inputs) into an output as reported in Figure 4.3. Thus the activation function is used to map the input of the neuron into the output. This allows the neural network to understand relations and complex schemes in the data. This activation function is fundamental to maintain the output of the neuron limited within a particular interval. Another advantage of this function is the fact that it introduces non-linearity in the data. In fact the activation functions are always non-linear except in the last layer. The non-linearity in the neural networks is important to form an approximator of universal function. It allows to learn any continuous function. If the activation function is linear, the network will become a simple model of linear regression. In this case it is like reducing the entire neural network to one neuron only; this will not be able to learn complex relationships in the data. In addition choosing the correct activation function is of fundamental importance to avoid the problems related to the updating of the weights. This problem may be present in those networks in which it is implemented a feedback cycle to continuously update the neural networks. In the neural networks during the posterior propagation, each weight is updated proportional to the partial derivative of the cost function which represent the error. These updates are obtained by multiplying various partial derivatives. If these derivatives are too small, the update is near to the 0. In this case the weights will be not able to update and the convergence will be very slow or absent. This problem is known as escape gradient. At the same way the updates must be not too high otherwise the algorithm would overcome the minimum and this is called explosive gradient. There are many different activation functions available, represented together with the formula and the pros and cons in the Table I. The rectified linear unit (ReLu) for example are more complex activation functions that are used for deeper networks. These ones, contrarily to the classical activation functions like the unit step and the sign, are convex and thus have a larger area with non-zero derivative [36].

Neural networks, as every ML model, need to be trained and tested with specific training set and test set. Then it is needed to evaluate the performance of the network for example in predicting a parameter set θ. This is done through the loss function $L(\theta)$ or the cost function $J(\theta)$. The loss function measures how well the model performs on a single training example. The cost function instead can be defined as the average of the loss functions for the entire training set. It measures the average error for the entire training set. As already said the neural network is able to learn from the experience and in particular from its own errors. This is possible thanks to the back propagation algorithm, which propagates the errors from the output layer backward and updates the variables layer by layer. Actually back propagation is how neural networks learn. The aim of the backpropagation is to minimize the error function of the model, thus the loss function or the cost function. This is possible by optimizing these values (weights and biases) with gradient descent based optimization algorithms. In fact with the optimization algorithms, it is possible to find the least minimum error value; the value for which the predicted output is as close as possible to the actual output. The type of loss function that need to be used, depends on the type of problem. For regression problems can be used the mean squared error loss or the mean absolute error. For binary classification problems can be used the binary cross entropy, the hinge loss or the squared hinge loss. Instead for multi class classification loss functions can be used the multi-class cross entropy loss or the sparse multiclass cross-entropy loss. The graph of the cost function for a neural network is characterized by the presence of multiple local minima. The aim is to find the global minimum in the graph and it can be done through many available optimization algorithms. One possible algorithm is the gradient descendent algorithm which has been already presented. This algorithm, as previously said, consists on deriving the gradient descendent by differentiating the cost function. For example the Equation (11), represents the binary cross-entropy loss function. This is used in binary classification tasks which have to answer a question

$$J = -\frac{1}{n}\sum_{i=0}^{n}[y_i * \log\left(\sum_{j=1}^{m} w_j x_j + b\right) + (1 - y_i) * log\left(1 - \sum_{j=1}^{m} w_j x_j + b\right)] \qquad (11)$$

with only two choices, like disease present or not. Where n is the output size, i is the index associated to the neurons inside that layer and j is the index associated to the position of each element inside the input vector $x = [x_1, \ldots, x_m]$. y_i is the real output expected by the neuron i, x_j is the input given to the neuron i and w_j is the weight associated to the input j. b is the bias which is a constant number. The bias is used to delay the triggering of activation function. Controlling this value helps in defining the value at which the function will trigger. All together $(w_j x_j + b)$, represents the output generated by the neuron j. Actually w_j is the weight of the layer (average of all weights of the neurons in that layer). It is like doing the linear regression model across all the different neural nodes in the different

Table I. Activation function (funct.) with their relative formulas, pros and cons.

Activation funct.	**Formula**	**Pros**	**Cons**
Sigmoid	$S(x) = \frac{e^x}{e^x + 1}$	-Continuous and differentiable. -Limit the input between 0 and 1. -Clear prediction for binary classification.	-Escape gradient. -Not centred around 0. -Computationally costly
Softmax	$\sigma(z) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$ With: i=1,….,K $z=(z_1,\ldots, z_K)$	-Used for the multi-class classification at output layer.	-Computationally costly.
Hyperbolic tangent	$\tanh x = \frac{\sinh x}{\cosh x}$	-Continuous and differentiable everywhere. -Centred around the 0. -The output is limited between -1 and +1.	-Escape gradient. -Not centred around 0. -Computationally costly
ReLU (Linear rectified unit)	$f(x) = x^+ =$ $= \max(0, x =$	-Easy to calculate. -Fast and efficient.	-Gradient explosion. -Not centred in 0. -Can definitely deactivate some neurons.
Leaky ReLU	$f(x) =$ $= \begin{cases} x \ if \ x > 0 \\ 0.01x \ otherwise \end{cases}$	-Easy to calculate. -Do not deactivate any neuron.	-Gradient explosion. -Not centred in 0.
ELU (Exponent linear unit).	$f(x) =$ $= \begin{cases} x \ if \ x > 0 \\ a(e^x - 1) \ otherwise \end{cases}$	-Do not deactivate any neuron	-Computationally costly. -Gradient explosion. -The value a must be decided.

layers. Then the gradient descent of the neural network is given by the partial derivative of the cost function with respect to the predicted parameters vector of that specific layer like reported in the Formula (12):

$$Gradient\ descent = \frac{dJ}{d\theta} \tag{12}$$

Where θ is a vector containing all the parameters given as output by the examined layer, thus the predicted output. The cost function and the gradient descent is calculated separately for each layer. The aim is to minimize as much as possible the cost function by using the backpropagation which allow to update weight and biases in order to obtain outputs as much equal as possible to the target values. What has been reported until here is a brief presentation of the DL architecture and of how it works. The deep learning techniques are acquiring much importance, especially nowadays, in the medical fields where the possible applications are many and very promising. The performances are improving even more thanks to the advancements in technologies, like the graphic elaboration units (GPUs), processors etc. One of these applications is the tomographic image reconstruction. This is an example of inverse problem in which externally measured data are linked to internal structures in a complicated way and processed to reconstruct internal features in cross sections or volumetrically. The use of the deep learning for this purpose has been demonstrated to be feasible and promising. It has competitive performances for CT image reconstruction of low-dose CT [30].

For example in a paper reported by Zhu et al. is proposed to learn the entire reconstruction operation only from raw data and corresponding images. The idea at the basis is to model an autoencoder-like dimensionality reduction in raw data and reconstruction domain. Then both are linked using a non-linear correlation model. The entire model can then be converted into a single network and trained in end to end manner. The end to end training characterizes the deep learning by processing the data without the need of having a prior feature extraction. Despite the promising results of this approach, learning operators completely data driven, present the risk that undesired effects may occur. For this reason it is important to integrate the prior knowledges to the structure of the operators. The solution is to design a neural network inspired by iterative algorithms that minimize the energy function step by step. Thus it can be used the concept of variational networks that allows to map virtually all the iterative reconstruction algorithms into deep networks by using a fixed number of iterations. The physical simulation is another emerging field of application of the deep learning. For example Han et al. tried to convert MR volumes to CT volumes. In addition also the image analysis is an area in which the deep learning can be used. This is related to the CAD systems. It is not only acting as a support for quantifying the evidence towards the diagnosis, but it is the diagnosis itself that need to be predicted. Actually the CAD is an hot topic. Especially for complex diagnosis, the actual deep

networks, that result immediately in a decision, are not well suited. This is because it is difficult to understand the evidence. Hence approaches that link observations to evidence are needed. A particular problem related to the image analysis is the image detection and recognition. It is for example a fundamental aspect for the cancer detection in CT images. It is needed to have an efficient parsing because often the images are volumetric. A good deep learning solution for this task is a neural network-based boosting cascade. This process the entire volume to reliably detect anatomical structures. In literature there are many methods available and one of these replace the search process by an artificial agent. It is able to detect anatomical landmarks using a deep reinforcement learning. An important advantage of this method is that it is fast, in fact it is able to detect hundreds of landmarks in a complete CT volume in few seconds. The image detection and recognition by using deep learning solutions is an important task also in histology with cell detection and classification. In addition also image segmentation benefited from the recent development in deep learning. As already said, the image segmentation is an important step in CAD systems thus it need to be as much efficient as possible. The aim of the image segmentation is to determine the outline of an organ or of an anatomical structure. Approaches based on convolutional neural networks (CNN) seem to dominate in this field. However an interesting scope of research is represented by the revisitation of traditional segmentation approaches fused with deep learning in an end-to-end fashion. Yet another interesting class of segmentation algorithms is the use of recurrent networks for medical image segmentation [37].

4.1 Convolutional neural networks

The main limitation of the deep learning is that it requires a lot of data in input and the performances of the DL model increase only logarithmically with the amount of data. Another problem linked to deep learning is the high amount of time required. An improvement to all these limitations is represented by the convolutional neural network. Its architecture emerges from the deep learning, but it is computationally more efficient. These CNNs are able to find solutions that are equal or even better than many state of the art algorithms. In addition the computational costs at inference time are much lower, especially in the field of medical image analysis (detection, segmentation, registration, reconstruction and physical simulation tasks) [37].

The CNNs are a sub-class of artificial neural networks that have become dominant in various computer vision tasks. It is actually the most established algorithm among various deep learning models. CNNs are attracting a lot of interest across a variety of fields, including the radiology and in

particular the radiological image analysis. CNN is designed to automatically and adaptively learn spatial hierarchies of features through back propagation by using multiple building blocks such as convolutional layers, pooling layers and fully connected layers. The architecture of CNNs, represented in the Figure 4.5, includes several building blocks such as convolutional layers, pooling layers and fully connected layers. The first 2, convolution and pooling layers, perform feature extraction. Whereas the third layer maps the extracted features into the final output. The step where input data are transformed into output data through these layers is called forward propagation. The convolutional layers play a fundamental role in CNNs because allow feature extraction through a stack of mathematical operations included in this layers. In particular it consists of a combination of linear and nonlinear operations, such as convolution and activation functions (non-linear by nature). The input to the CNN is digital images, 2D or 3D, composed by pixels and voxels respectively. In the case of 2D images, the information is stored in a grid, an array (matrix) of numbers. The features from the images can be extracted through kernels, which are sets of learnable parameters, thus optimal feature extractors which is applied at each image position. The fact that the kernels look at different positions in the images, makes CNNs highly efficient for image processing since features may be everywhere. During the training phase the kernels are optimized in order to minimize the difference between the output given by the CNN and the ground truth label. This can be achieved through the back propagation and the gradient descent. The “deeper” go inside the CNN and the more complex can become the features. The CNN is fed with a 2D image which is seen by the computer as an array of numbers comprised between 0 and 255 (in the case of an 8 bit processor). Each of these numbers correspond to the brightness associated to every single pixel. This image then enters the convolution layer for feature extraction. In the convolution layer a kernel, represented by a small array of numbers, is applied across sub-matrices, tensors, of the input (of the same dimension of the kernel). Also the

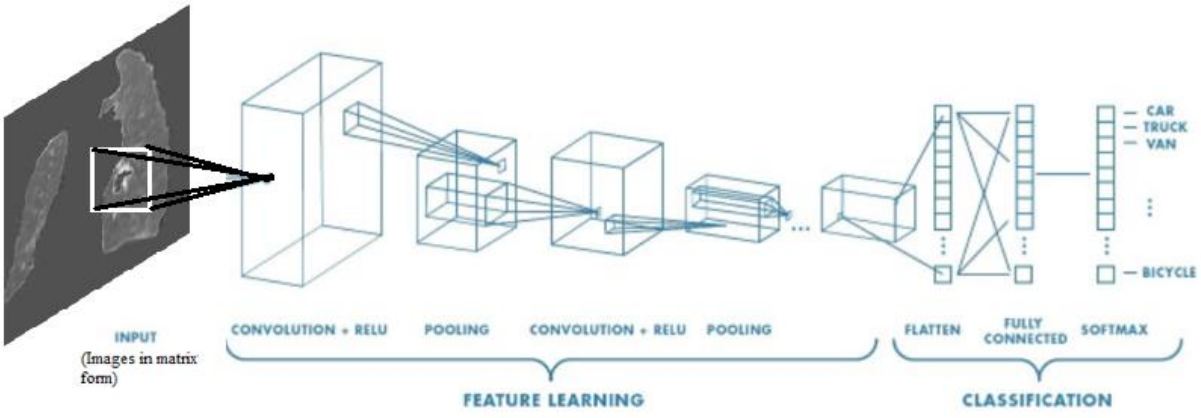

Figure 4.5 Basic architecture of Convolutional Neural Network

tensor is constituted by an array of numbers. An element-wise product between each element of the kernel and the input tensor is calculated at each location of the tensor and summed to obtain the output value in the corresponding position of the output tensor, called feature map. The feature maps have the same dimension of the kernel. This procedure is repeated applying different kernels (also called filters), to form an arbitrary number of feature maps, which represent different characteristics of the input tensors. Different kernels can, thus, be considered as different feature extractors. Two key hyperparameters that define the convolution operation are size and number of kernels. The former is typically 3 × 3, but sometimes 5 × 5 or 7 × 7. The latter is arbitrary, and determines the depth of output feature maps. When deal with kernels, there are 2 key hyperparameters that define the convolution operation. Hyperparameters are variables that need to be defined and set a priori, before the starting of the training process. These 2 hyperparameters are size and number of kernels. Typically kernels have dimension of 3 × 3, but sometimes 5 × 5 or 7 × 7. The number of kernels instead are arbitrary, and determines the depth of output feature maps. Usually it is needed also a padding in order to let the feature maps having the same dimensions of the input tensor. This paddling consists on adding rows and columns of zeros on each side of the input tensor, so as to fit the centre of a kernel on the outermost element. The main advantage of CNNs is the sharing of the same key, thus of kernels across all the image positions. This weight sharing characterizes the convolutional operations as follow:

- Let invariant the local feature patterns extracted with the translating kernels.
- Lean spatial hierarchies of feature patterns by down sampling in conjunction with a pooling operation, resulting in capturing an increasingly larger field of view.
- Increases model efficiency by reducing the number of parameters to learn in comparison with fully connected neural networks.

Then the output of the convolutional layer is then passed through a non-linear activation function. Some of the possible activation functions are reported in Table I, but the most commonly used is the ReLU activation function. The output of the activation function, in turn, pass to the pooling layer which down-sample by reducing the in-plane dimensionality of the feature maps. In conclusion can be said that in the convolutional layer, the parameter, thus the variable that is optimized automatically during the training process, are kernels. Instead the hyperparameters, variables defined by hand, are kernel size, number of kernels, stride, padding and activation function.[38].

Polling consists of aggregating pixel values of neighbourhoods using a permutation invariant function, typically the max or mean operation [34].

This process introduces 2 important aspects:

- Translational invariance to small shifts and distortions. Thus for the subsequent processing the exact location of the examined object is not important. For example if the model if a classifier which has to detect the presence of a lung cancer nodule, it is needed to detect whether such suspicious nodule is present or not. The position of this one is not required.
- Decreased number of subsequent learnable parameters. Thus the memory requirements are reduced.

At pooling layers there are no learnable parameters, whereas filter size, stride, and padding are hyperparameters in pooling operations, similar to convolution operations. The most popular form of pooling operation is max pooling, which extracts patches (sub-matrix with the same dimension of the output which is usually 2 x 2) from the input feature maps, and give as outputs the maximum value in each patch, and discards all the other values. The max pooling operation is like filtering. The most used size of the max pooling is 2 x 2 with a stride of 2 (shifting positions of the filter inside the original matrix). Another pooling operation is the global average pooling. This type of pooling performs an extreme type of down sampling, where a feature map with size of height × width is down sampled into a 1 x 1 array by simply taking the average of all the elements in each feature map, whereas the depth of feature maps is retained. This operation is typically applied only once before the fully connected layers. Applying the global average pooling allow to reduce the number of learnable parameters and enables the CNN to accept inputs of variable size. At the level of the pooling layer there are no optimized parameters but the hyperparameters are the pooling method, the filter size, the stride and the padding. The output feature maps of the final layer, which can be convolution or pooling, is usually flattened. The output is transformed into a one-dimensional (1D) array of numbers (or vector), before entering into one or more fully connected layers, also known as dense layers. Here every input is connected to every output by a learnable weight. Once the features extracted by the convolution layers and down sampled by the pooling layers are created, they are mapped by a subset of fully connected layers to the final outputs of the network, such as the probabilities for each class in classification tasks. The final fully connected layer typically has the same number of output nodes as the number of classes. The activation function applied to the last fully connected layer is usually different from the others. The activation function to be used, depends on the task, for example for a multiclass classification, it is used the softmax function. This normalizes the real values obtained as output from the last fully connected layer to target class probabilities, where each value ranges between 0 and 1 and all values sum to 1. In the fully connected layer the parameters that can be retrieved are the weights and the hyperparameters are the number of weights and the activation function to use. After having defined the network it is needed to train it. Training a network is a process of finding kernels in convolution layers and weights in fully connected layers.

These values should minimize differences between output predictions and given ground truth labels on a training dataset. The process is exactly the same as reported for DL models. Firstly it is evaluated the model performance through forward propagation on a training dataset with initial value of kernels and weights. The performance is quantified through the loss function. These loss values are then used for optimizing the learnable parameters by the back propagation algorithm and the gradient descent optimization algorithm. As previously reported there are many different types of loss function and in this case it is defined as an hyperparameter thus need to be chosen by the user. The loss function to be used, depends on the task, for example for a regression problem with continuous values, it is used the mean squared error, for a multiclass classification instead is the cross entropy. Then after having "quantified" how much is the error of the network, it is needed to optimize the parameters to have the best performance possible of the network. The aim is to reduce the error as much as possible. For this purpose it is used an optimization algorithm such as the gradient descent. This iteratively updates the learnable parameters, kernels and weights. The gradient of the loss function provides the direction in which the function has the steepest rate of increase, and each learnable parameter is updated in the negative direction of the gradient with an arbitrary step size determined. This step size depends on a hyperparameter called learning rate. Mathematically the gradient is a partial derivative of the loss with respect to each learnable parameter, and a single update of a parameter is formulated as in Equation (13):

$$w := w - \alpha * \frac{\partial L}{\partial w} \tag{13}$$

Where w represents each learnable parameter, α stands for a learning rate, and L stands for loss function. The learning rate is one of the most important hyperparameters to be set before the training starts. For practical reasons such as memory limitations [38].

In addition this value needs to be a trade-off, otherwise there can be the problem of exploding gradient (if too high) or of the creation of a stagnation point(if too low) [37].

Usually it is used the stochastic gradient descent (SGD), which consists of calculating the loss function with respect to parameters retrieved from the training of a training data subset, called mini batch. Here is an additional hyperparameters which is the size of the mini batch. After having trained the network it is needed to test it with the test set. This is used only once at the end of the project to evaluate the performances of the model. The test set consists of data that the model has never seen, before [38].

