Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale
in INGEGNERIA BIOMEDICA

Tesi di Laurea Magistrale

A Deep Learning Approach for Segmentation of Ovarian Adnexal Masses

Relatori

Prof. Filippo MOLINARI

Prof. Massimo SALVI

Tutor aziendale

Daniele CONTI

Candidata

Cecilia MARINI

Anno Accademico 2021-2022

Abstract

Ovarian cancer is the eighth most widespread cancer in the world among women. Due to the absence of a specific symptomatology and the lack of a defined screening protocol, this disease is usually diagnosed at an advanced stage, leading to the increase in the corresponding mortality rate. At present, **transabdominal sonography (TAS)** and **transvaginal sonography (TVS)** are generally recognised as the main diagnostics techniques for the first identification of the neoplasm. However, it is well known that their diagnostic effectiveness can be seriously compromised by the intrinsic noisy and operator-dependent nature of ultrasound images. In addition, the identification of ovarian structures is a time-consuming task in clinical practice, almost always repetitive and prone to errors when manually performed by medical doctors. New diagnostic approaches based on artifical intelligence algorithms have started to be investigated for the automatic detection, segmentation and classification of medical images, aiming at the development of **Computer Aided Diagnosis and Detection (CAD)** models. Although these protocols are commonly employed in the examination of other image acquisition modes, such as CT or MRI, their application is also expanding to ultrasound images. Nowadays, the majority of artificial intelligence algorithms in gynaecological ultrasound is mostly focused on the classification of ovarian mass types. Indeed, despite many advances have already been made to identify anatomical structures of the same district, such as ovarian follicles, very few has been done concerning ovarian mass segmentation. An increasingly widespread idea in medical imaging is that the automatic segmentation of ovarian masses could be consistently helpful for the development of CAD systems. For instance, the emerging field of radiomics could strongly benefit from the development of such segmentation protocols: the identification of ROIs within medical images is a mandatory step for the extraction of quantitative descriptors useful for tumor discrimination. The purpose of this thesis project is the development of an automatic algorithm that deals with the **segmentation of ovarian masses**. Given the success of Deep Learning models in the medical field, and, in particular, of **Fully Convolutional Neural Networks (FCNNs)** in image segmentation tasks, the learning model proposed in this work makes use of a **modified U-Net network** with **MobileNetV2** as encoding block and simple transpose-deconvolutional-based upsampling as decoding block. Being one of the few attempts to the segmentation of ovarian masses, its future potential has been here evaluated on an easier segmentation task, that is the identification of **cysts of unilocular serous type**. The model showed to be perfectly suitable to perform the proposed task and confirmed the improvement over the current state-of-the-art. Its performance have been further improved with both the introduction

of Data Augmentation and a refining post-processing stage. The success of this segmentation algorithm encourages as a next step its subsequent application to the segmentation of histotypes with more complex morphology than those covered in the present study. Another possible implementation could be aimed at multiclass segmentation of different cystic components. The automatic identification of these substructures would contribute to the development of a more interpretable and less "black-box" algorithm for differential diagnostics of ovarian lesions.

Contents

List of Tables

List of Figures

Acronyms

AI
Artificial Intelligence

CAD
Computer Aided Detection and Diagnosis

CT
Computer Tomography

DL
Deep Learning

FCNN
Fully Convolutional Neural Network

NN
Neural Network

ML
Machine Learning

MRI
Magnetic Resonance Imaging

PET
Positron Emission Tomography

TAS
Transabdominal Sonography

TVS

Transvaginal Sonography

US

Ultrasound

Chapter 1

Ovarian Tumor

1.1 Epidemiology and risk factors

Ovarian cancer is the eighth most common cancer among women; in 2020, around 313 959 cases and 207 252 deaths (66%) were reported worldwide for this disease [1]. In Italy, according to an estimate drawn up in 2020 by the **Italian Association of Medical Oncology** (AIOM) and the **Italian Cancer Registry Association** (AIRTUM), ovarian cancer affects approximately 5200 women every year [2]. Currently, it is considered one of the most difficult cancers to treat among gynaecological malignancies, due to the absence of symptoms in the initial stages and the lack of an accurate screening strategy, leading, in most cases, to a late diagnosis. The five-year survival rate stands at 43%, which is significantly low if compared to the same statistic registered for breast cancer (around 87%) [2]. When symptoms are present, they can include abdominal or pelvic pain and swelling, but also changes in normal bowel function (with bloating or constipation). However, the ambiguity of these symptoms, which are common in several other conditions, makes it difficult to correctly diagnose the disease.
Age is one of the main **risk factors** of ovarian cancer. In postmenopausal women, the likelihood of cancer increases greatly, with a peak incidence of 50-69 years. Other risk factors concern aspects of the female reproductive cycle: an early period, or more generally a high number of ovulations throughout life, can be a predisposing factor for the onset of an ovarian neoplasia. The ovary is damaged during ovulation, increasing the risk of uncontrolled proliferation within tissue renewal. Other cancer-related endocrine factors include infertility, nulliparity, and late pregnancy, elements related to ovarian stimulation. Having a family history of ovarian and breast cancer has also been shown to increase women's risk of ovarian cancer. Hereditary genetic mutations in the genes BRCA1 (Breast Cancer 1) and BRCA2 (Breast Cancer 2) are

frequently responsible for the disease transmission and manifestation. According to the European Society of Medical Oncology (ESMO), the probability of developing OC increases by 24-40% for women with BRCA1 mutation and by 11-18% for women with BRCA2 mutation, respectively [3]. Lastly, lifestyle factors, like obesity and alcohol consumption, contribute to increase the risk of developing ovarian cancer, as it happens for other malignancies.

[4]

1.2 Anatomy of ovaries

In order to achieve a better understanding of the subsequent classification of ovarian tumors, a brief description of the ovary is reported (Figure 1.1).

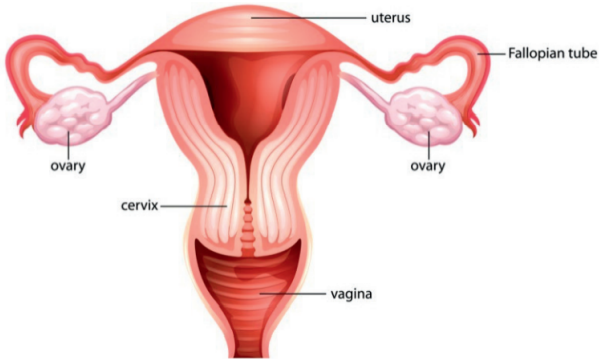


Figure 1.1: Anatomy of the female reproductive organs, showing the uterus, fallopian tubes and ovaries.

The ovary is an organ composed of two glands, known as ovaries, which together with the uterus, the fallopian tubes, the vagina and the vulva form the female reproductive apparatus. The ovaries, also called "female gonads", are two organs of ovoid shape whose size can vary from 3.0 to 5.0 cm, placed on the sides of the uterus and near the lateral pelvic wall. These glands perform two main functions: they deal with the production of oocytes, indispensable for reproduction (*gametogenic function*), and secrete hormones that regulate the stages of female reproductive life, i.e. estrogen, progesterone and partly androgen (*endocrine function*).

Histologically, the ovary consists of three layers (Figure 1.2):

- a **superficial epithelium**, often called improperly germinative epithelium, composed of epithelial cells of celomatic origin (cylindrical or flat);
- an intermediate zone, the **cortex**, made of fibroblastic stroma, a connective tissue responsible for the production of steroid hormones;
- an inner zone, the **medulla**, also composed of connective tissue, in which are distributed the nerve fibres, blood vessels and lymphatics that branch out inside the organ.