Because models obtained through deep learning and conventional machine learning are known to face the overfitting problem, its performance should be evaluated with cases which are not included in the training phase. Overfitting refers to a situation where a model learns statistical regularities

specific to the training set. To avoid this problem it is used the validation set. This set is used to estimate the generalization error during or after the training [39].

In contrast to the training set, the validation set is never used to actually update the parameter weights, but it is used for fine tuning hyperparameters and model selection. Hence, the loss of the validation set allows an estimate for the error on unseen data. During optimization, the loss on the training set will continuously fall. However, as the validation set is independent, the loss on the validation set will increase at some point in training as can be seen in the Figure 4.2. This is typically a good point to stop updating the model before it overfits to the training data [38].

Thus the purpose of the training and validation is to let the model give in output values as much similar as possible to the ground truth values, but at the same time the designed CNN must be able to perform well also on unseen data. To prevent the model to "specialize" itself on the training data, these should be used as input to the network only once, avoiding overestimation of the model performances. The best solution for reducing overfitting is to obtain more training data. A model trained on a larger dataset typically generalizes better, though that is not always attainable in medical imaging. The other solutions include regularization with dropout or weight decay, batch normalization, and data augmentation, as well as reducing architectural complexity. These are techniques to mitigate the overfitting, but there is still a concern of overfitting to the validation set due to information leakage during the hyperparameter fine-tuning and model selection process. Batch normalization is a type of supplemental layer which adaptively normalizes the input values of the following layer. This layer performs standardizing and normalizing operations on the input of current layer coming from previous layer. In fact normalization is a data pre-processing tool used to bring the numerical data to a common scale without distorting its shape. At the same time the batch normalization improves the gradient flow through the network, allowing higher learning rates, and reducing the dependence on initialization. Data augmentation is also effective for the reduction of overfitting, which is a process of modifying the training data through random transformations, such as flipping, translation, cropping, rotating, and random erasing, so that the model will not see exactly the same inputs during the training iterations. This is actually a step of input data preparation which allows to increase the numbers of data available for training. In spite of these efforts, there is still a concern of overfitting. Therefore, reporting the performance of the final model on a separate (unseen) test set, and ideally on external validation datasets if applicable, is crucial for verifying the model generalizability [38].

For testing, output values from the trained CNN (giving as input the test set), are compared with teaching data. The performance of the model is evaluated with appropriate methods, such as sensitivity, specificity, area under the receiver operating characteristic curve (AUC), Dice similarity

coefficient score, etc. For the testing phase the amount of calculations is not as large as for training. CNN models are very powerful tools in medicine, especially in radiology for the analysis of radiological images. In particular it is a promising radiomics method, a tool that extracts a large number of features from radiological images. The interpretation of these images and the extraction of features done by radiologists, suffer from subjectivity and is performed in qualitative way. The increase in data size of radiological images and the need to retrieve quantitative parameters (features) from them, led to the development of new techniques which involve the use of CNNs. Multiple CNN applications, like lesion detection, lesion evaluation, segmentation of images, classification etc, represents a powerful radiomics tool [39].

When deal with medical images, it is better to use a slightly different architecture of the CNN. The default CNN architecture can easily accommodate multiple sources of information or representations of the input, in the form of channels presented to the input layer. This idea can be taken further, and channels can be merged at any point in the network. Under the intuition that different tasks require different ways of fusion, multi-stream architectures are being explored. These models, also referred to as dual pathway architectures, having 2 main applications: multi-scale image analysis and 2.5D classification; both relevant for medical image processing tasks. For the detection of abnormalities, context is often an important cue. The most straightforward way to increase context is to feed larger patches to the network, but this can significantly increase the amount of parameters and memory requirements of a network. A first important task that can be addressed by the use of CNNs is the image/exam classification. Image or exam classification was one of the first areas in which deep learning made a major contribution to medical image analysis. In exam classification, the input is one or more images and the output is a single diagnostic variable, like disease present or not. The diagnostic images can be seen as a sample. In this field an important role is played by the transfer learning which consists of the use of pre-trained networks (usually on natural images). These pre-trained networks are then "adjusted" to perceive the requirements of the task. There are mainly 2 types of transfer learning strategies:

1. Use of a pre-trained network as feature extractor.
2. Fine tuning of a pre-trained network on medical data.

Summarizing, in exam classification CNNs are the current standard techniques.

The CNNs can be exploited for the segmentation task, it is typically defined as identifying the set of voxels which make up either the contour or the interior of the object(s) of interest. The segmentation of organs and other substructures in medical images allows quantitative analysis of clinical parameters related to volume and shape. Furthermore, it is often an important first step in computer-aided detection pipelines. The most well-known, in medical image analysis, is U-net, which presents

two main architectural novelties. The introduction of an equal amount of up sampling and down sampling layers. U-net combines up sampling with skip connections between opposing convolution and deconvolution layers. This concatenate features from the contracting and expanding paths. From a training perspective this means that entire images/scans can be processed by U-net in one forward pass, resulting directly in a segmentation map. This innovative CNN takes into account the full context of the image, which can be an advantage in contrast to patch-based CNNs. Furthermore, in literature can be found an example of full 3D segmentation achieved by feeding U-net with a few 2D annotated slices from the same volume. Other authors have also built derivatives of the U-net architecture. Milletari et al. proposed a 3D-variant of U-net architecture, called V-net, performing 3D image segmentation using 3D convolutional layers with an objective function directly based on the Dice coefficient. Anatomical object localization (in space or time), such as organs or landmarks, is an important pre-processing step in segmentation task. This is of fundamental importance in the clinical workflow for therapy planning and intervention. Localization in medical imaging often requires parsing of 3D volumes. To solve 3D data parsing with deep learning algorithms, in particular by using pre-trained CNNs, several approaches have been proposed that treat the 3D space as a composition of 2D orthogonal planes. Concluding, localization through 2D image classification with CNNs is nowadays the most popular strategy to identify organs, regions and landmarks, with good results. After having segmented the organ or the body part of interest it is needed to detect objects of interest or lesions in images. The object or lesion detection is the key part of the diagnosis and is one of the most labour-intensive for clinicians. Typically, the task consists of the localization and identification of small lesions in the full image space. There has been a long research tradition in computer-aided detection systems, designed to automatically detect lesions. These systems aim to improve the detection accuracy and/or to decrease the reading time required by clinicians Interestingly. The first object detection system using CNNs has been proposed in 1995, using a CNN with four layers to detect nodules in x-ray images. Most of the published deep learning object detection systems still uses CNNs nowadays to perform pixel (or voxel) classification, after which some form of post processing is applied to obtain object candidates. The incorporation of contextual or 3D information is also handled using multi-stream CNNs. Subsequently after having detected the object it is necessary to identify it; this is the object classification task. It usually focuses on the classification of a small (previously identified) part of the medical image into two or more classes like nodule classification in chest CT. For an accurate classification of the lesion it is necessary to have local information on lesion appearance and global contextual information on the location of the lesion. This combination is typically not possible in generic deep learning architectures. Thus it is more convenient to use multi-stream CNN architectures to resolve the task in a multi-scale fashion.

There are many proposals in literature. For example Shen et al. used 3 CNNs, each of which takes a nodule patch at a different scale as input. The resulting feature outputs of the 3 CNNs are then concatenated to form the final feature vector. Incorporating 3D information is also often a necessity for good performance in object classification tasks in medical imaging. As images in computer vision tend to be 2D natural images, networks developed in those scenarios do not directly leverage 3D information. Authors have used different approaches to integrate 3D in an effective manner with custom architectures. Setio et al. used a multi-stream CNN to classify points of interest in chest CT as a nodule or non-nodule. Up to nine differently oriented patches extracted from the candidate were used in separate streams and merged in the fully connected layers to obtain the final classification output. Finally the aim of lesion classification is to obtain high-quality labelled medical data, avoiding the necessity of hand-crafted annotations made by clinicians. However this task is not so popular due to the high complexity introduced by the need of incorporating contextual or three-dimensional information. Finally a task that can be accomplished by using CNNs is the image generation and enhancement. This task may range from removing obstructing elements in images, normalizing images, improving image quality, data completion and pattern discovery. In image generation, 2D or 3D CNNs are used to convert one input image into another. Typically, these architectures lack the pooling layers present in classification networks. For example has been possible to retrieve CT imaged form the magnetic resonance images (MRI). In addition with multi-stream CNNs super-resolution images can be generated from multiple low-resolution inputs. This strategies present many advantages like inferring missing spatial information. The anatomical application areas of the CNNs are many and between them can be found the chest. In thoracic image analysis of both radiography and computed tomography, the detection, characterization, and classification of nodules is the most commonly addressed application. In particular in chest CT, a very common research topic is the detection of textural patterns indicative of interstitial lung diseases. Together with the detection of abnormalities, it is also necessary to infer the nature of these ones. Thus another important anatomical application context for the CNNs, is the digital pathology and histology. The growing availability of large scale gigapixel whole-slide images (WSI) of tissue specimen has made digital pathology and microscopy a very popular application area for deep learning and in particular CNNs techniques. The developed techniques applied to this domain focus on three broad challenges:

1. Detection, segmentation and classification.
2. Segmentation of large organs.
3. Detection and classification of the disease of interest at the lesion level [34].

In literature are present many studies in which are reported successfully built models used to obtain clinically useful information in certain tasks. For some tasks, CNNs achieved higher performance

compared to conventional machine learning methods. And for some tasks, CNNs achieved the performance almost equal to radiologists. Transfer learning might be useful in developing CNN models for relatively rare diseases. However the deep learning techniques are evolving rapidly, together with an increase in performances of some CNN models. Despite all these advantages, the deep learning and CNN based models also present some limitation. For example the optimal structure and hyper-parameters of CNNs and the numbers of cases needed to train models differ from task to task. Another problem is the necessity of many cases for training process [39].

4.1.1 Convolutional neural network for lung cancer diagnosis: State-of-the-art

A major challenge in lung cancer screening is the detection of lung nodules from CT images and scans. CNNs are the state of the heart methods in lung nodule identification and classification. However CNNs have any drawbacks most of which are due to the routing data procedure. The routing is the process by which information coming from one layer are related to the subsequent layer. CNNs perform routing via pooling operations such as max-pooling and average pooling. This causes the discarding of some important information such as location and pose of the lung nodule. For this reason in literature many alternative CNN architectures are proposed. These innovative solutions are based on CNNs basic structure but present some modifications. The alternative convolutional neural networks proposed differ from the dimension of input data. Some work on CT images, thus on 2D images (pixels) and others work with CT scans, so 3D images (voxels) [40]. In Table II and in Table III are reported the studies analysed in the following paragraphs.

Table II. Analysed studies regarding lung cancer detection from CT images.

Study	*Aryanet al. [40]*	*Agarwal et al. [43]*	*AL-Huseiny et al [44]*
Publication year	2017	2021	2021
Input type	2D CT slices	2D CT slices	2D CT slices
Input dimension	32 x 32 pixels (grayscale)	227x227 (RGB)	512x512 (grayscale)
CNN type	2D CapsNet	AlexNet	GoogleNet
Model task	Lung nodules detection	Classification of lung nodules in malignant and benign	Classification of lung nodules in malignant and benign
Advantages	Dynamic routing	Combine feature extraction and classification in just one model	Computationally inexpensive
Disadvantages	Time consuming	Need of high computationally costly supervised technique to reduce false positive	Need of images pre-processing utilizing region of interest extraction
Accuracy (%)	89.44	96.00	94.38

Table III. Analysed studies regarding lung cancer detection from CT scans.

Study	*Afshar et al. [45]*	*Al-Yasriy et al. [46]*	*Alder et al. [47]*
Publication year	2020	2020	2020
Input type	3D CT slices	3D CT slices	3D CT slices
Input dimension	-	227x227x3	224x224x3
CNN type	3D multi-scale CapsNet	3D AlexNet	3D deep ConvNet
Model task	Classification of lung nodules in malignant and benign	Lung cancer detection and classification	Lung nodules detection
Advantages	Higher accuracy than classic 3D CapsNet	Improvement of existing CNN model	-
Disadvantages	Lower accuracy in malignant nodule classification	Training process on a quite small dataset	Region-based approach requiring careful manual interaction for seed point initialization
Accuracy (%)	93.12	93.55	-98.00 for solid nodule -99.50 for part-solid nodule -97.20 for non-solid nodule

4.1.1.1 Lung cancer diagnosis from computed tomography images: 2D CapsNet based model

Skilled radiologists have a high degree of accuracy in diagnosis of lung cancer from CT images. However, there are some remaining problems in the detection of this disease in the early phase, problems that cannot be corrected with current methods of training and high levels of clinical skill and experience. These problems would cause for example the miss rate in the detection of small pulmonary nodules, the missed detection of minimal interstitial lung disease and the inability to detect changes in pre-existing interstitial lung disease [41].

As previously said there are many steps involved in identification and classification (malignant or benign) of lung nodules. It is needed to start from segmentation which allow to "select" the organ of interest which in this case are the lungs and then identify the suspicious nodules. Actually the identification of the candidates nodules is not a mandatory step, but many examples proposed in literature show that this helps improving the performances of the classifier. In fact the latter one has worse performances when working with the entire organ which is a too wide region to inspect. Then the selected nodules enter the classifier which define if the nodules are malignant or benign. This is the main working principle behind many work proposed in literature concerning the lung cancer detection. In the following there will be proposed many CNN networks for the classification task of lung cancer. The CNNs are powerful tools for solving this task. In particular there are some accessible pre-trained models which constitute the state of the art in solving this task. One example of 2D CNN is the 2D Capsule Network (CapsNet) recently introduced by Sabour et al. [40]. This network presents a new architecture that tries to address the CNNs' shortcomings. The idea is to encode the relative relationships (e.g., locations, scales, orientations) between local parts and the whole object. Encoding these relationships allow the model to understand the entire 3D space. This enables CapsNet to recognize objects from different 3D views that were not seen in the training data. CapsNets employs a dynamic routing mechanism to determine where to send the information. The original algorithm has been successfully used for training the network on hand-written images of digits and achieved state-of the-art performances. The drawback of the classical CapsNet is the huge amount of time required by the dynamic routing operations. This prevents the use of CapsNet for higher dimensional data such as 2D CT images. In literature can be found some applications of modified CapsNet for the detection of lung cancer from CT images (2D). One well performing solution for the lung cancer detection and classification tasks is the fast Capsule Network whose structure is reported in the Figure 4.6. The fast CapsNet is based on the same architecture of the classical CapsNet but with some additional modifications. The network is composed by more capsules. A capsule is defined as a group

of neurons (whose outputs form an activation vector). They predict the presence and the pose parameters of a particular object at a given pixel location. In particular the object's pose information (like size, orientation, position etc) is given by the direction of the activation vector while the length of this vector represents the estimated probability that the object of interest (in this case the lung nodule) exists. The main basic structure and the relationships between capsules can be seen in the Figure 4.7. The relationship between i-th capsule in a lower layer and j-th capsule in the next higher layer is encoded using a linear transformation matrix Wij. The propagation of information is regulated by the Equation (14):

$$\widetilde{u_{j|i}}=W_{ij} * u_i \tag{14}$$

Where the vector $\widetilde{u_{j|i}}$ represents the belief of i-th capsule in a lower layer about j-th capsule in the higher layer. For example $\widetilde{u_{j|i}}$ can represent the predicted pose of the lung nodule according to the detected pose of one feature of the nodule itself. During the training, the network will gradually learn a transformation matrix for each capsule pair to encode the corresponding part-whole relationship. Having computed the prediction vectors, the lower-level capsules then route their information to parent capsules that agree the most with their predictions. The mechanism that ensures that the outputs of the child capsules get sent to the proper parent capsules is named dynamic routing. The decision is taken based on the value of the routing coefficient c_{ij} between the i-th capsule in the lower layer and the j-th capsule in the higher layer where $\sum_j c_{ij} = 1$ and $c_{ij} \geq 0, \forall j$. When c_{ij}= 1, all information from capsule i will be sent to capsule j, whereas when c_{ij}= 0, there is no information flowing between the two capsules. Dynamic routing method iteratively tunes the Cloud-YLung for Non-Small Cell

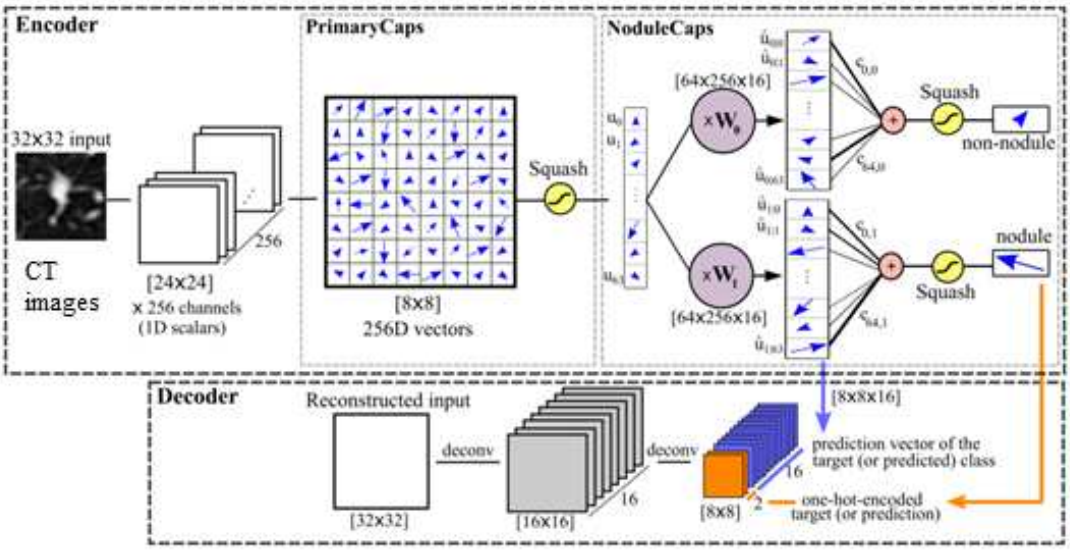

Figure 4.6. Architecture of the proposed CapsNet.

Lung Cancer Histology Classification from 3D Computed Tomography Whole-Lung Scans coefficients and routes the child capsules' outputs to the appropriate capsule in the next layer so that they get a cleaner input, thus determining the pose of the objects more accurately. The decision to send the information obtained

as output by the capsule i to the parent capsule j is made by multiplying the adjusted routing parameters by the prediction vector before sending to the higher level capsules. The output of each parent capsule v_j is computed as the weighted sum of all predictions from child capsules, then passed through a squash non-linearity. Squashing makes sure that the output vector has length no more than 1 (so that its length can be interpreted as the probability that a given feature being detected by the capsule) without changing its direction. Thus a child capsule will send more information to the parent capsule whose output v_j is more similar to its prediction $\widetilde{u_{j|i}}$. The network contains two main parts: encoder and decoder. The encoder is composed by three layers: two convolution and one fully-connected. The encoder works on the pixels given in input to identify the presence of lung nodules. The decoder tries to reconstruct the input from the final capsules, which will force the network to preserve as much information from the input as possible across the whole network. This effectively works as a regularizer that reduces the risk of over-fitting and helps generalize to new samples. In decoder, the outputs of the final capsules are all masked out (set to zero) except for the ones corresponding to the target (while training) or predicted (while testing) class. In the Figure 4.6 it's illustrated the proposed architecture, which consists of 3 different layers. The input data are 2D images, taken as middle slices of the 3D volumes along the x-axis (as x-axis contains more information according to radiologist's feedback). The first layer is a 2D convolution layer with 256 filters of size 9 and stride 1. The first layer receives an input image of 32 x 32 pixels and gives in output 24 × 24 × 256 tensor which are the basic features extracted from input image. The next layer is the PrimaryCaps layer which applies 256 convolutional filters on the input to obtain 8×8×256 tensor. This tensor is considered to be 8 × 8 thus 64 capsules, each with 256D (256 dimensions as

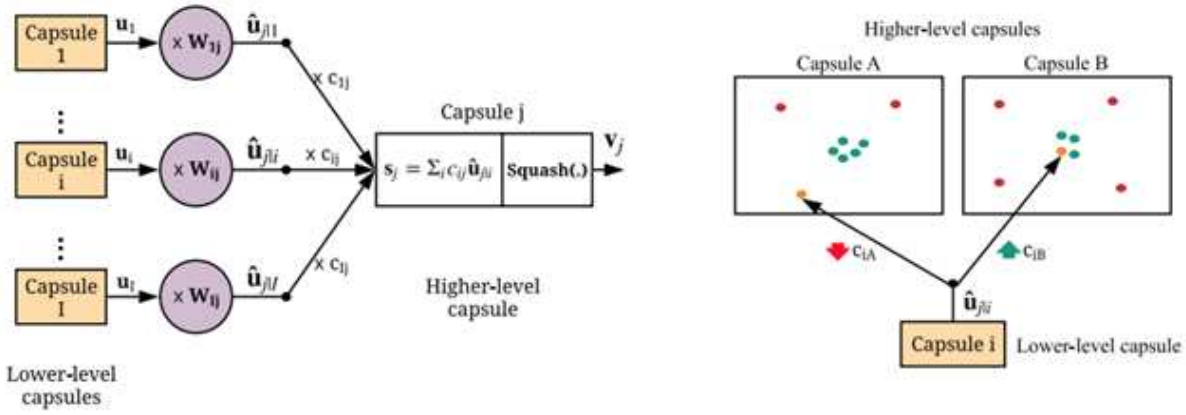

Figure 4.7 Interaction between capsules and information path decision.

matrix of 16 x 16). This results in only one capsule at each pixel, therefore, effectively enforces consistent dynamic routing. These vectors will then be passed through a squashing non-linearity to get the final output of the lower-level capsules. The proposed architecture is a departure from the original CapsNet. While the total number of parameters of the two networks are the same, the fats CapsNet drastically reduces the number of PrimaryCaps from 2048 to 64, thereby decreases the number of voting coefficients by 32 times. In the final layer, called NoduleCaps, 256D output from PrimaryCaps layer are multiplied with its own 256 × 16 transformation matrix which maps the output to the 16D space. Finally, the routing parameter tunes the coupling coefficients and decides about the amount of contribution they make to the NoduleCaps. During training, the coefficients are initialized uniformly, meaning that the output from the PrimaryCaps layer is sent equally to all the capsules in the NoduleCaps layer. Then the routing between the two layers is learned iteratively [40].

4.1.1.2 Lung cancer diagnosis from computed tomography images: AlexNet based model

Another different architecture of the CNN that can be exploited for the lung cancer detection is called AlexNet. This is a Convolutional Network which contains 8 layers. It is able to automatically extract the distinctive features from input images and classify them. This network showed, for the first time, that the features obtained by learning can transcend manually-designed features, breaking the previous paradigm in computer vision. The architecture of AlexNet is very similar to the LeNet architecture. In AlexNet's first layer, the convolution window shape is 11×11. A larger convolution window is needed to capture the object in the CT images. The convolution window shape in the second layer is reduced to 5×5, followed by 3×3. In addition, after the first, second, and fifth convolutional layers, the network adds maximum pooling layers with a window shape of 3×3 and a stride of 2. Moreover, AlexNet has ten times more convolution channels than LeNet. After the last convolutional layer there are two fully-connected layers with 4096 outputs. These two huge fully-connected layers produce model parameters of nearly 1 GB. Due to the limited memory in early GPUs, the original AlexNet used a dual data stream design, so that each of their two GPUs could be responsible for storing and computing only its half of the model. Fortunately, GPU memory is comparatively abundant now, so it is rarely needed to break up models across GPUs these days. AlexNet uses the ReLU activation function. The simplicity of this function makes model training easier when using different parameter initialization methods. Another important improvement introduced by AlexNet is the data augmentation in the training loop with flipping, clipping, and colour

changes. This makes the model more robust, and the larger sample size effectively reduces overfitting [42].

In literature can be found an application of this network for lung cancer detection, proposed by Agarwal et al. This study uses a modifies version of AlexNet CNN (Figure 4.8) to classify lung cancer from CT images following a specific framework. First of all has been extracted the green channel from the original colour CT image. The second step consists of extracting lung regions from the CT images using a multilevel thresholding process. Third the affected and non-affected regions are separated using thresholding and morphological segmentation methods. Finally these segmented tumour regions are given as input to the AlexNet CNN which classifies the tumour as malignant or benign. The strength of this network resides on the fact that can combine feature extraction and classification in just one model. This alternative AlexNet architecture achieve an higher accuracy than that of the classical CNNs [43].

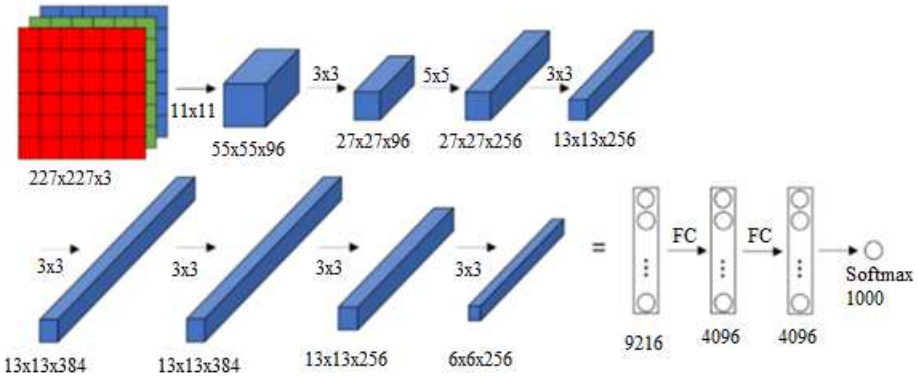

Figure 4.8 Proposed AlexNet architecture.

4.1.1.3 Lung cancer diagnosis from computed tomography images: GoogleNet based model

Another CNN architecture that can be exploited for detection of lung cancer is the GoogleNet. It is a convolutional neural network with a complex architecture. The model of this network was originally developed to account for high feature representations by using million everyday object images included in a huge dataset, the ImageNet dataset. It has the capacity to classify patterns of around ar 1000 images. One advantage is that it utilizes 12 times less parameters than AlexNet. Similar to other neural networks employed in computer vision applications, this model accepts images as input and produces labels of one of its learned classes together with the level of confidence as output. The architecture of GoogLeNet is built of 22 layers including 9 inception modules. Better results in the lung cancer detection are achieved by using slightly modified versions of the classical GoogleNet. For example in literature can be found a modified version proposed by AL-Huseiny et al. (Figure 4.9) which makes uses learnable filters with sizes ranging from (1x1) to (5x5) to perform convolution in parallel. This helps capture feature of different levels of details. The inputs given to this network are CT images with dimensions of 512 x 512 and the output is a label assigned to each image which

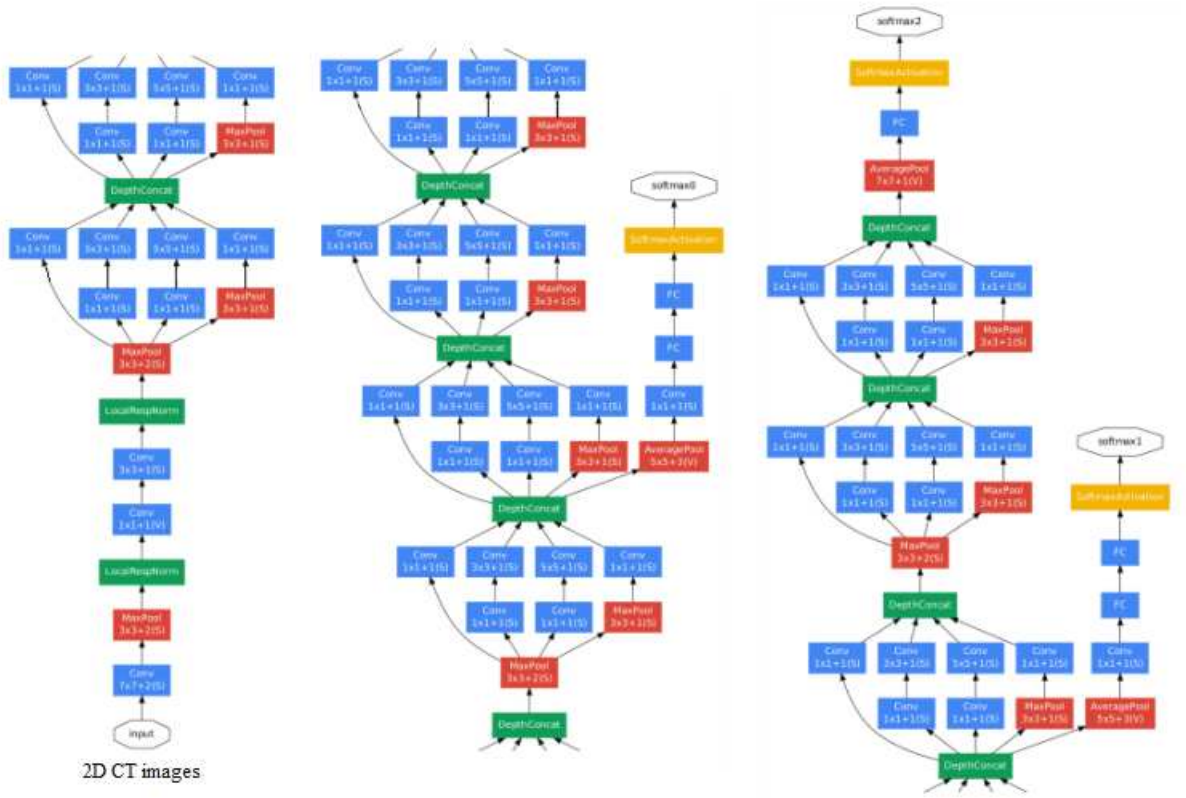

Figure 4.9 GoogleNet proposed architecture.

classify the nodule as malignant or benign. Experimental results show that the trained model has gained an overall accuracy of 94.38% on the validation data which is a good result [44].
These are only some of the available CNN networks, in fact in literature there can be found many other proposals for lung cancer detection starting from 2D CT images. Some of these are 2D ResNet-50, VGG, Inception, DenseNet etc [38].

4.1.1.4 Lung cancer diagnosis from computed tomography scans: 3D CapsNet and 3D multi-scale CapsNet based model

The detection of lung cancer can be performed in the same manner described previously but using 3D CT scans, given in input to 3D CNNs. 2D images dismiss the 3D spatial size, indicating that it is actually incapable of making complete usage of the 3D information furnished by CT scans. Thus a solution to this problem is the use of 3D CNNs, which are able to exploit the full nature of the 3D CT data. The 3D convolutional neural network is able to make use of the full nonlinear 3D context information for detecting lung cancer nodules. 3D CNN is essentially the same thing of the 2D CNN except for the fact that the input will be 3D data, instead of being a singular layer. Also the filters are 3D. So instead of detecting 2D features such as edges and corners, it will detect the same features but in the 3D fashion. This represents a very critical aspect especially in lung nodules detection, whose dimensions are of the order of few millimetres (mm). There are many different types of CNNs that can be employed for the classification of lung nodules. One example of 3D CNN is the so called 3D CapsNet. The 3D version of the proposed architecture is structurally similar to the explained 2D version. In the first two layers, 2D convolutions are replaced by 3D convolution operators capturing the cross-correlation of the voxels. The number and size of the filters and the strides are the same as before. This results in an $8 \times 8 \times 8$ volume of primary capsules, each sees the output of all $256 \times 9 \times 9 \times 9$ units of the previous convolution layer. Therefore, this proposed faster architecture gives 512 PrimaryCaps. This will limit the number of required routing coefficients, help the dynamic routing perform faster and perform significantly better. NodulesCaps are the same as in the 2D network, fully-connected to PrimaryCaps with the dynamic routing happening in the middle. Similar to the proposed 2D fast CapsNet, the decoder takes the prediction vectors of all PrimaryCaps (only the ones routed to the correct class), concatenates it with the one-hot-encoded label of the target to form a tensor of shape $8 \times 8 \times 8 \times 18$. It then reconstructs the input from this tensor by applying two consecutive 3D fractionally-strided convolution layers with 16 and 1 filters, respectively. In conclusion can be said that CapsNet is a promising alternative to CNN. Experimental results demonstrate that CapsNet performs well when the training size is large. The proposed fast CapsNet has more or less the same

accuracy of the standard CapsNet but the time required is 3 times less. The accuracy of the 3D fast CapsNet is higher than that of the 2D fast CapsNet [41].

Capitalizing on the success of CapsNet in biomedical domains, a lot of models use this type of CNN to construct more complex and more accurate prediction systems. For example in literature can be found a model for lung tumour malignancy prediction which makes use of CapsNet. The proposed framework is named 3D Multi-scale Capsule Network (3D-MCN). This fed the network with 3D inputs, providing information about the nodule in 3D. In particular the strength of 3D-MCN is that it is able to capture the nodule's local features, as well as the characteristics of the surrounding tissues. In addition it benefits from CapsNet-based design, which allows to work with a small number of training samples thanks to all the possible rotations and transformations of the underlying objects. The proposed 3D-MCN model (shown in Figure 4.10), takes 3D patches of the nodules at three different scales as inputs and predicts the nodule's malignancy. The rationale is that the morphological characteristics of the nodule are not the only ones predicting its malignancy, and incorporation of information obtained from the surrounding tissues and vessels play a critical role in determining the type of the nodule. This 3D-MCN model for lung nodule classification consists of three independent CapsNets, each of which takes nodule patches at a different spatial scale as input. This is why it is defined as a multi-scale learning architecture. Here, scale refers to the visible area of the tissue

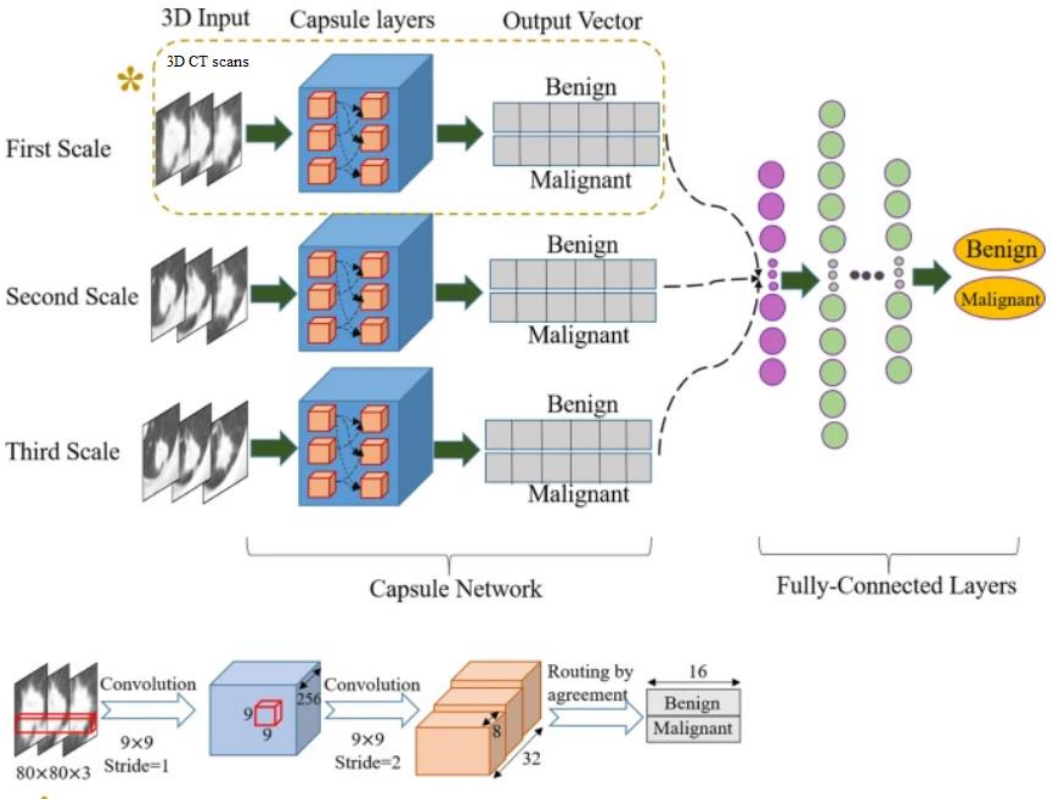

Figure 4.10 Proposed 3D-MCN network.

surrounding the nodule. Each input is a 3D nodule crop cantered at the nodule annotation. The output vectors of the three CapsNets are concatenated, and the result goes through a fusion module consisting of a set of fully connected layers. The final output is the probability of the nodule being benign or malignant. The proposed 3D-MCN model, after the testing, was shown to be advantageous over single-scale models whit an higher accuracy [45].