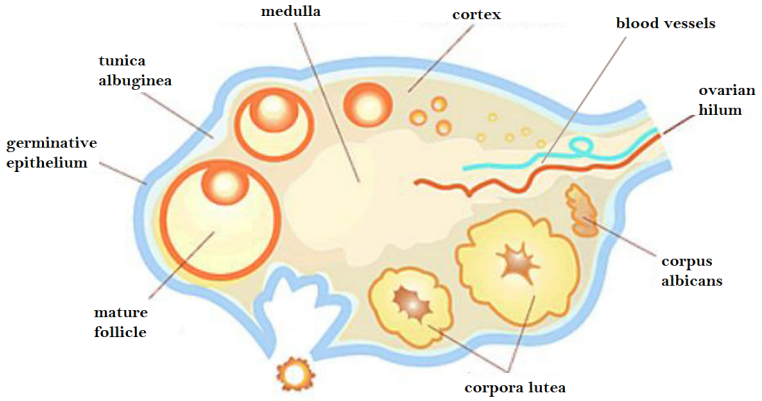


Figure 1.2: Ovarian structure.

In the outermost area of the cortical zone, immediately beneath the superficial epithelium, the stroma is rich in collagen fibres and low in cells (*tunica albuginea*). Instead, the stroma surrounding the ovarian organelles (ovarian follicles and luteal bodies), has a high cellular component, capable of differentiating into different endocrine elements, partly responsible for the production of female hormones in synergy with the cells of the granulosa (theca cells), and partly of male hormones (scattered steroidogenic elements). Stroma also contains germ cells, which give rise to oocytes.

[5]

1.3 Classification of ovarian adnexal masses

Ovarian adnexal masses are the hallmark of ovarian cancer and originate from the uncontrolled proliferation of cells in the organ; epithelial cells are the most common to undergo this abnormal growth, but germ cells and stromal cells can also give

rise to a tumour.
Ovarian tumours can thus be classified according to the type of cells from which they originate. Three main histotypes are distinguished:

- **Epithelial tumours**: originate from the surface celomatic epithelium and constitute more approximately 90% of malignant ovarian neoplasms; they affect women of both reproductive and post-menopausal age [6];
- **Germinal tumours**: originate from germ cells (oocytes), and represent about 5% of malignant ovarian neoplasms; in 40-60% of cases they are diagnosed in women under the age of 20 [2];
- **Sex-cord stromal tumour**: originate from the cortical stroma, and account for about 8% of all ovarian cancers [7]. Over half of these cancers occur in women over the age of 50, although the affected age group remains wide. Since the cells of these tumours are responsible for the production of ovarian steroid hormones, this class of tumours is usually associated with diseases characterized by hyperandrogenic or hyperestrogenic manifestations [8].

The categories described above belong to the group of *primary ovarian tumours*, as they originate from the three constituent elements of the ovary itself. *Secondary ovarian tumours* refer to extra ovarian neoplasms whose metastasis reach the ovary. Based on distinctive features of the mass, each of these categories has a number of more specific subtypes. Figure 1.3 below presents the most widespread classification of ovarian cancer.

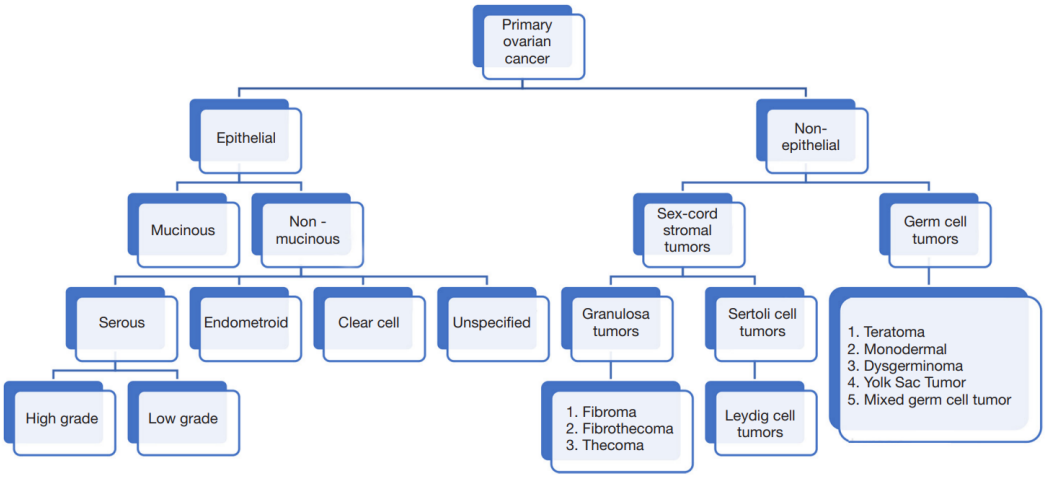


Figure 1.3: Classification of ovarian cancer.

The presence of an adnexal mass does not always coincide with the presence of a neoplasm: as with other types of tumours, it is always possible to distinguish between malignant and benign tumours. Consequently, each histotype can be further classified according to its benignity. In addition to the two standard categories (benign - B and malignant - M), in the context of ovarian masses, the borderline category (BOT) is also introduced. These tumours, also known as low malignant potential (LMP) tumours, are ovarian masses with intermediate characteristics, but generally benefit from a better prognosis than the malignant counterpart. A previous study of literature carried out within the company led to the following classification of the most common histotypes, reported in Table 1.1.

Epithelial Tumours (Type I)	
Mucinous Cystadenoma	B
Endometrioma	B
Brenner Tumour	B
Borderline Endometrioid Tumour	BOT
Borderline Mucinous Tumour	BOT
Borderline Bowel Mucinous Tumour	BOT
Mucinous Adenocarcinoma	M
Endometrioid Adenocarcinoma	M
Clear Cell Tumour	M
Epithelial Tumours (Type II)	
Cystadenofibroma	B
Serous Cystadenoma	B
Borderline Seerous Tumour	BOT
Low grade Adenocarcinoma	M
High grade Adenocarcinoma	M
Stromal Tumours	
Teratoma	B
Monodermal	B
Dysgerminoma	M
Yolk Sac Tumour	M
Mixed Germ Cell Tumour	M
Germinal Tumours	
Tecoma-Fibroma Group	B
Sertoli-Leydig Tumour	B/M

Table 1.1: The most common histotypes and their associated malignancy risk.

1.4 IOTA Standard

The accurate discrimination of malignant and benign masses is of paramount importance for the definition of appropriate treatment. Although the exact diagnosis requires a biopsy of the mass, most frequently the physicians need to provide a diagnosis through less invasive examinations, such as ultrasound based-screening or blood tests, before entering a surgical context.

Nowadays, doctors rely on different certified protocols to provide a pre-operative diagnosis, including **IOTA** (International Ovarian Tumor Analysis), **O-RADS** (Ovarian-Adnexal Reporting and Data System) and **GI-RADS** (Gynecologic Imaging Reporting and Data System) prediction models. Several studies are currently investigating the diagnostic accuracy of these system [9, 10]; among the mentioned, IOTA standard is probably the most widely used.

The work of the IOTA group, a team of experts in ovarian diseases, started in 2000, with the definition of terms useful for the description of adnexal masses, given the lack of a standard reference for this purpose [11]. In **Table 1.2** and **Table 1.3** some of the main characteristics of the ovarian masses addressed by the IOTA group are reported.

In 2008, the IOTA Group proposed the **Simple Rules** model, a diagnostic prediction system based on ultrasound-detectable features of ovarian masses [12]. Following a study upon ultrasound data from 1066 clinical cases in different countries (Italy, Belgium, Sweden, France and the United Kingdom), five typical traits of benign and malignant tumors were identified (Figure 1.4). The proposed classification model is extremely simple: a mass is diagnosed as benign if it shows at least one of the typical characteristics of a benign tumor, and none of the characteristics of malignancy; the opposite applies to a neoplasm. It follows that any mass not conform to these rules makes the prediction model not applicable.
Although the Simple Rules model allowed to label the ultrasound image, it could not provide the physician with the confidence with which the resulting class could be assigned to the tumour.