4.1.1.4 Lung cancer diagnosis from computed tomography scans: 3D AlexNet based model

An alternative CNN architecture for lung cancer detection is a 3D alternative to the classic 2D AlexNet, presented earlier. This has a quite complex structure as reported in Figure 4.11. Firstly, at the layer zero, which in the Figure 4.11 represents the input image, the size is 227 x 277 x 3, which are respectively relates to height, width and depth. Then a convolution process is done at the first layer with 96 filters whose size is 11 x 11 (as the 2D version) and the amount of stride is 4, without padding. Therefore, the result from this layer is an image with size of 55 x 55 x 96. At the second layer, a max-pooling operation follows the convolution process. This operation with a filter size of 3 x 3 and a stride equals to two, gives an image with size of 96 x27 x 27. The depths remain the same since this

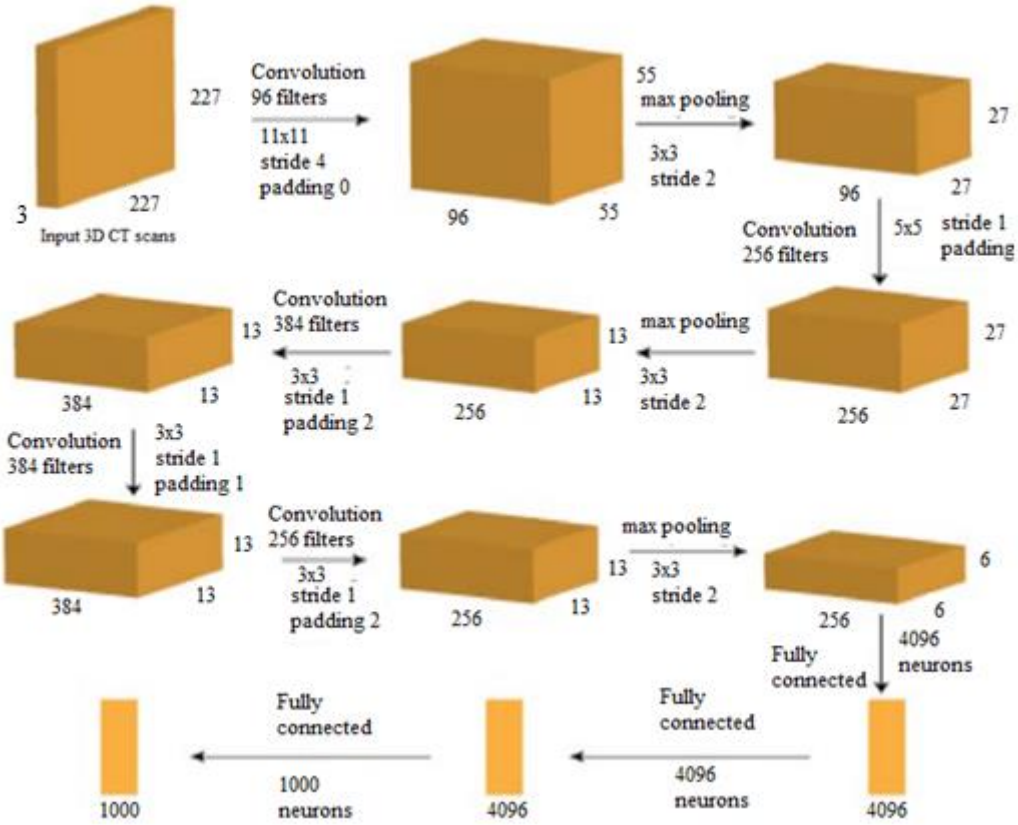

Figure 4.11. Proposed 3D AlexNet architecture.

operation is performed on every layer independently. At layer three, is performed a convolution process with 256 filters of size of 5 x 5 and stride equals to 1. At this stage it is introduced the padding with a factor of 2 for restoring the original size of 27 x 27 but with a depth of 256. Again, Max pooling with a filter whose size is size of 3 x 3 and stride 2 is performed, to down sampling the image size to 13 x 13 x 256. At the layer five, the convolution operation is employed with a 384 filters of size of 3 x 3 and here the stride and padding equals one, so the depth is increased to 384. Again, at the next layer, a convolution operation is done with the same size as the last convolution operation, and same stride and padding amount used, so the results are not changed. At the seventh layer, a convolution operation is performed too. But with 256 filters and a size of 3 x 3. Stride and padding equals 1. Therefore the results are the same of the previous layer, but with a depth size of 256. At layer 8 it is introduced again max pooling with a size of 3 x 3, and stride 2. This yields an image with a size of 6 x 6, and a depth of 256. The ninth layer, which is the fully connected layer, multiply the image coming as input for getting a 9216 pixels. These pixels will be fed into all 4096 neurons of AlexNet, At the next layer the previous steps is repeated. The last layer is also a fully connected layer but with 1000 neurons [46].

4.1.1.5 Lung cancer diagnosis from computed tomography scans: 3D deep ConvNet based model

Recently, several works proposed an alternative, called deep ConvNets, for nodule detection. In particular one of the most computationally effective network, is the 3D multi-scale ConvNet. This is characterized by shared weights for different scales which allow to learn scale-invariant features. The 3D multi-scale ConvNet is proposed for nodule classification. In this network, the basic convolution operation has been applied in three directions (x, y, z) at the same time. The convolutional filters used are 3-D, and have been applied over the input 3-D data for automatic feature extraction for the subsequent classification. These networks are more expensive in terms of computation efficiency than 2-D-based network. Convolution operations based on 3-D need more calculations and more memory. The main advantage of 3-D CNN is that it produces multi-view features with the help of 3-D filters. For example Dou et al. proposed a 3-D CNN-based architecture for lung nodule detection. Here researchers have used three 3-D CNNs to encode spatial information and representative features using hierarchical architecture. For the first architecture, a receptive field of size $20 \times 20 \times 6$ has been used, for second architecture, the size was $30 \times 30 \times 10$, and for third architecture, it was $40 \times 40 \times 26$. Finally, the three CNNs have been merged and act as a feature extractor for nodule detection. The

proposed architecture was validated by participating and achieved a sensitivity of 94.4%, for nodule detection [47].
As said in the precvous section, these are only some of all the available networks. Some others pre-trained CNNs are: 3D ResNet-50, VGG, Inception, DenseNet etc [38].

4.2 Interpretable convolutional neural networks

DL as represented by the artificial CNNs has achieved great success in many important areas. However, the black-box nature of CNNs has become one of the primary obstacles for their wide acceptance in the majority of critical applications such as medical diagnosis and therapy. Due to the huge potential of DL, interpreting neural networks has recently attracted much research attention. There is still no unique consensus on the exact meaning of interpretability. For example some researchers explore post-hoc explanation for models, while some others focus on the interplay mechanism between algorithms. Generally speaking, interpretability refers to the extent of human's ability to understand how the model achieved the task for which it was created. The implications of interpretability in different levels can be categorized in:

- Simulatability: it is considered as the understanding over the entire model. The simpler the model is, the higher simulatability the model has. For example, a linear classifier or regressor is totally understandable. To enhance simulatability, some facilities of the models can be changed or crafted regularization terms can be used.
- Decomposability: it is to understand a model in terms of its components such as neurons, layers, blocks, and so on. Such a modularized analysis is quite popular in engineering fields. For example in machine learning, a decision tree is a kind of modularized methods, where each node has an explicit utility to judge if a discriminative condition is satisfied or not. Each branch delivers an output of a judgement, and each leaf node represents the final decision after computing all attributes. Modularizing a neural network is an advantage which allow the optimization of the network design since it is known the role of each component of the entire model.
- Algorithmic Transparency: it is to understand the training process, architecture, and dynamics of the model. In particular the fact that deep models do not have a unique solution hurts the model transparency. If it is possible to understand how learning algorithms work, DL research and applications will be accelerated.

In addition to the fact that it is difficult to uniquely define the interpretability concept, here are also some problems linked to many factors. Firstly the data wildness; although it is a big data era, high quality data are often not accessible in many domains. Other problems rise also from algorithmic complexity. DL is based on highly non-linear algorithms thus there is an high variability of the model. In addition it is also needed to take into account that the number of trainable parameters of a deep model can be on the orders of hundreds million or even more. Despite all these problems, interpretability of neural networks is necessary for letting people rely more on these promising approaches. For this reason a good interpretation model must be built. In the work proposed by Feng-Lei Fan et al. [39] five rules-of-thumb are proposed: exactness, consistency, completeness, universality, and reward. Exactness means how accurate an interpretation method is. It is a matter of understanding if the model is just limited to a qualitive description or provide also a quantitative analysis. Generally, quantitative interpretation methods are more desirable than qualitative counterparts. Consistency means that there is no contradiction in an explanation. For multiple similar samples, a fair interpretation should produce consistent answers. Universality is a concept linked to the rapid development of DL. Such diverse DL models play important roles in a wide spectrum of applications. A driven question is whether it is possible to develop a universal interpreter that deciphers as many models as possible so as to save fatigue and time. But this is technically challenging due to the high variability among models [48].

4.2.1 Interpretable convolutional neural networks in lung cancer domain: state-of-the-art

Many examples of CNNs to identify the presence of lung cancer tumour from CT images and scans are available in literature. However as previously said, especially in clinics there is a certain reluctance against what cannot be properly explained. For this reason it is necessary to explain how these algorithms succeed in such tasks. Some works are proposed in literature which exploit Gradient-weighted Class Activation Mapping (Grad-CAM).

It is a technique to visualized important sections for available classes using guided propagation. Grad-CAM uses the gradients of any targeted class and passes it through to the last convolutional layer in the network. This allows to highlight the important sections in the image for prediction. Grad-CAM calculates the gradients with respect to the feature map generated from the convolutional layer [49].

4.2.1.1 Interpretability when using computed tomography images

Some works that can be found in literature refers to the interpretability of ML models whose aim is to identify and classify lung cancer from CT slices.

An interesting study is the one proposed by Renard Elyon et al. which investigate the goodness of Grad-Cam based interpretability on different types of classification networks. A simple CNN was used as baseline model and two deep CNN (DCNN) ResNet50V2 and Xception were used for transfer learning. The CNN has a simple structure and its purpose is to discriminate between malignant and benign tumours. It consists of three two-dimensional convolutional layers for feature extraction with kernel size of 3 and ReLU as activation function. In between the convolutional layer, there is a max-pooling layer for down sampling the images and at the end of the convolutional layer, there is a flatten layer to flattened the input from two-dimensional tensor to one-dimensional tensor. After the flatten layer, there is a fully connected layer with hidden neurons of 64 and a three-way output layer with softmax as the activation function. Before the output layer, there is a dropout layer that randomly sets the input unit to 0 to help prevent overfitting. ResNet50V2 is a variant of ResNet50 model that has 48 convolutional layers with one max-pooling and average-pooling layer. It has the convolutional layers organized into a residual block which create a residual connection between each layer. The implementation of residual block could prevent the vanishing gradient problem. The residual block consists of a bottleneck layer with kernel size of 1x1 continued with the 3x3 convolutional layer and ended by a 1x1 convolutional layer. The bottleneck layer is used to reduce the computational cost of the network by minimizing the number of channels that are fed into the 3x3 convolutional layer. Xception instead uses modified depth-wise separable convolutional layers. It consists of a 1x1 pointwise convolutional layer and is followed by a 3x3 depth-wise convolutional layer. It has 36 convolutional layers whose purpose is feature extraction. The convolutional layers are organized as 14 modules. Between each module there is a linear residual connection. Transfer learning is a method that uses models that had been trained for a particular task and utilized its acquired knowledge to solved another task. In this case it is done using pre-trained models (ResNet50V2 and Xception) that have been trained using ImageNet dataset. After feature extraction layers have been added classifications layer consisting of one fully connected layer with 512 hidden neurons. The dataset used for training and testing the classifier has been taken from IQ-OTHNCCD lung cancer dataset. It is composed by 1190 image; 416 are normal CT slices, 120 are benign tumour images and 561 are malignant tumour images. All images are in JPEG format grayscale and have dimensions of 512x512.The splitting proportions of the dataset to create training, validation and test

sets are 80:10:10. All the images before being fed into the model have been pre-processed by resizing (224 x 224) and normalizing them. After the training of the network it has been applied the Grad-CAM technique whose block diagram is reported in Figure 4.12(A). The class activation maps has been plotted for each model. These heatmaps help localizing the section of the images with an high probability of the existence of tumour. In Figure 4.12(B) are represented the activation aps in form of heatmap superimposed with lung tumour CT slice. Resnet50V2 and Xception have a more defined focus on the targeted location than the simple CNN alone. The lack of focus presented by the CNN is due to the fact that the network is shallower. This implies that the network is not able to extract low level features on the image. Finally it is possible to notice that the class activation maps from the two transfer learning models were accurate in detecting abnormal lung features generated from the tumour presence. It is obviously not perfect but it is accurate enough to help radiologists to pinpoint the location of lung tumour [49].

From a practical point of view it is created a "sub-model" that maps the input image to the activations of the last convolutional layer. Then it is created another "sub-model" that maps the activations of the last convolutional layer to the last class prediction. Finally the gradient of the top predicted class for the input image is calculated with respect to the activations of the last convolutional layer. The

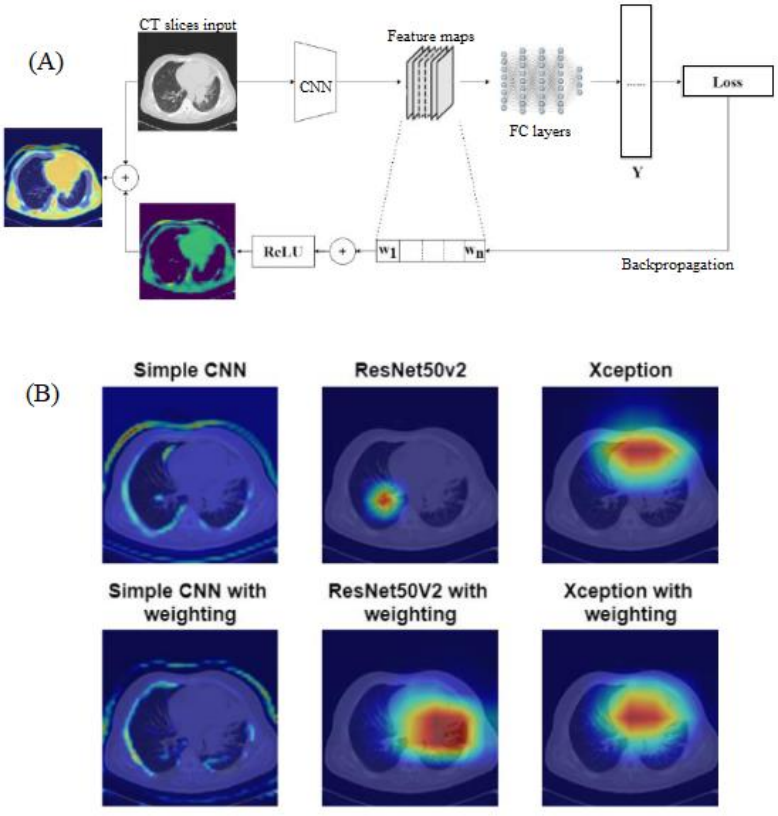

Figure 4.12. (A) Grad-CAM architecture. (B) Heatmap visualization over CT slices.

choice of mapping the activations of the last convolutional layer is based on some works found in literature. These assert that deeper representations in a CNN are able to capture higher-level visual constructs. Furthermore, convolutional layers retain spatial information which is lost in fully connected layers (typically used in class activation mapping (CAM)), so it is expected that the last convolutional layers has the best compromise between high-level semantics and detailed spatial information. The neurons in these layers look for semantic class-specific information in the image (object parts). Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision of interest. In order to obtain the class-discriminative localization map $L^c_{Grad-Cam}$ of width u and height v for any class c. First compute the gradient of the score for the class c, y^c with respect to feature map activations A^k of a convolutional layer as reported in the Equation (15):

$$Gradient = \frac{\partial y^c}{\partial A^k} \tag{15}$$

These gradients flowing back are global-average-pooled over the width and height dimensions (indexed by i and j respectively) to obtain the neuron importance weights αkc as reported in the Equation (16):

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{16}$$

This weight represents a partial linearization of the deep network downstream to the last layer and captures the 'importance' of the feature map k for a target class c. Finally it is performed a weighted combination of forward activation maps to get $LcGrad{-}CAM$ (Equation 17).

$$L^c_{Grad-Cam} = ReLU(\sum_k \alpha_k^c A^k) \tag{17}$$

This results in a coarse heatmap of the same size of the selected convolutional layer thus (8 x 8 x 512). The ReLU activation function has been applied to the linear combination of maps because it is needed to extract only those features that have a positive influence on the class of interest, for example the pixels whose intensity must be enhanced to increase y^c [50].

Practically it is generated the gradient of the top predicted class with respect to the output feature map of the last convolutional layer. After this it is created a vector where each entry is the mean intensity of the gradient over a specific channel of the map. Each channel is then multiplied by the importance weight α_k^c. Finally the channel wise mean of the resulting map is the heatmap associated to the class activation.

4.2.1.2 Interpretability when using computed tomography scans

As previously said, working with CT slices imply a loss of tumour spatial information. For this reason it is better to use CT scans. Also in this case, many studies in which Grad-Cam is applied to lung CT scans are available.

Here is presented the work proposed by Eali Stephen Neal Joshua et al.[51]. Here Grad-Cam++ is used together with squeeze and excite network (SENET) to provide a revolutionary method for differentiating malignant from benign lung nodules on CT scans. The new SENET Grad Cam++ module, which combines the features calibration and discrimination benefits of SENET, has been shown to have a substantial potential for improving feature discriminability in lung cancer classification. The proposed technique is divided into three major stages:

1. Image data gaining from the "LUNA16" database.
2. Nodule Cataloguing using Squeeze and Excitation extraction of nodules.
3. Feeding classified image to Gard-Cam++ Activation function for the future classification.

This study used the Lung Image Database Consortium image collection (LIDC–IDRI). Thin-slice CT scans are optional by the American College of Radiology for nodule identification and categorization. After some dataset adjustments like rejecting scans with missing slices, a total of 888 CT images with 1004 nodules has been obtained. To ensure that the learning model is correctly trained, 3D patches (32 x 32 x 32) containing lung nodules are clipped from raw CT images and inserted in the network to serve as training examples. SENETS are a type of CNN that, despite having limited computational power, increases channel interdependence. Through the use of channel adjustments for each conversion block, it has been possible to fine-tune the weighting on each feature map. Using these filters, CNNs can extract hierarchical information from images. Higher layers are capable of detecting complex geometric shapes, whereas lower layers are only capable of detecting borders of the tumour. They gather all of the information necessary to complete a task in a short period of time. This is accomplished through the combination of space and image data. Following that, the filters will combine data from all of the output channels that are currently available. When the network generates output charts, it assigns the same weight to each of the channels. SENET intends to alter the channel weighting by incorporating a content-aware mechanism into its network. In CNNs SENETS are a novel kind of building component that enhances channel dependence while requiring low computational effort. The basic thing is that SENET requires to have convolutional blocks with parameters assigned to each of the channels.

This allows the network to adjust the weighting of each feature map as necessary. When simple CNNs analyse images, convolutional filters are used to collect hierarchical information contained within the images. Higher degrees of intelligence may be able to recognise complex geometrical patterns that are undetectable to lesser levels of intelligence. For this reason Eali Stephen Neal Joshua et al.[41] proposed a model CNN model which integrates SENET (Figure 4.13). This solution allows to merge the spatial and channel data contained inside a single image. In this configuration the different filters will search for spatial components in each individual input channel before aggregating the data over all available output channels to get a final conclusion. When the network is constructing the output feature maps, it provides equal weight to each of the channels that it has. SENETS is a revolutionary network that includes a content-aware technique for adaptively weighing each channel. A single parameter is assigned to each channel and processed as if it were a linear scalar value. For each channel feature maps are considered as a single numerical value, which is then used to get a global understanding of that channel. This results in an n-dimensional vector, where n is the number of convolutional channels. Data is then fed into a two layer neural network, which generates a result with a size equal to or greater than that of the input vector. Now that the n values have been determined, they may be used as weights on the original feature maps, allowing for accurate scaling of each channel. The method proposed here is used to offer visual explanations by highlighting discriminative areas in the model, which is a very strong strategy for finding out how to make the model interpretable. Essentially, has been developed a 3D CNN for the classification of the lung nodule into malignant or benign which simultaneously provides visual insights into the model's decision-making processes. This study proposes Grad-CAM++ techniques for visualization based on a single module categorization [51].