In order to overcome this issue, an advanced system was released in 2016. This new prediction model not only can return the confidence percentages of the classification but, if the tumour has been identified as malignant, also the relative probability of belonging to a certain stage is given. The algorithm is called **ADNEX**, and it differs from the previous one for the greater number of clinical cases that have been

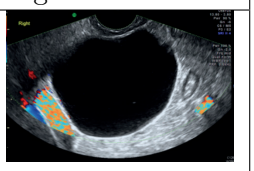
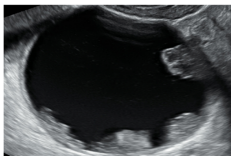
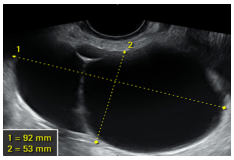
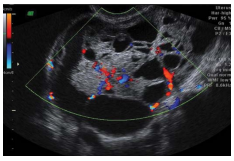
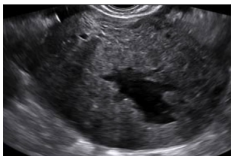
Terminology	Description	Figure
Unilocular cyst	A unilocular cyst without septa and without solid parts or papillary structures	
Unilocular solid cyst	A unilocular cyst with a measurable solid component or at least one papillary structure	
Multilocular cyst	A cyst with at least one septum but no measurable solid components or papillary projections	
Multilocular-solid cyst	A multilocular cyst with a measurable sold component or at least one papillary projection	
Solid tumour	A tumour where the solid components comprise 80% or more of the tumour when assessed in a two-dimensional section	

Table 1.2: Standard morphological terminology for ovarian lesions.

included for its development (about 6000), certifying its clinical validity, and for the use of metadata instead of images. Among the included descriptors considered useful for differential diagnosis are the age of the patient, the level of CA-125 found in blood tests, the size of the lesion and a number of other parameters.
The use of numerical descriptors instead of qualitative descriptors may be beneficial if a detailed mass identification is sought; however, the retrieval of the required information (blood exam results, manual measurements of the mass) inevitably leads to an increase in the time needed for diagnosis. The performances of the two models have proved to be generally optimal, with an AUC of around 83% for the Simple Rules [13] and around 90% for the use of the ADNEX model [14]. The IOTA protocols constitute the first preoperative differential diagnosis system and are now widely used also by not-experts physicians [15], thanks to their high reproducibility.

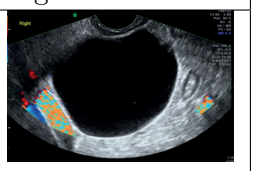
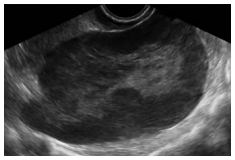
Terminology	Description	Figure
Anechoic	Completely black cyst	
Low-level	Homogeneous low-level echogenic cyst, as seen in mucinous tumors	
Ground glass	Homogeneously dispersed echogenic cystic content, as often seen in endometriotic cysts	
Hemorragic	Cyst with internal thread-like structures, representing fibrin strands	

Table 1.3: Cystic content terminology.

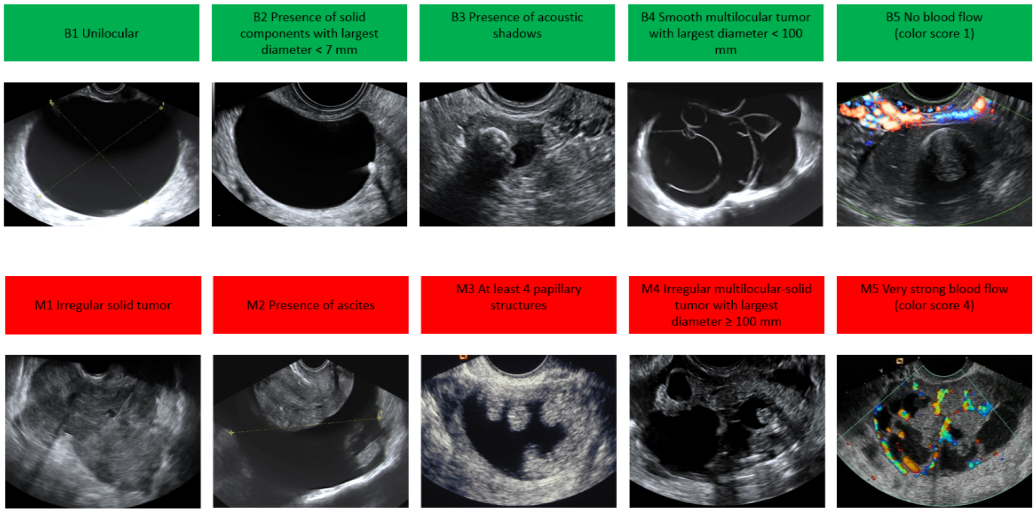


Figure 1.4: IOTA B-rules and M-rules with visual examples.

1.5 Diagnostic and screening approaches

One of the main reasons for the low survival rates following the identification of ovarian neoplasm, even after the treatment, is the late diagnosis of the disease. Unfortunately, to date, there is no screening protocol for early detection of ovarian cancer, since no diagnostic method attempted so far showed enough sensitivity.
In most cases, the diagnostic process starts with a medical visit carried out by the family doctor. According to the clinical history of the patient and the reported symptoms, the physician can prescribe further medical investigations or a specialist examination.
Blood tests are commonly prescribed in case of suspected ovarian cancer, as the finding of specific substances (tumour markers) can be a clue of the presence of a malignant tumour. As concerns ovarian cancer, high values of the **CA-125 antigen** (Carbohydrate Antigen 125) have been historically associated with the presence of the neoplasm, with some desire for it to be used as a "screening" tool. This protein is produced by neoplastic cells of epithelial ovarian cancer, the most frequent typology. Nevertheless, elevated serum CA-125 levels are only seen in 50% of patients presented with early-stage ovarian cancer [16, 17], and are common in several other diseases, non-ovarian gynecologic cancers and other non-epithelial malignancies [18, 19, 20]. Despite its flaws, the CA-125 marker is still widely used for differential diagnosis purposes [21, 22], and for monitoring the progress of anticancer therapies as it has proven helpful for detecting relapses [23].
Together with CA-125, **HE4** (Human Epididymis Protein 4) is currently being studied for the diagnosis of ovarian cancer. Several publications have reported that the combination of information derived from both markers is more effective in diagnosis than CA-125 alone [24, 25, 26].
Although the use of markers has been quite successful in the differential diagnosis of ovarian cancer in recent years, usually, the first examinations performed in the diagnostic process of ovarian tumours, are **Transvaginal Sonography** (TVS) and **Transabdominal Sonography**(TAS). These techniques allow for the detection of adnexal masses, which are the typical manifestations of this cancer. TVS and TAS can be commonly prescribed for screening, in the absence of symptoms, or for diagnostic purposes. Since ultrasound images are the primary data source for the thesis project, the sonography technique will be treated in more detail in the next chapter (Chapter 2).
Other imaging techniques that can be performed for diagnostic purposes are CT and PET/CT. However, their use is more often associated with preoperative assessments, to evaluate mass extension and consequently choose the best modality of intervention, or to monitor the effects of anticancer therapy.