This technique is an improvement of Grad-CAM. It can provide better visual explanations of CNN model predictions, in terms of better object localization as well as explaining occurrences of multiple object instances in a single image, when compared to state-of-the-art [52].

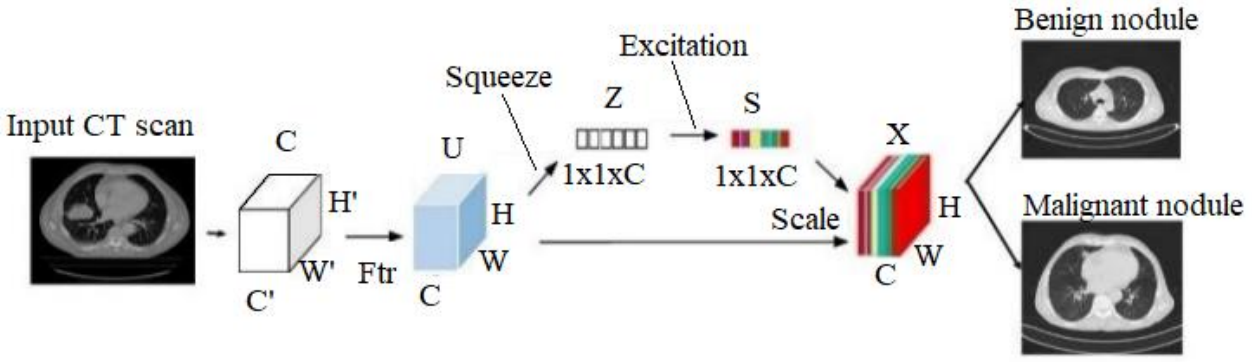

Figure 4.13 Representation of SENET integrated in CNN model.

Grad-CAM++ is supported by a large number of different CNN models. It generates a visual image that distinguishes between different classes of people. It also informs you of the model's failure modes, should they occur. To determine the importance of each neuron for a particular decision, Grad-CAM++ examines the information on the gradient that is passed into the final convolutional layer of the CNN, which is called the final convolutional layer [51].

5. An interpretable approach for non-small cell lung cancer computer-aided diagnosis from computed tomography scans

5.1 Preliminary experiment

This thesis is divided in two implementation sections according to the data used. The first preliminary experiment consists of the use of CT scans of whole lungs with a reduced dimension and a reduced resolution. Indeed for each scan only 10 slices are considered.

5.1.1 Data selection and pre-processing

The model is fed with 100 3D CT scans (50 ADC and 50 SCC) of whole lungs. These have been taken from the work proposed by Selene Tomassini et al. [53].

Data were retrieved by the openly-accessible NSCLC-Radiomics dataset contained in TCIA. All the scans belonged to anonymized subjects ranging from 45 to 88 years in age. Among them, 30% were females and 70% were males. The CT scans were pre-processed by following a workflow which aims to remove all disturbing information. Firstly, HU scale conversion was accomplished to describe values contained in each voxel, considering the intensity window from -1024 HU to 400 HU. Next, 1-mm resampling was performed to make slices within scans spatially homogeneous. Then has been paid attention also to the inter-slice distance z. In particular, when z was bigger than 1, some artificial slices were generated by interpolation; whereas, when z was smaller than 1, some slices were removed. Another important step is the intensity normalization to soften the intensity variation caused by the use of different scanners or scanning parameters during the acquisition. In the obtained image all the pixel values ranged between 0 and 1. After having accomplished all these passages needed for the optimization of the image, has been performed the automatic lung parenchyma segmentation for the elimination of non-lung tissues by exploiting a pre-trained UNet-like model, named UNet (R231). Then, each scan was automatically cropped by removing all black, non-informative voxels and also the non-informative slices were automatically cut from each scan. Finally, all scans were resized to 250 pixels × 200 pixels × 250 pixels, which was the average shape of the cropped-and-cut scans [53].

In the current experiment instead, the dimensions of CT scans used for training and testing the network have been reduced to 10 pixels × 128 pixels × 128 pixels. In a total of 100 CT scans constituted in equal amount by ADC and SCC scans, the 80% of them has been used for creating a training set and the remaining 20% a test set. In particular, the Stratified Shuffle Solit cross-

validator has been used here. It provide train/test indices to split data in train/test sets for each of the 5 splits. Together with the images, the model must receive in input one label for each CT scan because this is a supervised learning. These labels are contained in an array composed by 100 rows and 2 columns. CT scans containing ADC are labelled with number 1 in first column instead CT scans containing SCC are labelled with number 1 in the second column.

5.1.2 Model structure

The model proposed in this thesis aims to discriminate the main NSCLC histotypes (ADC and SCC) from pre-processed 3D CT whole-lung scans taken from the study of Selene Tomassini et al. [53].

It is fed with the 100 3D CT scans described in the previous section and has been developed in Google Colab cloud service with a 12 GB RAM. The presented model is a sequential model and its structure can be seen in Figure 5.1. The sequential model is a linear stack of layers where each layer has just one input tensor and one output tensor.

The model begins with a time distributed CNN (VGG16 trained by scratch) layer. It is a wrapper that allows to apply a layer to every temporal slice of an input. Every input should be at least 3D, and the index one of the first input tensor will be considered to be the temporal dimension [54].

In this case the input is a numpy array of shape (100,10,128,128,1), this means that it is considered a batch of 100 CT scans composed by 10 slices each with a dimension of 128 pixels x 128 pixels as can be seen in Figure 5.2. 1 is the number of channels of the image which is in grayscale. This layer is used to apply the same convolutional model, which in this case is the VGG16, to each of the 10 slices independently. VGG16 is a CNN model proposed by Karen Simonyan and Andrew Zisserman at the Oxford University. They submitted the actual model in the 2014 ImageNet Challenge. It is composed by a stack of multiple (usually 1, 2, or 3) convolution layers of filter

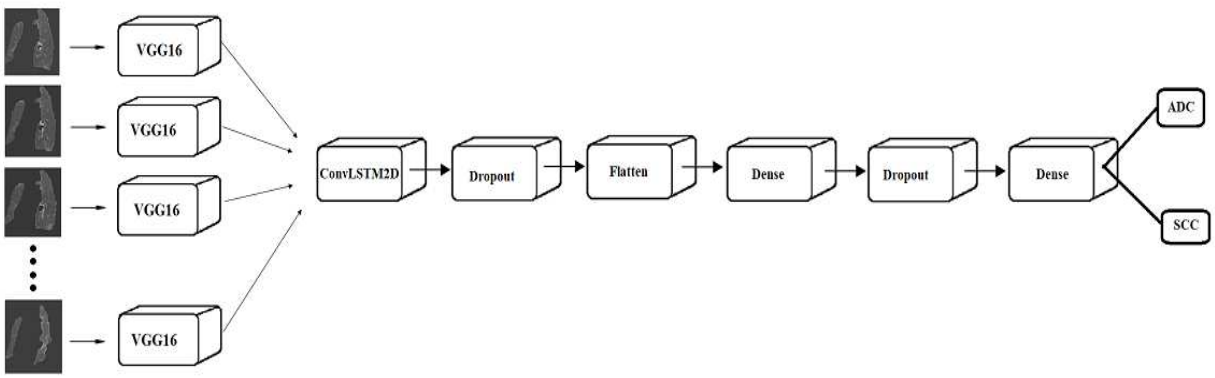

Figure 5.1. Representation of proposed model.

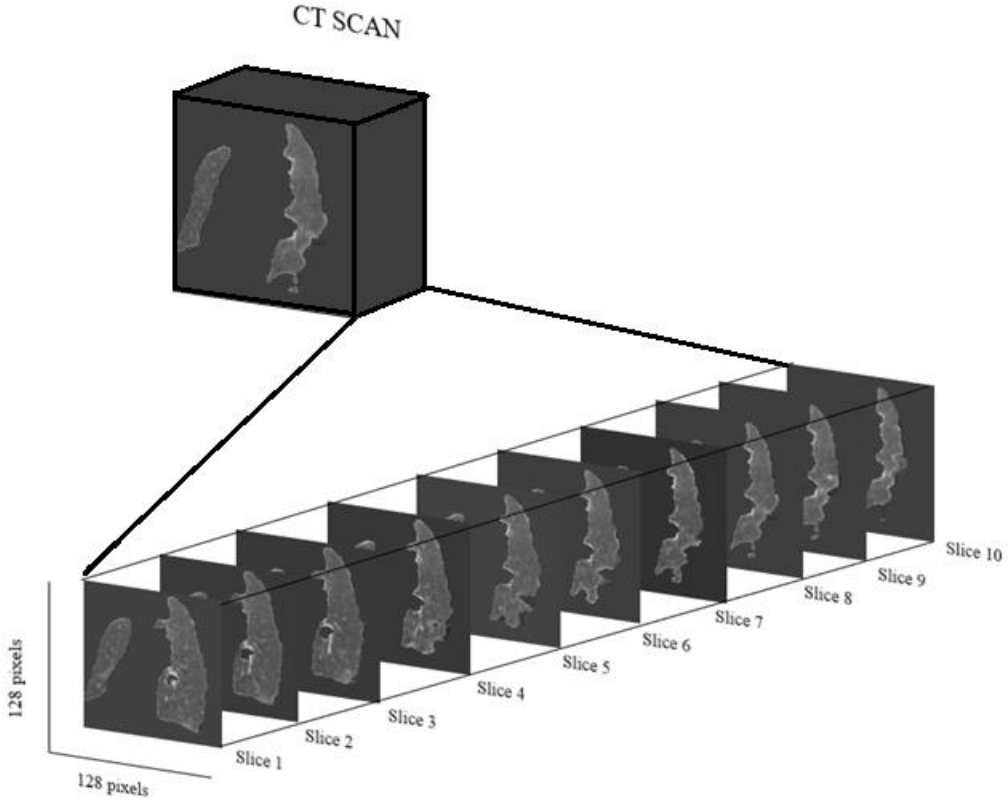

Figure 5.2. Representation of a CT scan composed by 10 slices of dimensions 128 pixels x 128 pixels.

size 3 x 3, stride one, and padding one, followed by a max-pooling layer of size 2 x 2. Different configurations of this stack were repeated in the network to achieve different depths. The convolution stacks are followed by three fully connected layers at the end, two with size 4'096 and one with size 1'000. The last one is the output layer with Softmax activation. The size of 1'000 refers to the total number of possible classes in ImageNet [55].

After the time-distributed CNN layer, there is a 2D Convolutional Long-Short Term Memory (ConvLSTM2D) layer. It is a special kind of recurrent neural network to handle long-term dependencies that allows information to persist [56].

The choice of using ConvLSTM2D here is linked to the fact that CT scans are composed by a series of single images (slices) collected over successive periods of time. Thus, CT scans can be considered to be time series. The strength of this recurrent neural network architecture is that it passes the previous hidden state to the next step of the sequence. Internally can be found 2D convolutional operations which allow to work with 3D data [57].

The layer analyses each CT slice "remembering" also what has already seen in the previously processed images. The parameters that need to be passed by the user to this layer are the kernel size and the number of filters.

Then, there is a Dropout layer, which randomly omit or drop some of the neurons in hidden or visible layers. It helps in regularizing the neural network making the model more robust to avoid overfitting. The dropout layer requires the dropout rate as parameter to be selected by the user. It

is a number comprised between 0 and 1 that represents the fraction of the input unit to be dropped [58].

The Flatten layer which follows has the function to flatten the input without affecting the batch size. After the Flatten layer, there is a Dense layer. It is commonly used in the final stages of the neural network. It helps in changing the dimensionality of the output from the preceding layer so that the model can easily define the relationship between the values of the data given as input to the model. The Dense layer is deeply connected with its preceding layer; this means that the neurons of this layer are connected to every neuron of the previous layer. Every neuron receives as input the outputs coming from every neuron of the preceding layer and a matrix vector multiplication is performed. This layer requires to specify the units number as parameter. It defines the dimensionality of the output vector exiting from the layer, thus it must be a positive number. Also the activation function to be used is a parameter that must be specified by the user. This is the function that will be utilized for the transformation of values given as input to neurons. It introduces the non-linearity for letting the network learn the existing relationship between input and output values. In this case has been chosen the ReLU activation function [59].

Then, there is another Dropout layer and, finally, again a Dense layer with Softmax activation function. In Figure. 5.3 can be seen the architecture of the model together with the input and output dimensions of each layer. Here is possible to notice how the input dimension is progressively reduced. On the top there is the time distributed input which is a layer that simply takes the CT scans of dimensions (None,10,128,128,1), where None is the batch size. Then the time distributed (VGG16) layer has as output a numpy array of dimension (None,10,4,4,512) which is the size of the output of the last layer in VGG16 which is a MaxPooling2D. Then in the successive layers there is a progressive modification of the array's dimensions until arriving at the last layer where the output dimension is (None,2). The number 2 has been chosen because the model has to classify the images previously given in input in one of the two class: ADC or SCC. Together with the models' layers and relative parameters it must also be selected the loss function and the optimizer for compiling the model. In this case the loss function used is the categorical cross entropy which is well suited to classification tasks. The chosen optimizer, that aims to reduce as much as possible the error function, is the stochastic gradient descent (SGD). It converges to the global minimum with a forward and backward propagation for every record. The selected metrics for the prediction is the accuracy.

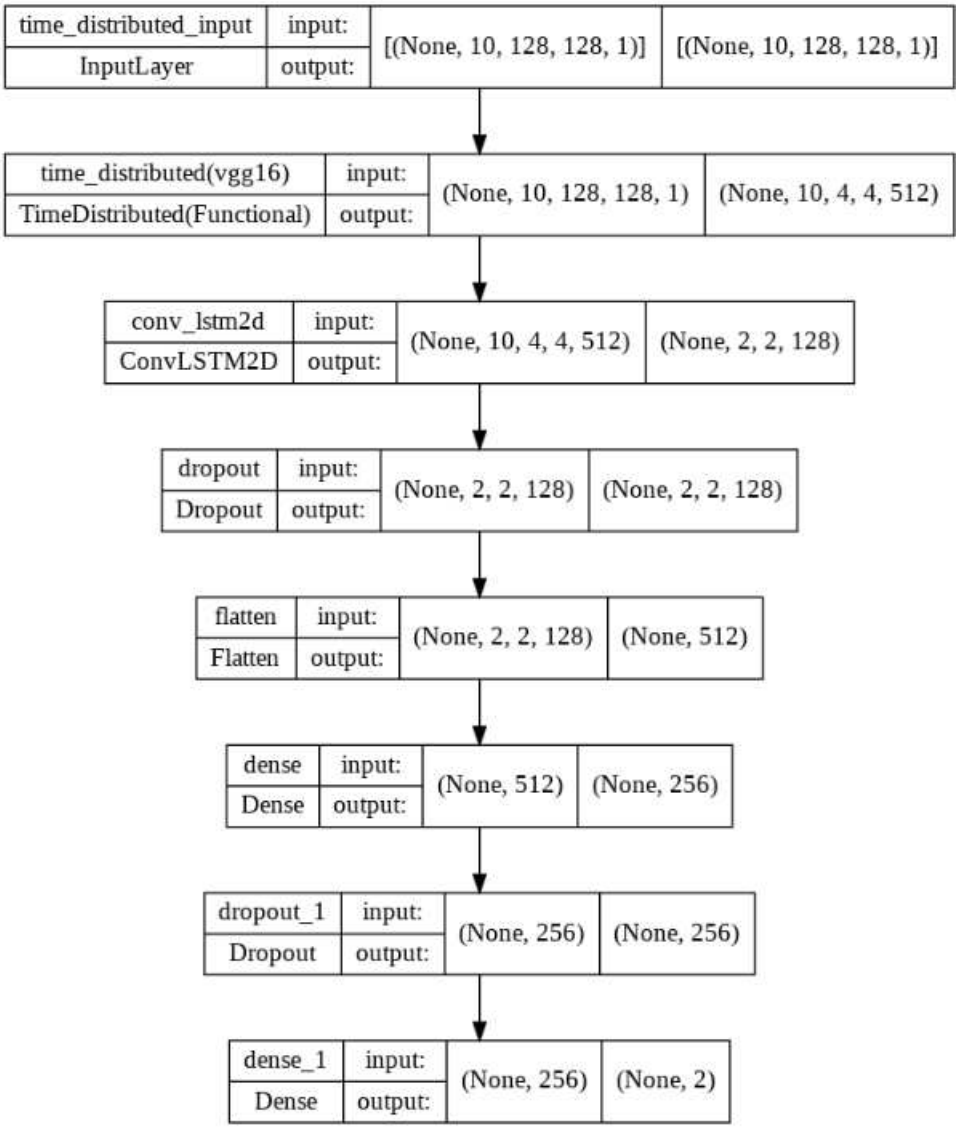

Figure 5.3. Representation of the proposed model together with input and output dimensions associated to each layer.

5.1.3 Hyperparameter tuning

A DL model is a mathematical model with a certain number of hyperparameters which are parameters that cannot be directly learned from the training process. These parameters express important properties of the model like its complexity or how fast it should learn [60].

The first purpose of this work is to tune the main model hyperparameters: the filter dimension of the second layer (ConvLSTM2D), the dropout rate of the two Dropout layers, the learning rate of the optimizer and the batch size. The last one is one of the most important parameters as it defines the number of samples to work through before updating the internal model parameters [61].

Hyperparameter tuning is performed to find those parameters that let the model achieve the best performances possible. Here, the hyperparameter tuning has been accomplished by exploiting the Gaussian Process minimization (GP_minimize). This can be considered as a black-box which takes the hyperparameters as inputs. Inside this black-box some performance metrics are used to reduce the regression error in order to have a better model. The GP minimization is nowadays very attractive thanks to the increasing computational power together with a reduced cost. It is an algorithm that maps hyperparameter values to a number that indicates how well the model fits the data with the currently tested hyperparameter value. Then, using a Gaussian process regression, the confidence boundaries between the mapped values are computed. These boundaries indicate an estimated range of outcomes that exists between the already evaluated hyperparameters. In particular, the application of this method to the proposed model, requires steering the algorithm to search through a certain space by initializing the exploration with few random observations. Then the algorithm searches and estimates the confidence intervals for the hyperparameter value that has the highest potential to decrease the estimation error. Firstly has been created an objective function with which evaluate the goodness of the fit. In this case it is the model itself which is dependent on the five hyperparameters that need to be tuned. This function, at each step of the process, returns the estimation error associated to the hyperparameters values tested. Then it has been defined the searching space for every hyperparameter. It is the range in which the optimal hyperparameter values should be searched. The values which constitute the searching space of each hyperparameter, and the type of these data are reported in Table IV. Then, GP_minimize evaluates a certain number of random points within the selected searching space by passing them to the objective function [62].

The amount of random points to be evaluated must be selected by the user; in this case the number of hyperparameters values combination to be tested is 20. On the 20 random points a Gaussian process regression is performed, which returns the estimated confidence bounds. In these regions, the algorithm looks for the hyperparameter values that are expected to improve the most the

Table IV. Data types and values of hyperparameters searching space.