CT is the most accurate technique in the detection of cancer metastases, for this reason it is frequently employed for staging purposes (*clinical staging*). Thanks to the high spatial resolution, CT is known to predict the success of cytoreduction operations effectively [27], whose main issue is the punctual localization of the tumour. The CT-based diagnostic approach instead has seen two major limitations: the impossibility of tracing small lesions (less than 2cm) [28], and the low discrimination of the different tissue components within the mass.
Combined **PET/CT** is a particular imaging mode, where the patient undergoes PET and CT in a single session, and the functional and anatomical images are merged together to maximize the information content of the examination. Unlike CT scan, PET reveals functional changes in tissues showing the distribution of glucose FDG (2-(F-18)-floide-2-deoxy-D-glucose); typically, malignant cells have higher absorption of FDG due to increased glycolytic turnover than the surrounding healthy tissue, thus PET is often exploited for cancer detection. However, this imaging technique is not recommended for primary detection of ovarian cancer, given the high rate of false positives due to absorption of FDG by healthy organs; endometrial absorption is also dependent on alternating ovulatory and menstrual phases in premenopausal women [29]. Although it is not the preferred technique for cancer detection, the PET/CT combination is playing a key role in treatment planning and follow-up, and is very successful in relapse detection [27, 30].
Finally, as with any neoplasm, the true diagnosis is made by **biopsy**. Depending on which treatment the patient is subjected to after the identification of the mass, the biopsy can be performed in different ways. In **fine-needle aspiration** (FNA), the patient is given local anaesthesia, and the doctor, with the help of imaging techniques (ultrasound or CT), guides the needle towards the tumour; the content of the mass is extracted and then analyzed in the laboratory. If the patient undergoes **laparoscopy** or **laparotomy** directly, a mass tissue sample is taken during the operation and then analyzed in the laboratory to certify the benignity or malignancy of the mass.

Despite ongoing diagnostic and screening research, 75-80% of cancers are already at an advanced stage at the time of diagnosis (FIGO III-IV)[2]. The current recommended screening procedure involves a **transvaginal ultrasound** examination once a year along with **CA-125** tumor marker analysis. A recent UK study, however, found that this approach, while reducing the incidence of cancer diagnosed in stage III-IV, was not effective in reducing the mortality rate [31].

Chapter 2

Ultrasound Imaging

2.1 Introduction

Ultrasonography (also known as Echotomography) is an imaging technique that allows the investigation of the body's internal structures through the use of ultrasound. The procedure is **non-invasive**, provides **real-time** results, and does not require any special preparation from the patient: it is thus globally acknowledged as **first-level diagnostic examination**, extremely powerful for preliminary diagnosis. The use of ultrasound instead of ionizing radiation makes this technique safe for the patient, allowing the investigation to be repeated and performed longer without damaging biological tissues. Other advantages offered by Ultrasonography include the low cost of the equipment and its ease of transport, which make the examination accessible to many people and replicable even outside the hospital environment.
The equipment is composed of:

- a **transducer**(the probe), that acts as emitter and receiver;
- a **central console**(the computer), which continuously elaborates the data coming from the probe;
- a **video monitor**, that displays the final ultrasound image.

As concerns ovarian pathologies, **transabdominal** and **transvaginal pelvic sonography** are the most used ultrasound examinations. In abdominal ultrasound the probe is moved down the lower abdomen, previously covered with a conductive gel to facilitate the passage of ultrasound from the transducer to the tissue. In the transvaginal approach, instead, the probe, wrapped in a sterile shell and gel, is inserted inside the vagina, and is therefore closer to the organs of interest. This

latter mode allows to obtain more precise and detailed images of the anatomical structures, both for the proximity and for the reduction of the barriers (and therefore of the interferences) that separate the source of the ultrasound from the tissues of interest [32].

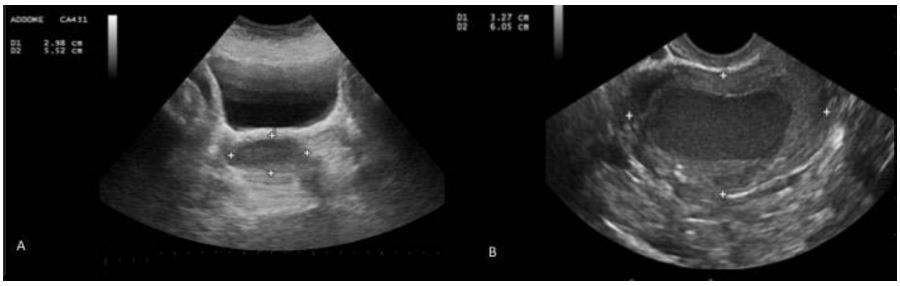


Figure 2.1: TAS (left) and TVS (right) showing a hypoechoic area located in the uterine shape; TVS modality reveals finer details of the mass (i.e. endometrial nature, presence of necrotic-colliquative material within the mass).

2.2 Ultrasound definition

Ultrasound is defined as a mechanical wave characterized by a frequency greater than the upper limit of human hearing ability. Like all sound waves, ultrasounds propagate through the conduction medium carrying vibrational energy, which moves the particles away and near, interchanging **compression zones** with **rarefaction zones** (Figure 2.2).

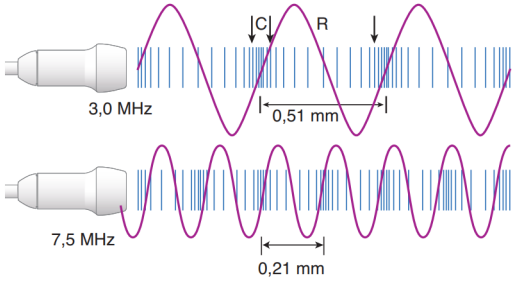


Figure 2.2: Sound wave representation.

Having wave characteristics, ultrasound can be defined according to the fundamental physical notations of wave mechanics: wavelength, frequency, and speed

propagation.

- **wavelength**: is the distance between two consecutive peaks. The wavelength range used in ultrasound is 1,5 to 0,1 nm;
- **frequency**: is defined as the number of complete oscillations, or cycles, which particles perform in the unit of time and is measured in Hertz (Hz). The frequency range used in ultrasound is between 2 and 15 Mega Hertz (MHz);
- **speed**: is the velocity at which the wave propagates and depends on the mechanical properties of the medium it passes through. The speed of propagation of a wave is given by:

$$v(m/s) = f(cycle/s) \cdot \lambda(m/cycle) \tag{2.1}$$

The propagation speed inside a medium depends on the resistance of the medium to the compression phenomenon, which in turn is defined by its density and elasticity ($firmness$). The speed is higher in tissues with increased stiffness and reduced density. Fortunately, **soft body tissues** share similar propagation rates and in clinical diagnostics it is therefore assumed that the average speed of ultrasound is **1540 m/s** (Table A.1). This hypothesis is critical for the ultrasound computer console to calculate the depth of a reflective surface; however, it remains an approximation: each tissue has actually its own propagation speed. Adnexal masses may contain different sorts of materials, including fibrous, blood, serum, and microcalcifications. It follows that within these cysts, assuming constant velocity, will inevitably lead to the production of artifacts in the resulting image.

Material or Tissue	**Speed (m/s)**
Air	331
Fat	1450
Water (50°C)	1540
Average soft tissue	1540
Brain	1541
Liver	1549
Kidney	1561
Blood	1570
Muscle	1585
Crystalline lens	1620
Bone	4080

Table 2.1: Ultrasound propagation speed in human body tissues.

The passage of ultrasound through the medium always encounters a resistance, which can be expressed in the form of **acoustic impedance**:

$$Z(kg/(m^2 \cdot s)) = v(m/s) \cdot \rho(kg/m^3) \tag{2.2}$$

The acoustic impedance value is a fundamental element of diagnostic ultrasound, as the amplitude of the return echo is proportional to the difference in acoustic impedance between the two adjacent tissues encountered by the ultrasound wave. As it is reported, the impedance depends on both velocity and density of the medium. If the speed is assumed constant (as previously stated), the acoustic impedance will be completely determined by the tissue density. Therefore, the information going back to the transducer and then turned into image, represents the difference in tissue densities encountered along the path of the ultrasound scanning line.

In accordance with the laws of physics that regulate the propagation of waves in media, ultrasounds mainly face:

- **Reflection**: occurs at the interface with a new medium; the incident wave reverses the direction of propagation and returns toward the transducer with the same incidence angle;
- **Refraction**: occurs at the interface with a new medium when the incidence of the wave is not perpendicular to the surface; it consists in a deviation of the incident radius;
- **Scattering**: occurs when the width or lateral dimension of the tissue boundary is less than one wavelength; if a large number of small tissue boundaries occurs, the scattering can radiate in all directions;
- **Attenuation**: is the result of an ultrasound wave losing energy and depends largely on the absorption (dispersion of energy in the form of dissipated heat) but also on reflection and scattering. The attenuation phenomenon is proportional to the characteristic frequency of the sound wave.

In order to maximize echo collection at the probe, specular reflection must be pursued, where incident US rays hit the surface perpendicularly. In the case of non-perpendicular incidence, some of the signals generated at the interface are lost due to refraction. Due to the numerous phenomena taking place at the interface, only a small part of the incident ultrasound will be correctly transmitted to the tissue, thereby maintaining the same direction and almost the original speed of propagation.

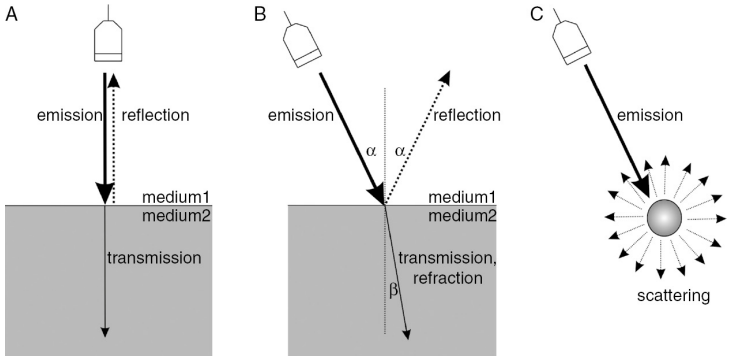


Figure 2.3: Specular reflection (A), Non-specular reflection (B), Scattering (C) phaenomena.

2.3 Ultrasound image acquisition

The process of image acquisition begins with ultrasounds generation, which take place within the transducer. The probe contains an **array of piezoelectric elements**, typically made of lead titanate-zirconate (PZT), capable of deformation when electrically stimulated. Viceversa, the crystals also manage to transform mechanical stress into electrical potential. This dual behaviour is to be found in the microscopic structure of the PZT crystals, whose asymmetry leads to shape and ions alterations following external stimulations. The mechanical vibration associated with the deformation of the crystal results in the production of the ultrasound. As the frequency of the generated US is inversely proportional to the thickness of the piezoelectric materials, depending on the value of the stimulation voltage and the morphology of the crystals, it will be possible to generate pulses of different frequencies, capable of reaching certain depths of exploration.

The ultrasound produced propagates following the principles governing the interaction of a mechanical wave in a medium; therefore, at each interface with a new tissue it is mainly reflected and refracted. The reflected component completely reverses its direction and returns to the probe, carrying with it the information content that will help provide the final image (echo). The component that is actually transmitted propagates in the original direction with much less energy than the original incident beam, mainly due to attenuation. Refraction phenomena can give rise to echoes that can return to the probe with information that is not representative of the actual depth of an object, and are therefore responsible for producing artifacts.

As anticipated, the echo produced is the more energetic the greater the difference in acoustic impedances in contiguous tissues. Consequently, if the second medium has a density much greater than the first, the signal returning to the probe will be highly energetic, and it will not be possible to investigate the structures below that interface, since the energy is mainly returned via reflection. In the event that the tissue discontinuity is such as to allow the further propagation of the ultrasound beam, this will continue until it encounters the next discontinuity, where beam splitting will occur again .

Therefore, the information collected and processed in real-time by the probe are:

- the **energy** with which the ultrasounds return to the detector, index of the attenuation:
- the **time** interval that separates the sending of the beam from its reception, an index of the depth at which the discontinuity was found.

Ultrasound reaching deeper depths will return to the probe highly attenuated, thus the computer console is equipped with complex **compensation systems** capable of amplifying the signal coming from distant sources. All the information collected by the transducer is processed by the central console to output the typical ultrasound image. The conical shape that defines the scan corresponds to the ultrasound fan, which is the area investigated by the set of ultrasounds produced by the probe. Its shape may vary depending on the probe used. The image is grayscale, and the intensity of each pixel is directly proportional to the intensity of the echo returning to the transducer. Areas that record more intense echoes are called **hyperechoic**, and are the brighter areas in the image. Since reflection is the prevailing phenomenon for these structures, hyperechoic zones correspond to high-density media (solid components, calcifications). The areas without echoes, called **hypoechoic**, are the darker areas. These indicate the presence of liquid collections (water, serum, urine).

2.4 Pulsed-echo

The ultrasound image is based on the pulse-echo principle. In order to create the image and keep it updated in real-time on the monitor, the transducer constantly sends and receives ultrasounds, using a pulsed mode to excite the crystals. This modality consists in stimulating the piezoelectric elements with a discontinuous current to have an equally discontinuous ultrasound emission. There will therefore

be a period during which the piezoelectric element is involved into ultrasonic beam generation, and will be unable to transduce stimuli from outside (**Pulse Duration**, PD), and a period during which the crystal is at rest, and can thus be stimulated by the incoming echo (**Receiving Period**, RP). The sum of these time intervals constitutes the **Pulse Repetition Period** (PRP), whose reciprocal is the **Pulse repetition Frequency** (PRF), i.e. the frequency with which the probe investigates the tissues.

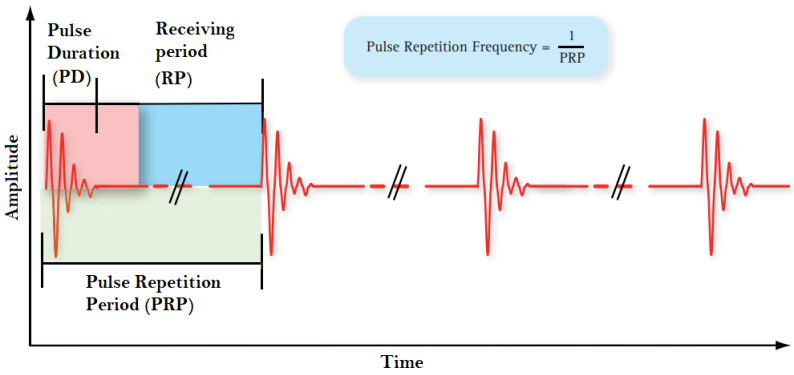


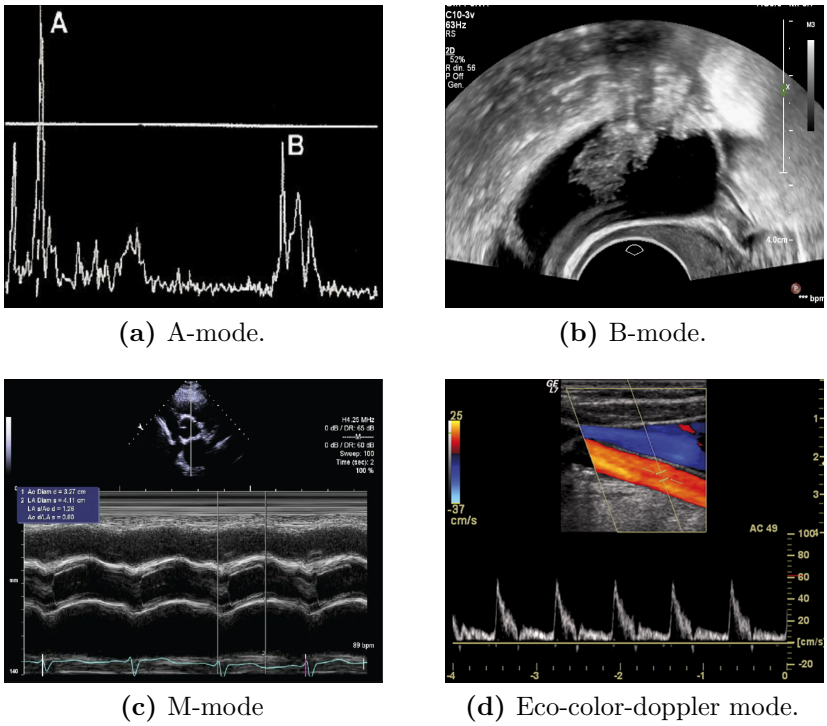
Figure 2.4: Echo pulsed mode.