Hyperparameter	Searching space data type	Searching space values
Filter	Categorical	[2,4,16,32,64,128,256]
Dropout rate 1	Categorical	[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8]
Dropout rate 2	Categorical	[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8]
Learning rate	Categorical	[1e-6,1e-5,1e-4,1e-3,1e-2]
Batch size	Integer	Low=5 / High=256

accuracy achieved by the model and to reduce as much as possible the loss function. In practice for each set of the five hyperparameters to tune, the model is trained with four epochs and then validated, respectively with a train and validation set constituted by the already cited ADC and SCC CT scans. For each epoch is performed only 1 step for the training process and 1 step for the validation process. The number of steps for training and validation to be used are calculated according to the Equation (18) and (19) respectively:

$$Training\ steps = \frac{Train\ length}{Batch\ size} \tag{18}$$

$$Validation\ steps = \frac{Validation\ length}{Batch\ size} \tag{19}$$

Where train length is the number of scans in the training set and validation length the number of scans in the validation set. GP_minimize performs twenty "calls", so it evaluates twenty different sets of hyperparameters and in output returns the loss and the accuracy achieved by the model during the training phase. In addition, it also returns the validation loss and the validation accuracy associated to the validation phase. The best hyperparameters set is selected according to the accuracy value achieved by the model in classifying the CT scans contained in the training set. Obviously the higher is the accuracy and the better is the model. After the hyperparameter tuning the model has been constructed with the values that provide the best CT scans classification accuracy. The model is then compiled and trained. The training process consists of 5 splits of 50 epochs each. The early stopping has been introduced to regularize the model; it stops the training when the validation error reaches a minimum.

5.1.4 Model evaluation

As previously said, the model performances must be evaluated on test set to understand how well it classifies the CT scans given in input. Here are taken into consideration:

- Loss.
- Accuracy.
- Sensitivity.
- Specificity.
- F1-scores.
- ROC curve AUC.
- Confusion matrices.

In particular, the metric values listed above for evaluating this classification model have been obtained for each of the five splits. Thus to evaluate the overall performances of the model has

been considered the average values among all the splits. In DL, as previously said, the loss is the function that the neural network try to minimize. It is the difference between the predicted value and the actual value. In this case the model is a classifier, so losses are expressed as probability because the predicted classes are based on probability. Here, losses have been calculated by exploiting the binary cross entropy loss function (Equation (11)).

The accuracy is the fraction of predictions the model got right. For binary classification, accuracy can also be calculated in terms of positives and negatives as reported in Equation (20):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

where TP are the true positives, TN are the true negatives, FP are the false positives and FN the false negatives. The measure of the accuracy is usually provided as percentage [66].

The sensitivity is a statistical index of diagnostic accuracy, also called true positive rate. It is defined as the ability of the model to correctly identify all the TP. The sensitivity is calculated with the Equation (21):

$$Sensitivity = \frac{TP}{TP + FN} \tag{21}$$

The specificity instead is the ability of the test to correctly identify all the TN. It is also called true negative rate and it is calculated with the Equation (22) [67]:

$$Specificity = \frac{TN}{TN + FP} \tag{22}$$

The F1 score is another ML metric that can be used in classification model. It is a proposed improvement of two simpler performance metrics. It is defined as the harmonic mean of precision (Equation 23) and recall whose expression equals the one of sensitivity (Equation 21) [68].

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

Finally the receiving operating characteristic (ROC) curve is a graph showing the performances of a classification model at different classification thresholds. This curve plots two parameters: true positive rate (TPR) which is equal to the sensitivity and false positive rate (FPR) which is defined with the Equation (24).

$$FPR = \frac{FP}{TN + FP} \tag{24}$$

To compute the points in an ROC curve, it is needed to evaluate a logistic regression model many times with different classification thresholds, but this would be inefficient. So an efficient, sorting-based algorithm that can provide this information is the area under the ROC curve (AUC). It

measures the 2D area under the entire ROC curve. AUC provides an aggregate measure of performance across all possible classification thresholds [69].

Finally the confusion matrix is used for evaluating the performance of a classification model. It is a N x N matrix, where N is the number of target classes. The matrix compare the values predicted by the model to the actual values. A representation of the confusion matrix is reported in Figure 5.4. It is a summary of the number of correct and incorrect predictions made by the classifier. A good model is the one which has high TP and TN rates together with low FP and FN rates.

		Predicted	
		Negative **(N)**	Positive **(P)**
Actual	Negative	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive	**False Negative (FN) Type II Error**	True Positive **(TP)**

Figure 5.4. Confusion matrix structure.

5.1.5 Model interpretability

The model is now able to classify the scans given in input by discriminating between ADC and SCC lung tumour cytotype with a certain accuracy. However, the CNN architectures are a sort of "black-box"; only the input and the output are known to the user. It is not possible to decode the decision made by the model. For instance if give in input to the model the CT scan reported in Figure 5.2, the output generated by the classifier is an array containing numbers between 0 and 1 arranged in one row and two columns. This represents the prediction made by the model. The number in the first column is the probability that the image is a CT slice containing ADC thus, the more this number is near to 1 and the higher is the probability that the CT scans contains ADC. In turn the number in the second column refer to the probability of having CT slices containing SCC. The more this number is near to 1 and the higher is the probability that the CT scans contains SCC. The problem is that it is not known how the model predicts a certain label (ADC or SCC). This is not acceptable in clinical settings because making diagnosis is high-risk. Hence, it is necessary to explain how the model performs the classification task. From here the necessity of introducing the model interpretability which consists of all those techniques which bring transparency in the model decisions. Decoding decisions is also important for investigating which are the features of the input image which had a major contribution in producing that decision [68].

Gradient-based methodologies are the most popular interpretability techniques that have been used to generate explainable predictions for medical image analysis tasks. In this work has been used the Gradient-weighted Class Activation Mapping (Grad-CAM) which manipulates gradients to highlight the pixels that contribute most to the prediction. It generates heatmaps to identify the region of importance in input images towards the model's final predictions. Visual explanations provide information about how the input CT is stored in the internal layers of the network and what strategy the network utilized to identify the tumour regions [69].

In the case of image classification a 'good' visual explanation from the model for justifying any target category should be:

(a) Class discriminative (i.e. localize the category in the image).

(b) High-resolution (i.e. capture fine-grained detail).

Pixel-space gradient visualizations such as Guided Backpropagation and Deconvolution are high resolution and highlight fine-grained details in the image, but are not class-discriminative. In contrast, localization approaches like CAM and Grad-CAM, are highly class-discriminative. This means that the tumour explanation exclusively highlights the tumour regions. Grad-CAM is

applicable to a wide variety of CNN model-families like CNNs with fully connected layers (e.g. VGG16) [61].

However it has some problems in working with models containing LSTM. For this reason the model previously presented has been modified a little bit. The time distributed layer and the ConvLSTM2D have been removed. The new model (Figure 5.5) has been built by "opening" the VGG16 and putting in series the layers taken by the original model (excluding the time distributed layer and the ConvLSTM2D layer). Between the last layer of the "opened VGG16" and the first layer of the original model have been inserted one additional convolutional layer and one average Pooling layer. These two additional layers are needed to match the array dimensions. In order to "maintain a connection" with the original model the weights of the two Dense layers have been taken by the old model and applied to the Dense layers in this new model. To verify that the new model performs as well as the previous one, it has been evaluated with the validation set. The loss and accuracy achieved by the new model are comparable and even better than the old ones. Also the input shape of the model has been changed and put equal to (128,128,1). This means that the model is fed not anymore with CT scans but with single CT slices grayscale of dimensions 128 pixels x 128 pixels. So the model now classify one slice at a time instead of ten slices together as in the previous model. In order to generate heatmaps has been firstly created a "sub-model" that maps the input image to the activations of the last convolutional layer. Secondly has been created another "sub-model" that maps the activations of the last convolutional layer to the last class

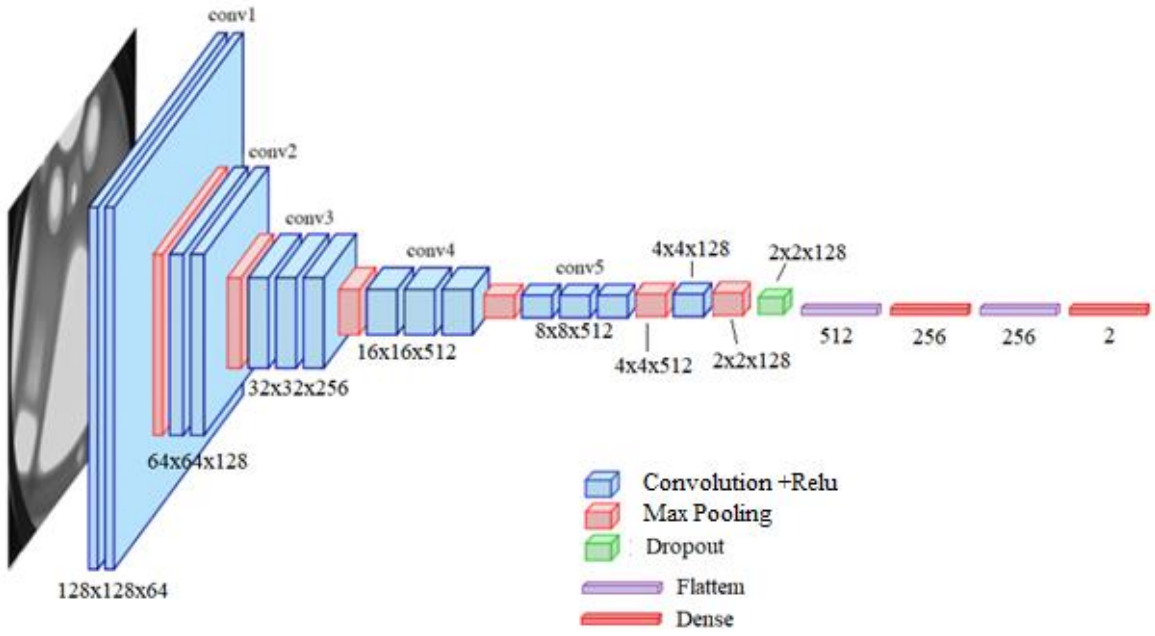

Figure 5.5. New model structure.

prediction. In this case the prediction layer is the last dense layer. Finally has been computed the gradient of the top predicted class for the input image with respect to the activations of the last convolutional layer (as reported in Figure 5.6). The heatmap has then been normalized between 0 and 1 for visualization purposes. Consequently it has also been colorized by using the jet colormap resulting in a RGB heatmap. In order to see where the network "focus its attention" for tumour cytotype classification, it is necessary to superimpose this heatmap to the CT slice analysed by the model. For this reason the heatmap is resized to equal the dimension of the CT slice which is [128,128]. In this work the dimensions of the data given in input to the model is [250,128,128,1]; this means that it will be generated one heatmap for each of the 250 slices which constitute one whole lung CT lung scan. These 250 heatmaps are then superimposed to the 250 CT slices. The obtained superimposed images are consequently used as frames for creating a video. This displays dynamically where the network focuses its attention for the tumour classification. The chosen frames velocity is 30 frames per second (fps).

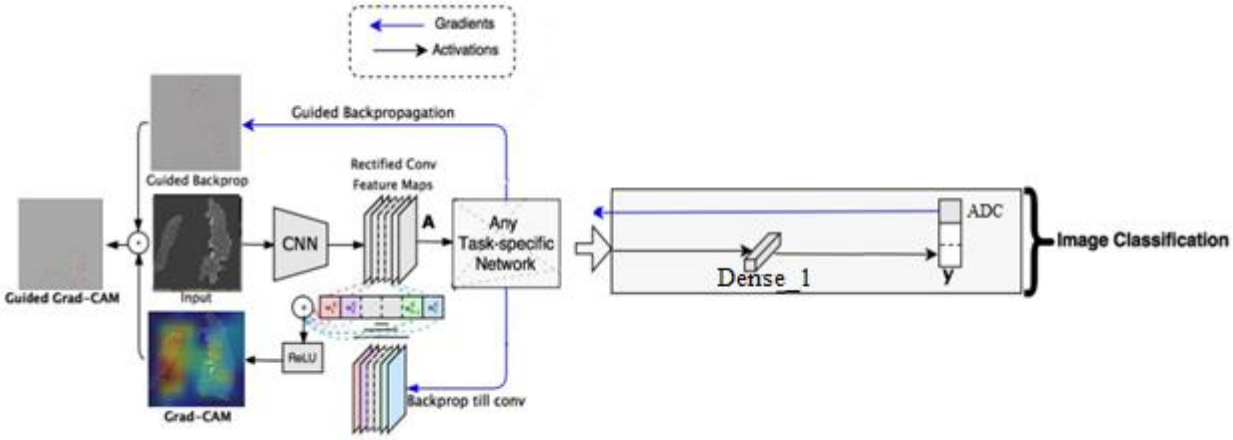

Figure 5.6. Grad-CAM algorithm schematization.

5.2 Final experiment

For the final experiment, the steps performed are the same of the preliminary experiment. The difference is that here are used the original CT scans with an higher number of slices for each scan and an higher resolution.

5.2.1 Data selection and pre-processing

The model is fed with 100 3D CT scans (50 ADC and 50 SCC) of whole lungs. These have been taken from the work proposed by Selene Tomassini et al. [53].

The main difference with respect to the preliminary experiment is that here are used the original data with an higher resolution of 250 pixels × 200 pixels × 250 pixels. Once pre-processed, 3D CT whole-lung scans were divided in train and test sets with a split ratio of 60:40, and 40% of the train set served as the validation set. Furthermore, to face the scarcity of train data and mitigate overfitting, the train set was quintupled through data augmentation, consisting in 15° left/right rotation and random in/out zooming from 0.8 to 1.2 [53].

Together with the images, the model must receive in input one label for each CT scan because this is a supervised learning. These labels are contained in an array composed by 100 rows and 2 columns. CT scans containing ADC are labelled with number 1 in first column instead CT scans containing SCC are labelled with number 1 in the second column.

5.2.2 Model structure

The model structure used for the final experiment has been developed in Google Colab Pro cloud service with a maximum RAM of 34 Gb. In addition the GPU has been used as hardware accelerator. All of this is necessary because, even if the model is the same of the one used for the preliminary experiment (reported in section 5.1.2), the data given as input are much “heavier”. The only difference in the model is that the input layer has been changed because the dimensions of the CT scans have been modified. Now the network has an input layer of dimensions (250,200,250,1) where each scan is composed by 250 slices each of dimension 250 pixels x 200 pixels. The model structure is reported in Figure 5.7. As a consequence of the modifications of input dimensions, also the number of parameters characterizing the model have changed. In the preliminary experiment there were 17,729,090 parameters (3,015,554 trainable parameters and 14,713,536 non-trainable parameters). Now instead, for what concern the model used for the final experiments, there are much more parameters: 23,103,968 (8,390,402 trainable parameters and 14,713,536 non-trainable parameters).

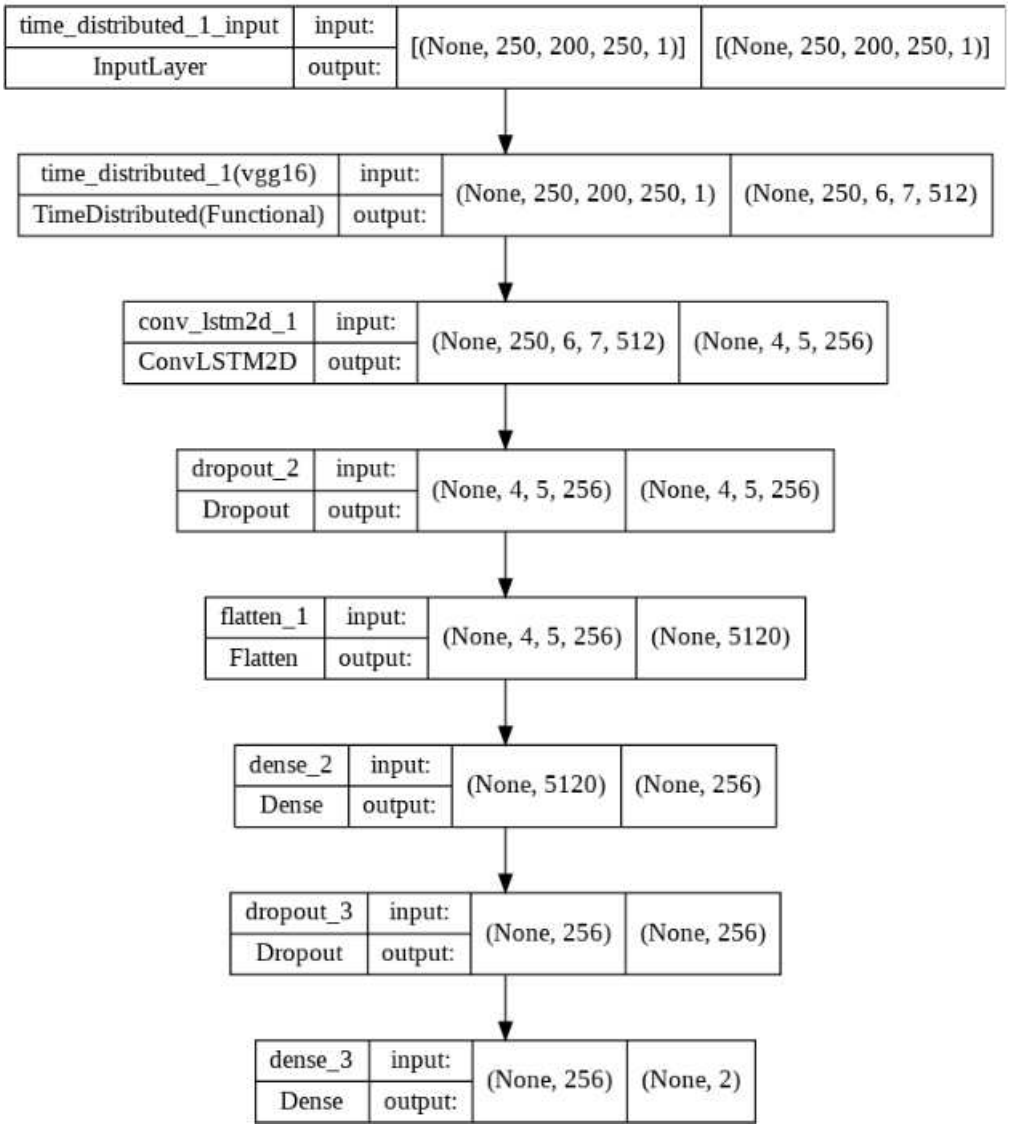

Figure 5.7. Model structure of the final experiment.

However this is very heavy from a computational point of view, thus all the layers belonging to the VGG16 has been “freeze”. This means that the parameters linked to the VGG16 are not trained. Consequently the number of total parameters is drastically reduced up to 14,753,370 (39,834 trainable parameters and 14,713,536 non trainable parameters). This is a very common practice when dealing with complex network. Often the “freezing” of layers is coupled with data augmentation.