The duration of the pulse strongly affects the axial resolution of the US beam: if the time interval between two echo signals originating from two different objects is less than the duration of the pulse, the return signals can overlap and prevent the correct display of the two objects as distinct.

2.5 Echography modes

There are several ways in which ultrasound can be arranged to display information extracted from tissues: these can vary according to the specification it is intended to examine.
A-mode (Amplitude - mode): consists of a one-dimensional scan in which a single transducer sends a single pulse that propagates through the tissue. It is the simplest acquisition of an ultrasonic echo and can be represented on a cartesian plane in which the vertical axis is an index of signal intensity and the horizontal axis represents the depth reached. It is still exploited today, but only in ophthalmology and neurology applications.
B - mode (Brightness - mode): is a two-dimensional scanning mode in which a transducer array transmits pulses to the tissue. The resulting image is the traditional

grayscale ultrasound image, in which each pixel takes on a color depending on the energy of the echo returning to the probe. The resulting image shows the anatomical substructures of the area of interest, represented in different shades depending on their density and depth. This is the most common modality used for diagnostic purposes in various applications (cardiological, breast, gynaecological).
M-mode (Motion - mode): is a modality based on B-mode. It offers the possibility to select a single transducer to isolate the sequence of grey tones related to a specific scanning line and to plot it over time. It's exploited to examine the movement of anatomical structures and is currently used in cardiology, for the analysis of valve movement.
Eco-color-doppler mode: this modality shares with the previous the usage of US to build the image, however, the focus is shifted from the morphology of anatomical structures to the analysis of blood flow. The theory of the Doppler effect is exploited for this purpose. The resulting image is a B-mode image superimposed with a colorimetric map of the blood flow, where the color represents directionality (red for approaching blood flow, blue for departing flow) and the brightness is proportional to the intensity.



(a) A-mode. **(b)** B-mode.

(c) M-mode **(d)** Eco-color-doppler mode.

Figure 2.5: Ultrasound modes.

2.6 Limitations

Ultrasound imaging is affected by several drawbacks, some related to the inherent limitations of the equipment, others to external factors. First of all, the **spatial resolution** of ultrasound imaging is among the lowest in diagnostic imaging. Spatial resolution is the ability of an ultrasound system to correctly detect and display adjacent structures. It can be broken into its components, axial resolution, measured along the direction of propagation of the beam, and lateral resolution, measured along the plane perpendicular to the direction of propagation. Both of these components depend on the frequency of US wave, and the overall effect is that the spatial resolution diminishes in value (optimal condition) when the frequency is high. However, the frequency can not be increased without taking into account the attenuation effect, therefore the common compromise sees as chosen frequency the highest allowing to reach the desired depth. The second major drawback of ultrasound imaging is its **applicability**, limited to soft tissues. Hard tissues show high reflection, preventing the ultrasound to propagate to underlying structures. This technique also suffers from numerous artifacts whether not correctly identified, they could lead to misdiagnosis. Given their variety and their relevance for segmentation tasks, they will be covered in the next section. Lastly, ultrasound imaging is a highly **operator-dependant** examination. The results vary significantly according to the person conducting the examination and his experience. This is one of the limitations artificial intelligence aims to overcome through the design of algorithms trained on the basis of the experiences of most expert physicians.

2.7 Artifacts

Image artifacts are frequently found in clinical ultrasound and consist of false or distorted representations of internal structures, resulting in an output image not corresponding to reality. The characteristics and causes of the most common artifacts are listed below; visual examples are provided in Figure 2.6.

- Beam width artifact: occurs when a highly reflective object falls within the distal zone of the beam and produces a detectable echo. The display assumes that this echo comes from within the focal area and displays it together with the correct one.
- Side lobe artifact: the main ultrasonic beam is accompanied by beams of lower energy that propagate in the radial direction in the vicinity of the transducer.

The presence of highly reflective structures in the propagation field of these minor beams can generate detectable echoes in the image.

- Mirror artifact: structures placed in close proximity to curved and highly reflective surfaces are reproduced in the image both in their real position and beyond the reflective interface acting as a mirror.

- Reverberation artifact: occur when the beam hits two nearby highly reflective surfaces perpendicularly. The echo generated by the first interaction can be reflected several times between the two interfaces before returning to the transducer. When this happens, the system records multiple echoes and brings them back to the image: the first recorded echo will be positioned correctly, the subsequent ones will be placed at a greater depth, depending on the time taken to return to the probe.

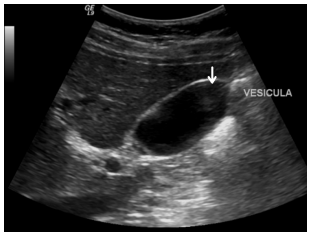
- Speed displacement artifact: the image reconstruction process works on the assumption that ultrasounds propagate through the body tissues with a constant speed, equal to 1540 m/s. As anticipated in Section 2.2, the real velocity differs from tissue to tissue even if slightly. The different propagation speed results in different times needed to return to the transducer, therefore a part of an identified object could appear broken or disconnected from its surroundings.

- Posterior acoustic shadowing: highly reflective or attenuating structures can cause complete ultrasonic beam reflection. Beyond the interface, there will be a zone free of echoes, thus completely anechogenic, the "shadow cone".

- Lateral acoustic shadowing: the tangential encounter of the ultrasonic beam with the ends of solid or liquid round formations results in the production of anechogenic bands, extending from the meeting point in the distal direction from the probe.

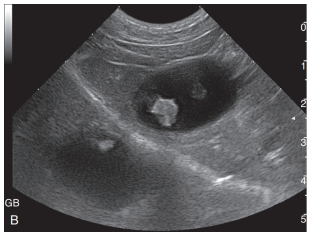
- Posterior enhancement: in presence of a homogeneous liquid collection, or a poorly reflective structure, the ultrasound undergoes minimal attenuation. At the lower wall of the same structure, the ultrasound will therefore appear more intense than those that are at the same depth but have not crossed the liquid zone. This artifact is presented as a hyperechoic nuance that follows a completely anechogenic structure.

In diagnostics, some of these artifacts are considered beneficial as their presence can sometimes help in the recognition of some peculiarities of the structures. The speed displacement artifact, for example, can help in determining the composition of the lesion (solid, liquid). Others remain simple noise. In the context of image segmentation, the artifacts are always attributed a negative connotation, since the

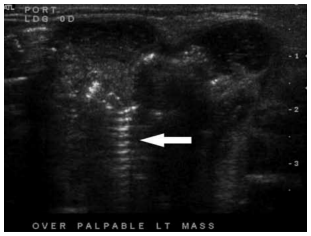


(a) Beam width artifact.

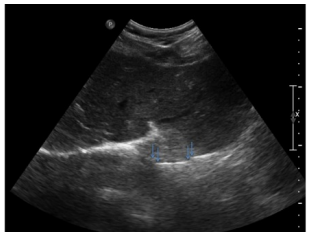
(b) Side lobe artifact.



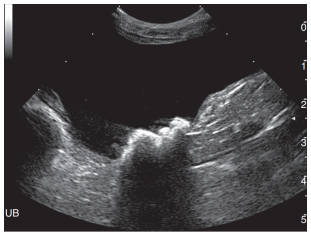
(c) Mirror artifact.



(d) Reverberation artifact.



(e) Speed displacement artifact.



(f) Posterior acoustic shadowing.

(g) Lateral acoustic shadowing.

(h) Posterior enhancement.

Figure 2.6: Common artifacts in US imaging.

algorithm that works to isolate individual structures, if not well trained, could behave indifferently for real and fictional representations.

Chapter 3

Artificial Intelligence

Artificial intelligence (AI) deals with the development of technologies exhibiting behaviours that normally require human intelligence. **Speech recognition**, **decision making** and **visual perception** are just some of the tasks that have been successfully tackled by AI. The corresponding algorithms can then be exploited for the implementation of complex technologies, such as automated guidance systems or banking services. The terms Machine Learning and Deep Learning are frequently misused in the field of artificial intelligence. Considering that both can be used in medical applications, a brief distinction is reported here.

Machine Learning (ML) is a branch of AI, whose models are capable of directly learning from a set of examples the salient traits needed to accomplish a given task, rather than resort to explicit coding instructions. The main feature of ML algorithms is thus the skill of autonomously adapting the learning according to variations in the input data.

Deep Learning (DL) is, in turn, a subcategory of Machine Learning that makes use of Deep Neural Networks to process a large amount of data for learning. The structure and depth of these models mimic the architecture of neuronal organization characteristic of the nervous system in the brain. In this perspective, it can be thus stated that Deep Learning is closer to the original definition of AI than Machine Learning is. The main advantage brought by this second class of algorithms is the automatic extraction and selection of the features useful to task fulfilment. This ability proves to be extremely powerful in high-feature problems, or generally when the combination of features that leads to the best performance is still not known.

While most ML algorithms are considered intrinsically interpretable by humans, DL models work at a higher-abstraction level, such that their internal behaviour is not well traceable. These models are thus often assimilated to **black-boxes**

[33, 34]. DL performs efficiently in the domain of high-dimensional data. In this regime, Deep Neural Networks typically outperform Machine Learning algorithms in most applications [35]. However, for low dimensional inputs, and especially in cases of limited availability of training data, ML algorithms can still produce optimal results [36]. Finally, while DL's performance may potentially outweigh human performance, problems requiring strong AI capabilities, such as common sense and intentionality, cannot be solved yet [37].

3.1 Artificial Intelligence and Medicine

Artificial intelligence is making its way into biomedical research and clinical practice, showing its potential in several applications, such as **personal screening**, **diagnosis**, **prediction of treatment response** and **prognosis**. The ability of these algorithms to integrate multiple streams of data, also from heterogeneous sources, and to use them for their continuous updating, is definitely encouraging the integration of AI in the health system as a powerful support in medical practice. AI, and specifically Deep Learning, possess the ability to "learn" a behavior after training, much like a doctor learns a clinical practice over time. Since experience derives from dealing with a wide variety of cases and situations, also AI algorithms need consistent amounts of data to improve their own skills.
Data availability is a recurring issue in DL: DL models are intrinsically data-hungry. In principle, this should not pose a problem in healthcare, where a great amount of data is constantly produced every day. The data format can vary from simple numerical values, such as concentrations resulting from blood tests, to modular information, such as medical records, or to images, like the outcome of a radiological examination. Therefore, an extraordinary amount of data would already be available for AI applications, taking into account not only data collected in recent years, but also the thousands of clinical histories of patients stored over the past years. The medical sector would thus be one of the most suitable for the application of DL technology. However, despite the huge amount of collected data in health care, only a small portion of them can directly feed DL models. The reasons behind such a data scarcity include the lack of adequate resources by health systems to share vast amounts of medical images, the lack of an automatized system for labelling and the bureaucratic procedure needed to approve the usage of the data[38].
Coupled with the data problem, the **ethical issue** is also slowing down the usage of AI within the medical field. Healthcare decisions made by intelligent algorithms have always been considered controversial. In order to shed light on the future ultimate entrance of AI within Medicine, the World Health Organization published the

guidance "Ethics and Governance of Artificial Intelligence for Health" [39]. Among the various points addressed, the first one directly concerns **human autonomy**, that is, how intelligent systems should be placed *in support* of physicians and not *in their place*, and how the decision-making task related to reporting should always be entrusted to the physician. These ethical issues bring to light a second concern, which is the growing apprehension of radiologists to lose their jobs.

Certainly, the introduction of AI in medicine will inevitably affect the work of physicians, however, the common hope among DL practitioners is that DL can lead to the evolution of the figure of doctors rather than their disappearance. Indeed, by being relieved of lower-level activities (such as segmentation activities), radiology physicians will thus be able to increase their work efficiency, allowing them to raise the number of cases processed within a time window, thereby reducing the waiting time for the performance of a diagnostic examination. At the same time, however, they will need to continue to monitor the activity of intelligent algorithms. Together with healthcare facilities, the European community is also making major strides in the area of artificial intelligence, working to introduce a CE marking system specifically dedicated to these systems. With such regulations, the use of AI in medicine could be revolutionized, allowing safer integration of automatic softwares into healthcare facilities.

3.2 Computer Aided Detection and Diagnosis

The use of images with artificial intelligent algorithms finds its main application in the development of **CAD** (Computer Aided Detection and Diagnosis) systems. Automatic interpretation of medical images has been investigated for the past few decades with the goal of minimizing operator dependence in the diagnostic process. The physician's analysis of a medical image can be influenced by various subjective factors, including the operator's experience but also fatigue and distraction caused by long hours of work. Furthermore, even intrinsic features of the image, such as its poor quality, can negatively affect the interpretation of the result of a radiological examination. The implementation of CAD systems could therefore help to improve the accuracy and the effectiveness of the diagnostic process and reduce the physician workload. Within the CAD models, two subgroups can be identified: **CADe** (Computer Aided Detection) and **CADx** (Computer Aided Diagnosis) algorithms. CADe methods are directed to the detection, localization and segmentation of elements within the medical image, and can therefore be used as a support tool for tracking lesions. Instead, CADx algorithms aim at the characterization of lesions and can be employed by radiologists to determine the nature of a given disease.

CAD system can make use of both ML and DL algorithms. Machine Learning classifiers have been exploited for **detection** [40] but also for **diagnostic purposes**[41]. Deep Learning models have great applicability both in **diagnostic**[42] as well as in **segmentation tasks**[43].

It follows, that according to the chosen model (ML or DL), the image workflow within the algorithm will be different.

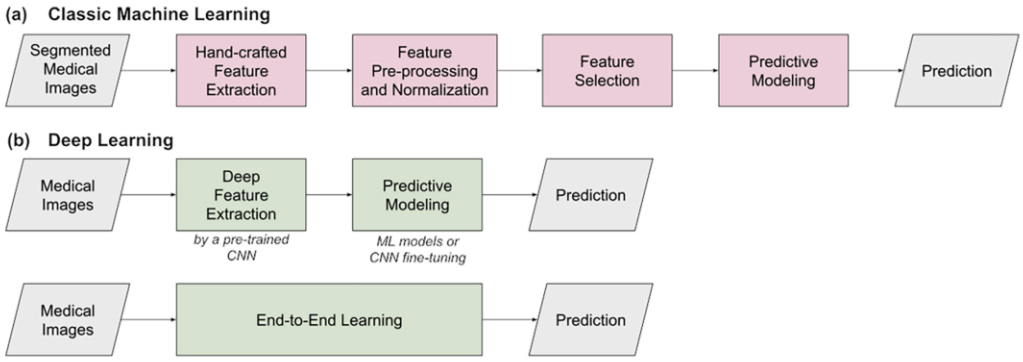


Figure 3.1: Workflow of different AI systems for CADx application.