5.2.3 Hyperparameter tuning

The parameters to be tuned in the final experiment are the same ones of the preliminary experiment except for the batch size. The batch size has been set to 1 in order to avoid the out of memory problem (OOM), linked to the higher number of data (with higher resolution) used. Also in this case the hyperparameter tuning has been accomplished by exploiting the GP_minimize. The values which constitute the searching space of each hyperparameter, and the type of these data are reported in Table V. These are the same values used in the preliminary experiment but the batch size has been excluded. In this case the number of hyperparameters values combination to be tested is 40. On the 40 random points a Gaussian process regression is performed, which returns the estimated confidence bounds. In practice for each set of the four hyperparameters to tune, the model is trained with four epochs and then validated, respectively with a train and validation set constituted by the original ADC and SCC CT scans. GP_minimize performs 40 "calls", so it evaluates 40 different sets of hyperparameters and in output returns the loss and the accuracy achieved by the model during the training phase. In addition, it also returns the validation loss and the validation accuracy associated to the validation phase. The best hyperparameters set is selected according to the accuracy value achieved by the model in classifying the CT scans contained in the training set. Obviously the higher is the accuracy and the better is the model. After the hyperparameter tuning the model has been constructed with the values that provide the best CT scans classification accuracy. The model is then compiled and trained. However in this final experiment the training data (and also the testing and validating data) are given to the model in form of batches. This is done because the data used here are a lot and with an higher resolution. The training process consists of just one split of 50 epochs each with 200 training steps and 20 validating steps according to the Equation (18) and (19) respectively. The choice of performing one split is to avoid overfitting problems. Also here the early stopping has been introduced to regularize the model and to stop the training when the validation error reaches a minimum. In the

Table V. Data types and values of hyperparameters searching space.

Hyperparameter	Searching space data type	Searching space values
Filter	Categorical	[2,4,16,32,64,128,256]
Dropout rate 1	Categorical	[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8]
Dropout rate 2	Categorical	[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8]
Learning rate	Categorical	[1e-6,1e-5,1e-4,1e-3,1e-2]

final experiment the test set, not present in the preliminary experiment, is exploited for performing the model prediction.

5.2.4 Model evaluation

The model performances are evaluated by taking into account the same metrics proposed in Section 5.1.4:

- Loss.
- Accuracy.
- Sensitivity.
- Specificity.
- F1-scores.
- ROC curve and AUC.
- Confusion matrices.

In addition here also another metric has been introduced; the recall. It is the measure of how the model correctly identify the TP, in fact it is also called true positive rate. Recall is calculated according to the Equation (21) used for the sensitivity.

5.2.5 Model interpretability

The steps performed for the model interpretability of the final experiment are the same ones reported in Section 5.1.5 regarding the preliminary experiment. The model used has the same structure reported in Figure 5.5. The only difference is that the input layer's shape has been modified according to the dimensions of the data used here (as written in Section 5.2.1). For the final experiment the gradient maps are generated for both the ADC scan and the SCC scan in order to evaluate the differences between them.

6 Results

6.1 Preliminary experiment

The results reported in this section refer to the preliminary experiment.

6.1.1 Hyperparameter tuning

The results of the hyperparameter tuning are reported in Table VI. From here has been selected the set of hyperparameters which provide the best model performances. In this case it corresponds to the call number 14, associated to an accuracy of 56.25%.

Table VI. Hyperparameter tuning results. Dropout1_rate is the dropout rate associated to the first Dropout layer. Dropout2_rate is the dropout rate associated to the second Dropout layer.

Call number	Filter	Dropout1_rate	Dropout2_rate	Learning rate	Batch_size	Accuracy
0	32	0.5	0.5	0.000100	10	50.00%
1	64	0.4	0.7	0.010000	180	50.00%
2	16	0.7	0.4	0.000100	105	50.00%
3	128	0.2	0.3	0.000010	176	50.00%
4	32	0.4	0.2	0.000001	206	50.00%
5	64	0.7	0.5	0.001000	90	50.00%
6	32	0.7	0.4	0.000001	55	50.00%
7	4	0.6	0.7	0.001000	163	50.00%
8	64	0.8	0.2	0.000010	141	50.00%
9	128	0.7	0.8	0.001000	250	50.00%
10	64	0.8	0.7	0.001000	117	50.00%
11	2	0.4	0.8	0.000010	5	50.00%
12	256	0.1	0.7	0.000100	226	43.75%
13	256	0.1	0.8	0.000001	21	43.75%
14	128	0.4	0.8	0.001000	223	56.25%
15	256	0.4	0.8	0.001000	184	50.00%
16	32	0.8	0.4	0.000010	226	50.00%
17	128	0.1	0.8	0.001000	167	50.00%
18	128	0.6	0.3	0.010000	212	50.00%
19	64	0.4	0.8	0.001000	246	50.00%

6.1.2 Model evaluation

The parameter's values retrieved from the previous section have been utilized for constructing the model. This has been compiled and trained achieving the performances reported in Table VII. In Figure 6.1 it is possible to see the ROC curve together with the AUC values. Here are reported the ROC curves and the AUC values associated to each split. Finally also the mean ROC curve and the mean AUC values are displayed. In Figure 6.2 are reported the five confusion matrices relative to each of the splits performed during the training process.

Table VII. Model training performances. In the last raw are reported the averaged values (avg.) over the 5 splits.

Split	Losses %	Accuracies %	Sensitivities %	Specificities %	F1_scores %
0	69.33	44.99	0.00	90.00	0.00
1	69.34	50.00	100.00	0.00	66.67
2	69.30	50.00	100.00	0.00	66.67
3	69.27	60.00	20.00	100.00	33.33
4	69.17	50.00	0.00	100.00	0.00
Avg.	66.67±0.01	51.00±4.89	44.00±46.30	58.00 ±47.50	33.33±29.81

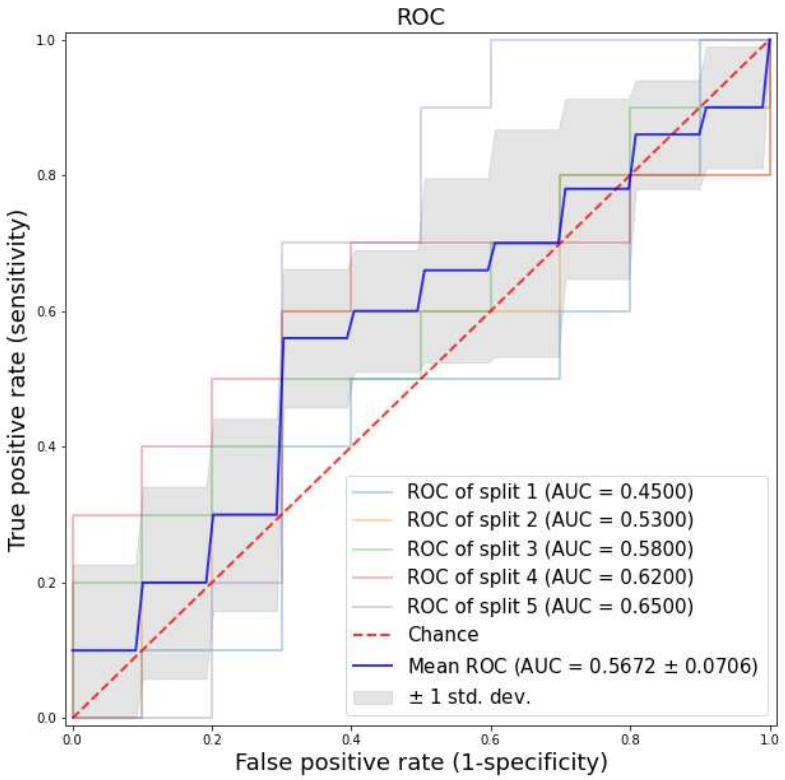

Figure 6.1. ROC curves and AUC values.

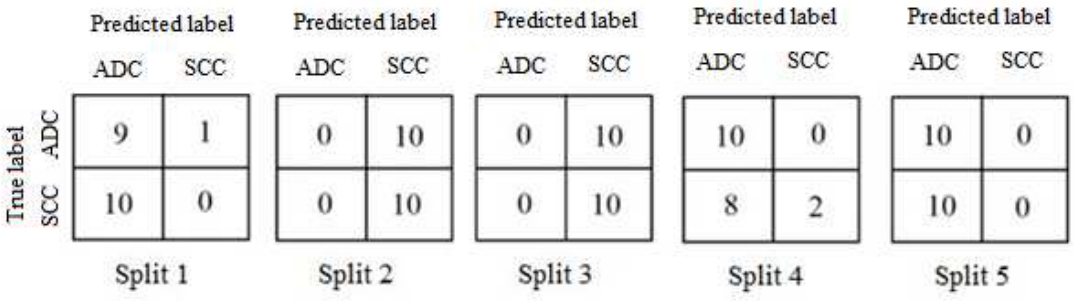

Figure 6.2. Representation of confusion matrices.

6.1.3 Model interpretability

As reported in section 5.1.5, it has been created the new model which achieved an avg. loss of 69.22%. This model has then been exploited for generating the heatmaps which have been then superimposed to the CT slices. Some of the most explicative frames of the generated video for the ADC scan are reported in Figure 6.3. In Figure 6.4 instead can be observed some frames of the generated video for the SCC scan.

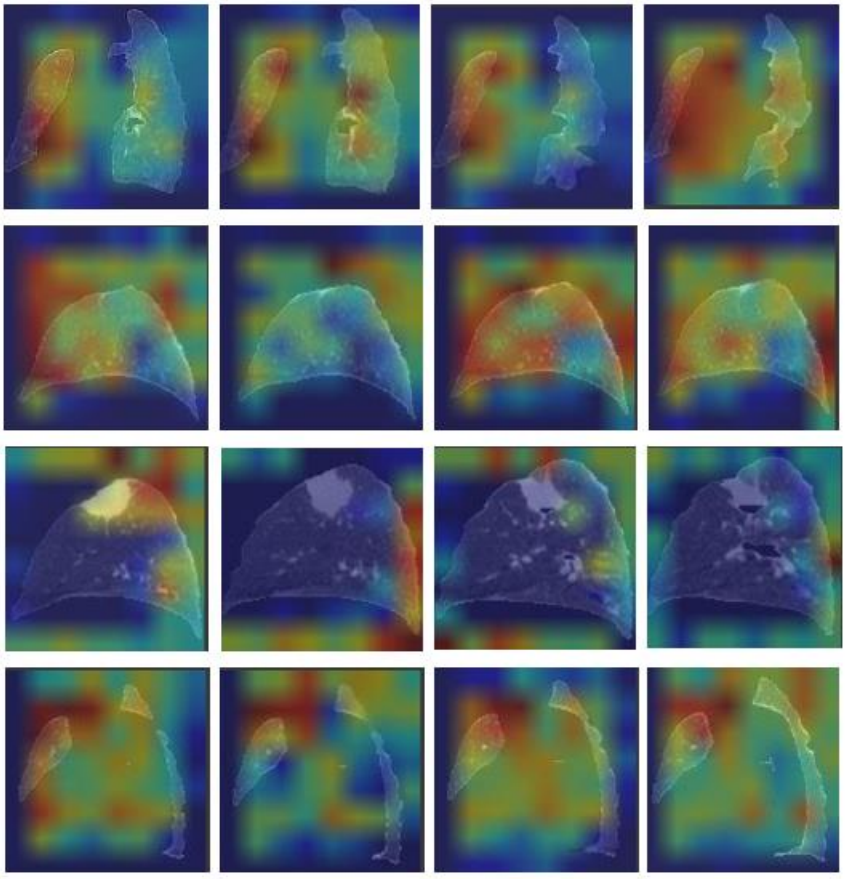

Figure 6.3. Representation of some frames of the generated video (heatmaps superimposed to ADC CT slices). Red colour represent those areas on which network mainly focus its attention. Blue colour represent those areas on which network does not focus its attention.

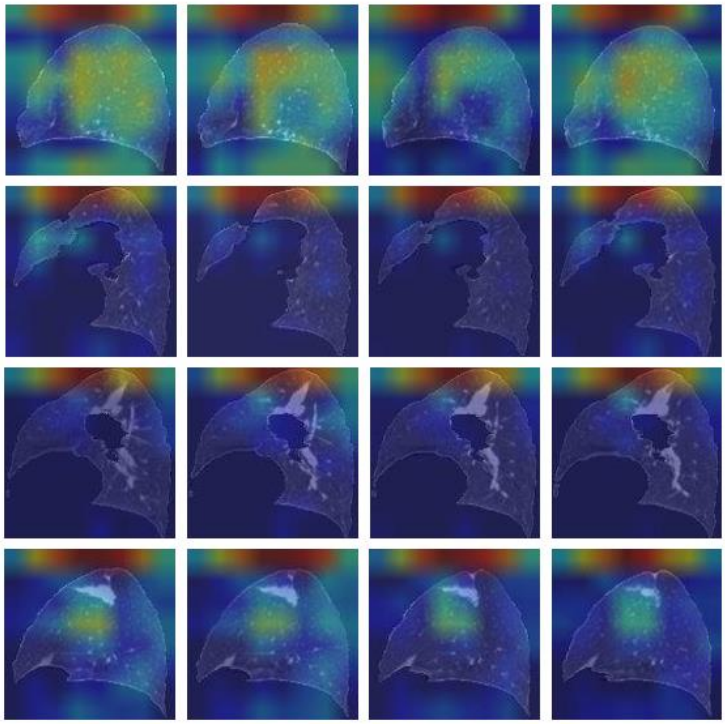

Figure 6.4. Representation of some frames of the generated video (heatmaps superimposed to SCC CT slices). Red colour represent those areas on which network mainly focus its attention. Blue colour represent those areas on which network does not focus its attention.

6.2 Final experiment

The results reported in this section refer to the final experiment.

6.2.1 Hyperparameter tuning

The results of the hyperparameter tuning are reported in Table VIII. From here has been selected the set of hyperparameters which provide the best model performances. In this case there are two calls that provide the same accuracy of 69.99%. So, both the set of parameters that provide this accuracy values have been tested and the call number 9 is the one liked to the best model performances.

Table VIII. Hyperparameter tuning results. Dropout1_rate is the dropout rate associated to the first Dropout layer. Dropout2_rate is the dropout rate associated to the second Dropout layer.

Call number	Filter	Dropout1_rate	Dropout2_rate	Learning rate	Accuracy
0	16	0.5	0.5	0.000100	50.00%
1	128	0.7	0.8	0.000001	50.00%
2	4	0.5	0.3	0.000100	50.00%
3	16	0.1	0.2	0.000100	40.00%

4	32	0.4	0.3	0.000001	50.00%
5	32	0.3	0.7	0.000001	40.00%
6	64	0.5	0.3	0.000100	50.00%
7	64	0.1	0.5	0.000010	30.00%
8	16	0.6	0.8	0.000100	60.00%
9	256	0.3	0.7	0.000100	69.99%
10	256	0.6	0.1	0.010000	50.00%
11	32	0.3	0.7	0.000100	60.00%
12	64	0.3	0.5	0.000100	40.00%
13	64	0.3	0.7	0.000100	40.00%
14	128	0.1	0.7	0.000100	50.00%
15	64	0.4	0.5	0.000001	50.00%
16	256	0.3	0.1	0.000100	60.00%
17	256	0.6	0.7	0.000100	50.00%
18	256	0.5	0.7	0.000100	50.00%
19	256	0.4	0.7	0.000100	50.00%
20	256	0.3	0.7	0.000100	60.00%
21	256	0.3	0.7	0.001000	50.00%
22	256	0.3	0.7	0.000010	50.00%
23	32	0.6	0.8	0.001000	50.00%
24	256	0.3	0.7	0.000010	60.00%
25	256	0.3	0.7	0.010000	50.00%
26	16	0.6	0.8	0.010000	50.00%
27	16	0.6	0.8	0.001000	40.00%
28	256	0.7	0.7	0.000100	50.00%
29	4	0.2	0.3	0.000010	60.00%
30	256	0.8	0.7	0.000100	50.00%
31	256	0.2	0.7	0.000100	50.00%
32	4	0.2	0.3	0.001000	60.00%
33	256	0.1	0.7	0.000100	50.00%
34	4	0.2	0.3	0.000100	50.00%
35	4	0.2	0.3	0.010000	50.00%
36	4	0.4	0.3	0.000010	40.00%

37	4	0.2	0.3	0.000001	69.99%
38	64	0.2	0.3	0.000001	60.00%
39	4	0.2	0.6	0.000001	50.00%

6.2.2 Model evaluation

The value identified with the hyperparameter tuning, presented in the previous section, have been utilized for constructing the model. This has been compiled and trained achieving the performances reported in Table IX. In the Table X and in Table XI instead are reported respectively the performances obtained as results of the model evaluation and of the model prediction. In Figure 6.5 it is possible to see the ROC curve together with the AUC value. In Figure 6.6 is represented the confusion matrix.

Table IX. Model training performances.

Losses %	Accuracies %	Sensitivities%	Specificities %	F1_scores %
69.03	62.50	0.00	0.00	0.00

Table X. Model evaluation performances. Support stands for the number of data in the validation set.

Cancer type	Precision %	Recall %	F1_scores %	Support
ADC	65.00	55.00	59.00	20
SCC	61.00	70.00	65.00	20

Table XI. Model prediction performances. Support stands for the number of data in the test set and avg. stands for average.

	Precision %	Recall %	F1_scores %	Support
Macro avg.	63.00	62.00	62.00	40
Weighted avg.	63.00	62.00	62.00	40

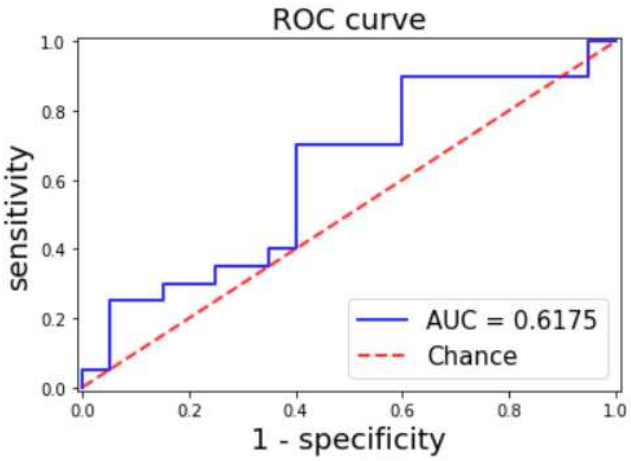

Figure 6.5. ROC curve and AUC value.

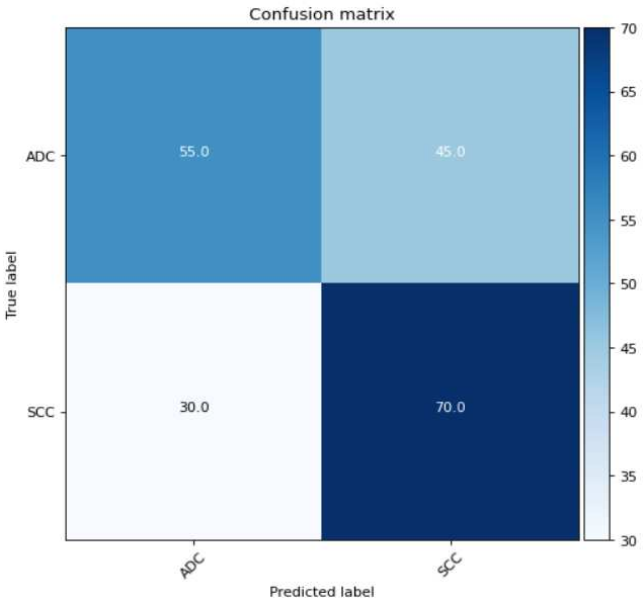

Figure 6.6. Confusion matrix.

6.2.3 Model interpretability

As reported in section 5.1.5, it has been created the new model which achieved an avg. loss of 69.81 %. This model has then been exploited for generating the heatmaps which have been then superimposed to the CT slices. Some of the most explicative frames of the generated video for the ADC scan are reported in Figure 6.7. In Figure 6.8 instead can be observed some frames of the generated video for the SCC scan.

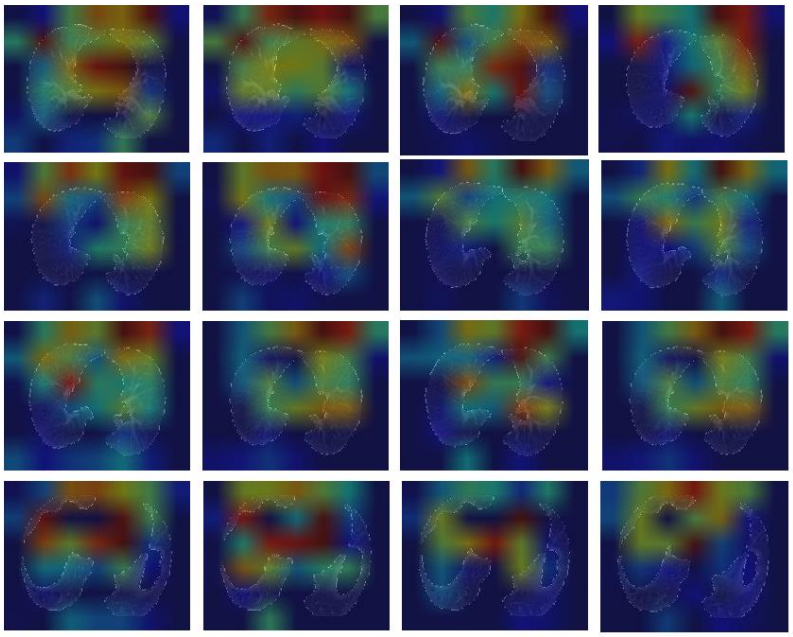

Figure 6.7. Representation of some frames of the generated video for the final experiment (heatmaps superimposed to ADC CT slices). Red colour represent those areas on which network mainly focus its attention. Blue colour represent those areas on which network does not focus its attention.

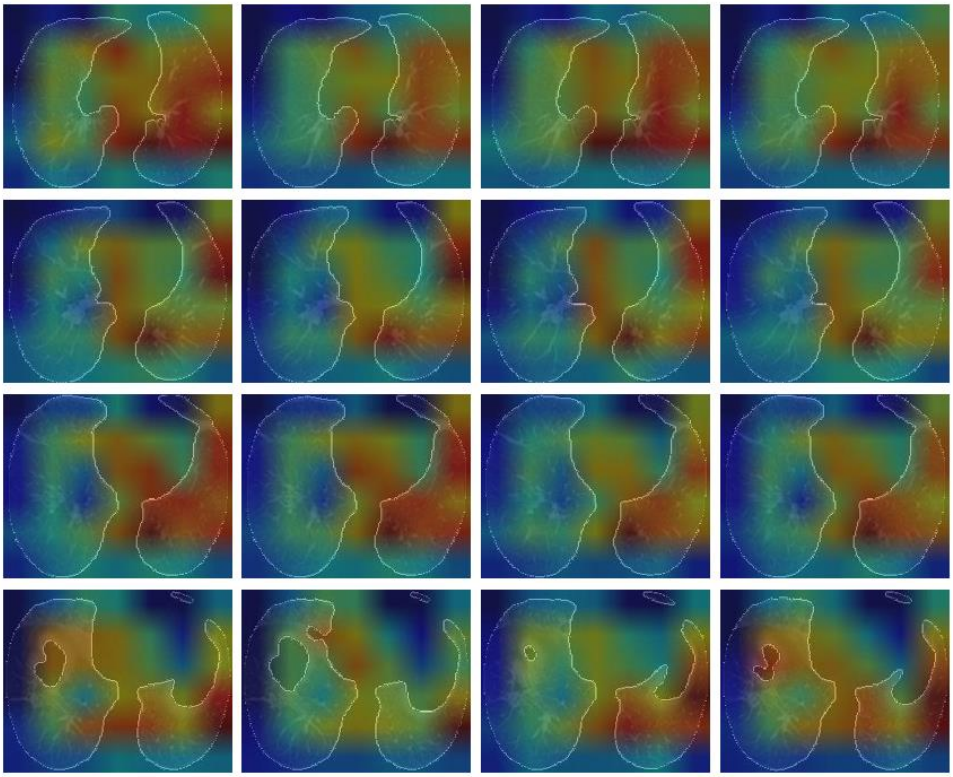

Figure 6.8. Representation of some frames of the generated video for the final experiment (heatmaps superimposed to SCC CT slices). Red colour represent those areas on which network mainly focus its attention. Blue colour represent those areas on which network does not focus its attention.

7 Discussion

The aim of this work was to propose an interpretable AI-based model, able to discriminate between ADC and SCC histotypes from whole lungs CT scans. Firstly the model has been tested in the preliminary experiment with a reduced number of data (only 10 slices per scan) characterized by a low resolution. Then the same model has been exploited in the final experiment with augmented data characterized by 250 slices per scan with higher resolution. For comparing the performances of the two experiments is taken into account mainly the AUC. This choice is linked to the fact that this is the only metric independent from the classification threshold (which is 50% in the binary case). In addition the AUC, together with the sensitivity, are the most important measures in the clinical field. In the preliminary experiment has been obtained an AUC of 56.72%. This means that the model correctly predict the histotype of tumour in the 56.72% of cases. Actually this is not a so good result because it is slightly better than "guessing by chance". The fact that the AUC is not so high can be observed also in the obtained activation maps (reported in Section 6.1.3); these are not confined to the area of the offending lung nodule. However, the generated activation maps are all focused inside the lungs and this is a positive aspect which indicates that the model is "not completely out of the way". Actually it was expected to obtain not very high performances because the data used are not so many and their resolution is poor. This is why augmented data with better resolution have been used in the advanced experiment. There, the AUC obtained is 61.75%. It can be observed an increase of five percentage points and this was predictable because of the higher quality data used here. A slight improvement is visible also in the activation maps (in Section 6.2.3) which are more confined. In particular, looking at the recalls reported in Table X, it is possible to notice that the model is able to correctly identify ADC histotype in 55% of cases and SCC in 70% of cases. Thus, the model better classify SCC. This is reflected in the activation maps generated when applying Grad-CAM algorithm to scans containing SCC tumour. The generated heatmaps are more confined in the external edge of the medial part of right lung which is where the offending lung nodule is. Concluding it can be said that the achieved performances are not very high but it is also necessary to say that the task is very demanding for the model. The aim of the model is to determine which type of tumour is present prior to lung biopsy and this is very complex. So it can be said that this is a very promising starting point for developing a network which is able to discriminate between SCC and ADC histotypes. Future improvements can allow to optimize the network for solving the task proposed here reaching better results. Maybe some attention can be put also to ADC tumour in order to better understand why the model presents problems in its identification.

Bibliography

[1] Massimo Franzin. Compendio di anatomia umana, 2007,seconda edizione.

[2] Martini, Timmons, Tallitsch. Anatomia umana. 2016. VI edizione. Edises,

[3] OpenStax College. Anatomy & Physiology. Vol.3. Textbook Equity Open Education.

[4] G. Ambrosi, P. Castano, R.F. Donato. Anatomia dell'uomo. 2006. Seconda edizione. Edi-Ermes.

[5] J.B. West. Fisiologia della respirazione (l'essenziale). 2017. VI italian edition. Piccin editore Padova.

[6] Joseph D.Bronzino. Medical devices and systems Cap.11. 2006. Edition 3.

[7] Khanpur. Handbook of biomedical instrumentation Cap. 20. 2003. Second edition.

[8] Francesco P. Branca. Fondamenti di ingegneria clinica Vol. 2. 2008. Springer Verlag.

[9] Dr. Eng. Sarah Hagi. CT generation RAD309. 2015.

[10] Kalender WA. X-ray computed tomography. Phys Med Biol. 2006 Jul 7;51(13):R29-43. doi: 10.1088/0031-9155/51/13/R03. Epub 2006 Jun 20. PMID: 16790909.

[11] Clark SB, Alsubait S. Non Small Cell Lung Cancer. 2021 Sep 9. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan–. PMID: 32965978.

[12] Bernard W. Stewart. Mechanisms of carcinogenesis: from initiation and promotion to the hallmarks. 2019. Chapter. 11.

[13] Alecsandru Ioan Baba and Cornel Câtoi. Comparative oncology. 2007. Chapter 2.

[14] http://helmberg.at/carcinogenesis.htm (12/03/2022).

[15] Chemical Carcinogenesis: Initiation, Promotion and Progression. NST110, Toxicology Department of Nutritional Sciences and Toxicology University of California, Berkeley.

[16] https://dralanjeans.wordpress.com/2011/05/17/the-three-stages-of-carcinogensis-part-2https://dralanjeans.wordpress.com/2011/05/17/the-three-stages-of-carcinogensis-part-2-promotion/promotion/ (12/03/2022).

[17] Rudin CM, Brambilla E, Faivre-Finn C, Sage J. Nat Rev Dis Primers. Small-cell lung cancer2021;7(1):3. Published 2021 Jan 14. doi:10.1038/s41572-020-00235-0

[18] https://www.yalemedicine.org/conditions/non-small-cell-lung-cancer (14/03/2022).

[19] Khevna Vasani1, Ayushi Shah. Lung cancer detection using CT scan images. International Research Journal of Engineering and Technology (IRJET). April 2021. Volume:08, Issue: 04. e-ISSN: 2395-0056. p-ISSN: 23950072.

[20] Dr.NouraAlHinai. Chapter 1 - Introduction to biomedical signal processing and artificial intelligence. 2020. Pages 1-28. https://doi.org/10.1016/B978-0-12-818946-7.00001-9.

[21] Purandare NC, Rangarajan V. Imaging of lung cancer: Implications on staging and management. Indian J Radiol Imaging. 2015;25(2):109-120. doi:10.4103/0971-3026.155831

[22] Suren Makaju. Prasad P.W.C. Abeer Alsadoon. A.K. Singh. Lung Cancer Detection using CT Scan Images. January 2018. DOI:10.1016/j.procs.2017.12.016

[23] https://www.ibm.com/cloud/learn/convolutional-neural-networks (30/03/2022).

[24] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learninghttps://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47algorithms-934a444fca47 (30/03/2022).

[25] Wafaa Alakwaa ,Mohammad Nassef ,Amr Badr. Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN). (IJACSA) International Journal of Advanced Computer Science and Applications, 2017. Vol. 8, No. 8.

[26] https://guandi1995.github.io/Pooling-Layers/ (02/04/2022).

[27] https://www.aegissofttech.com/articles/watershed-algorithm-and-limitations.html (02/04/2022).

[28] https://www.oracle.com/artificial-intelligence/what-is-ai/ (02/04/2022).

[29] https://www.ibm.com/cloud/learn/what-is-artificial-intelligence (02/04/2022).

[30] Karen Drukker, Pingkun Yan, Adam Sibley, Ge Wang. Biomedical image and analysis through deep learning. 2021. Pages 49-74. Doi: https://doi.org/10.1016/B978-0-12-821259-2.00004-1

[31] https://www.ibm.com/blogs/systems/ai-machine-learning-and-deep-learning-whats-thehttps://www.ibm.com/blogs/systems/ai-machine-learning-and-deep-learning-whats-the-difference/difference/ (04/04/2022).

[32] https://vitalflux.com/7-common-machine-learning-tasks-related-methods/ (04/04/2022).

[33] https://www.javatpoint.com/overfitting-in-machine-learning. (04/04/2022).

[34] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017 Dec;42:60-88. doi: 10.1016/j.media.2017.07.005. Epub 2017 Jul 26. PMID: 28778026.

[35] Ian Goodfellow and Yoshua Bengio and Aaron Courville. Deep Learning. Goodfellow-et-al2016. MIT Press.

[36] https://netai.it/guida-rapida-alle-funzioni-di-attivazione-nel-deep-learning (04/04/2022).

[37] Andreas Maier, Christopher Syben, Tobias Lasser, Christian Riess, Zeitschrift für Medizinische Physik. A gentle introduction to deep learning in medical image processing.Volume 29. Issue 2,2019. Pages 86-101, ISSN 0939-3889,

[38] Rikiya Yamashita & Mizuho Nishio & Richard Kinh Gian Do & Kaori Togash. Convolutional neural networks: an overview and application in radiology. Received: 3 March 2018 / Revised: 24 April 2018 /Accepted: 28 May 2018 /Published online: 22 June 2018. https://doi.org/10.1007/s13244-018-0639-9.

[39] Koichiro Yasaka, Hiroyuki Akai, Akira Kunimatsu, Shigeru Kiryu, Osamu Abe. Deep learning with convolutional neural network in radiology. Received: 28 December 2017 / Accepted: 26 February 2018 / Published online: 1 March 2018. https://doi.org/10.1007/s11604-018-0726-3.

[40] Mobiny, Aryan & Nguyen, Hien. Fast CapsNet for Lung Cancer Screening (2018).

[41] Shih-Chung B. Lo, Heang-Ping Chan, Jyh-Shyan Lin, Huai Li, Matthew T. Freedman, Seong K. Mun, Artificial convolution neural network for medical image pattern recognition. Neural Networks,Volume 8, Issues 7–8,1995,Pages 1201-1214,ISSN 08936080, https://doi.org/10.1016/0893-6080(95)00061-5.

[42] 7.1. Deep Convolutional Neural Networks (AlexNet) — Dive into Deep Learning 0.17.5 documentation (d2l.ai) (06/04/2022)

[43] Aman Agarwal, Kritik Patni, Rajeswari Devarajan. Lung Cancer Detection and Classification Based on Alexnet CNN. July 2021. DOI:10.1109/ICCES51350.2021.9489033. Conference: 2021 6th International Conference on Communication and Electronics Systems (ICCES)

[44] Muayed S AL-Huseiny, Ahmed S Sajit. Transfer learning with GoogLeNet for detection of lung cancer. Indonesian Journal of Electrical Engineering and Computer Science Vol. 22, No. 2, May 2021, pp. 1078~1086 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i2.pp1078-1086

[45] Afshar P, Oikonomou A, Naderkhani F, Tyrrell PN, Plataniotis KN, Farahani K, Mohammadi A. 3D-MCN: A 3D Multi-scale Capsule Network for Lung Nodule Malignancy Prediction. Sci Rep. 2020 May 14;10(1):7948. doi: 10.1038/s41598-020-64824-5. PMID: 32409715; PMCID: PMC7224210.

[46] Hamdalla F. Al-Yasriy et al. Diagnosis of Lung Cancer Based on CT Scans Using CNN. 2020 IOP Conf. Ser.: Mater. Sci. Eng. 928 022035.

[47] Alder A, Dey D, Sadhu AK. Lung Nodule Detection from Feature Engineering to Deep Learning in Thoracic CT Images: a Comprehensive Review. J Digit Imaging. 2020 Jun;33(3) 655677. doi:10.1007/s10278-020-00320-6. PMID: 31997045; PMCID: PMC7256172.

[48] Fan, F., Xiong, J., Li, M., & Wang, G. On Interpretability of Artificial Neural Networks: A Survey. IEEE Transactions on Radiation and Plasma Medical Sciences, (2021). 5, 741-760.

[49] Imawanto, Renard Elyon & Rifqialdi, Ghiffary & Yudith, Amadhea & Sinaga, Adrian. (2021). Computer-Aided Detection of Lung Cancer from CT-scan Images with Visual Insights using Deep Convolutional Neural Network.

[50] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. [Submitted on 7 Oct 2016 (v1), last revised 3 Dec 2019 (this version, v4)]. https://doi.org/10.48550/arXiv.1610.02391

[51] Eali Stephen Neal Joshua1 , Debnath Bhattacharyya2 , Midhun Chakkravarthy1 , Hye-Jin Kim. Lung Cancer Classification Using Squeeze and Excitation Convolutional Neural Networks with Grad Cam++ Class Activation Function. 9 August 2021. Vol. 38, No. 4, August, 2021, pp. 1103-1112. https://doi.org/10.18280/ts.380421

[52] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, Vineeth N Balasubramanian. GradCAM++: Improved Visual Explanations for Deep Convolutional Networks. https://doi.org/10.48550/arXiv.1710.11063

[53] Selene Tomassini, Student Member IEEE, Nicola Falcionelli, Paolo Sernani, Agnese sbrollinu, Member IEEE, Micaela Morettini, Member IEEE, Laura Burattini, Member IEEE, Aldo Franco Dragoni, Member IEEE. Cloud-YLung for Non-Small Cell Lung Cancer Histology Classification from 3D Computed Tomography Whole-Lungs Scans.

[54] https://keras.io/api/layers/recurrent_layers/time_distributed/ (03/06/2022).

[55] https://www.mygreatlearning.com/blog/introduction-tohttps://www.mygreatlearning.com/blog/introduction-to-vgg16/ -VGG%20%E2%80%93%20The%20Ideavgg16/#VGG%20%E2%80%93%20The%20Idea (03/06/2022).

[56] https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/ (03/06/2022).

[57] https://medium.com/neuronio/an-introduction-to-convlstm-55c9025563a7 (03/06/2022).

[58] https://machinelearningknowledge.ai/keras-dropout-layer-explained-for-beginners/ (03/06/2022).

[59] https://analyticsindiamag.com/a-complete-understanding-of-dense-layers-in-neural-networks/ (05/06/2022).

[60] https://www.geeksforgeeks.org/hyperparameter-tuning/ (07/06/2022).

[61] https://medium.com/geekculture/how-does-batch-size-impact-your-model-learning 2dd34d9fb1fa (07/06/2022).

[62] https://medium.com/sitechassethealthcenter/gaussian-process-to-optimize-hyperparameters-ofhttps://medium.com/sitechassethealthcenter/gaussian-process-to-optimize-hyperparameters-of-an-algorithm-5b4810277527an-algorithm-5b4810277527 (08/06/2022).

[63] https://developers.google.com/machine-learning/crash-course/classification/accuracy (09/06/2022).

[64] https://thebiologynotes.com/sensitivity-and-specificity/#sensitivity (10/06/2022).

[65] https://towardsdatascience.com/the-f1-score-bec2bbc38aa6 (10/06/2022).

[66] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc (10/06/2022).

[67] https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5 (10/06/2022).

[68] https://towardsdatascience.com/grad-cam-camera-for-your-models-decision-1ef69aae8fe7 (09/06/2022)

[69] Saleem H, Shahid AR, Raza B. Visual interpretability in 3D brain tumor segmentation network. Comput Biol Med. 2021 Jun;133:104410. doi: 10.1016/j.compbiomed.2021.104410. Epub 2021 Apr 19. PMID: 33894501.