The figure shows the possible strategies to deal with a predictive problem. As can be noticed, the use of a total Machine Learning approach requires starting from a pre-segmented image, different from the Deep Learning case. Furthermore, the use of ML classifiers for prediction requires feature extraction, which may or may not be done through Deep-type approaches. The main advantage of the use of an End-to-End Deep Learning strategy is to bypass the feature extraction and selection steps, speeding up the whole pipeline. However, the lack of knowledge concerning the elements believed meaningful for task completion leads to the algorithm resembling a "black box".

3.3 Machine Learning Approaches for medical imaging

The usage of traditional Machine Learning algorithms in CAD imperatively requires as primary step the definition of the features used to train the algorithm. **Radiomics** is the emerging field for the conversion of the image in mineable data

(features) and therefore plays a fundamental role in the employment of ML classifiers for CAD applications [44]. Radiomics is usually considered a CADx extension, as it performs organ characterization through the extraction of **qualitative features** from ROIs (Regions of Interest), previously manually or automatically identified. The computed features are characterized by a higher predictive value if compared to standard clinical predictors (such as age, sex, family history, and potential indicators of pathology not derived from the image). These features usually consist of morphometric characteristics (i.e., size, shape, and diameter of the lesion), as well as measurements of tissue heterogeneity (including first-, second-, and higher-order statistical descriptors), which efficiently represent the lesion. The whole collection of "radiomic" features is also known as "**tumour signature**". The two steps following feature extraction, are feature normalization and feature selection. The former helps to avoid model bias toward high-value features and is achievable through different methods, such as min-max rescaling or standardization. The latter serves to remove features with a low-diagnostic impact for the task to be performed. Once the maximum level of abstraction is reached, a ML classifier can be exploited for the desired task.

Among popular ML algorithms, **Logistic regression** is one of the most used, probably due to its simplicity and the capacity to lend robustness to classification.[45, 46]. Logistic regression employs the logistic function to differentiate binomial distribution and is usually used as a classifier.

Another frequently employed predictive models are **Decision Trees**; they offer a substantial advantage as they reproduce human reasoning by choosing to take a decision (a ramification) when the hypothesis is confirmed ("if-then"). If more than one decision tree is used to build the prediction model, the algorithm becomes **Random Forest**[47], which has usually higher performances. Moreover, unlike standard decision trees, in random forests only a subset of the total number of examples is used to train each tree within the forest (*bootstrapping*) and each tree is trained by random selection of all available features($feature\,bagging$). The result is an uncorrelated forest of trees whose prediction is more accurate than that of any single tree.
Support Vector Machines (SVMs) are supervised learning models used for binary classification tasks that use a representation in the space of examples based on their features to find a law that can separate them based on their class. If an immediate division is not achievable, SVMs are able to map the input data into a multidimensional space where the examples are linearly separable. They are mainly used for CADx and have proven to be effective in the diagnosis of breast cancer, when coupled with an accurate method of feature selection [48].

Lastly, also **Artificial Neural Networks** (ANNs) belong to the category of ML algorithms employed in the classification of medical images [49]. ANNs are massively parallel systems with large numbers of interconnected simple processors. Their main feature is the ability to use interconnected layers to autonomously extract the features needed to accomplish a given task. They will be discussed in more detail in the next chapter, as they constitute the crossing point from the ML to the DL.

3.4 Deep Learning Approaches for medical imaging

In Machine Learning approaches, the search for distinctive features and their selection has a great weight in determining the final performance of the model, which therefore depends only partly on the type of classifier chosen. Thanks to Deep Learning models, this issue is left behind, in favour of the research of the optimal Neural Network architecture achieving the best performance for a specific task. Among the Deep Learning architectures employed in medical context there are Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Generative Adversarial Neural Networks (GANs) and Autoencoders (AEs).

Convolutional Neural Networks are the most popular neural network architecture for medical image processing. Their distinctive feature is the use of layers that implement convolution operation (linear) and activation layers (non-linear) that act as powerful extractors of the image characteristics. End-to-end CNN architectures directly associate images to a target class and have been employed to perform image classification tasks for both screening and diagnosis purposes [50, 51]. In particular, several CNN architectures pre-trained on large natural image data sets, such as ImageNet, have been used to classify medical images by making use of already trained layers to overcome data scarcity problems [52]. Moreover, Convolutional Neural Networks show their potential also in segmentation tasks, and therefore can play a role even in the CADe field. The U-Net, a convolutional network introduced for the first time in 2015 [53], represents the starting point of various segmentation approaches in the medical field [54].

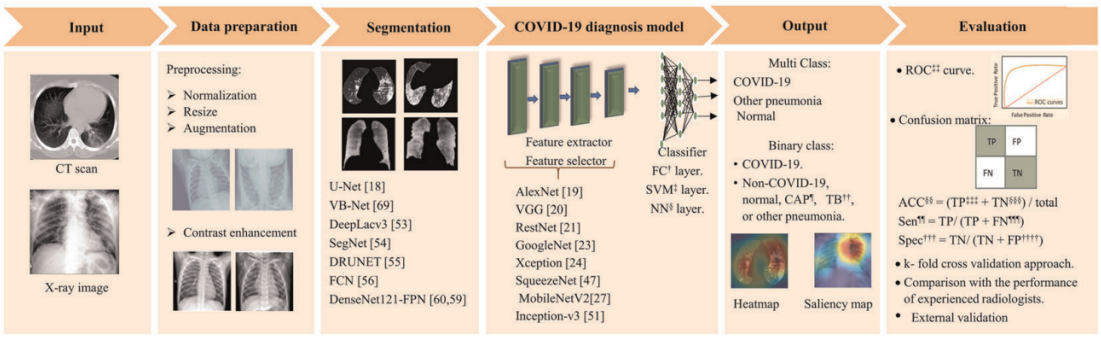
Recurrent neural networks have also been combined with CNNs to extract spatial-temporal features from imaging data series. RNNs introduce a recurrent layer whose main function is to provide the Network with a memory ability. Therefore, these Networks allow for the processing of new data while being aware of older inputs. In medical image analysis, they have been used to predict the shape

of anatomical structures in new images based on a sequential series of previous images [55].

Generative Adversarial Networks are frequently employed for the generation of realistic datasets starting from original medical images (data augmentation), addressing the problem of data scarcity. Their architecture is composed of two adversarial networks, one dedicated to new data generation and the other to its discrimination control. Lately, they have also seen application in segmentation tasks, in which they attempt to generate a segmentation map close to the ground truth [56].

Autoencoders are also frequently employed in medical image processing. These are unsupervised models whose hidden layers are used partly to compress the input into a low-dimensional representation (encoding path), and partly (decoding path) to reconstruct the original input from the learned features. They have been used for anomaly detection and also segmentation tasks [57].

As can be noticed, Deep approaches are able to perform different tasks directly from raw data, without the need for a pre-segmented image. Consequently, since most of them can effectively handle segmentation, they could be used in radiomics approaches as powerful ROI identifiers. The combination of DL and ML algorithms could therefore be a winning strategy for CAD systems, avoiding the black box problem usually associated with Deep Learning algorithms and speeding up the whole image pipeline (Figure 3.2).



†FC: Fully Connected; ‡SVM: Support Vector Machine; §NN: Neural Network; ¶CAP: community-acquired pneumonia; ††TB: Tuberculosis, ‡‡ROC: Receiver Operating Characteristic; §§ACC: Accuracy; ¶¶Sen: Sensitivity "the proportion of COVID-19 which was correctly identified"; †††Spec: Specificity "the proportion of those without COVID-19 which was correctly identified"; ‡‡‡TP: True Positive; §§§TN: True Negative; ¶¶¶FN: False Negative; ††††FP: False Positive.

Figure 3.2: Illustration of CAD pipeline for COVID-19 diagnosis based on the combination of classic Machine Learning and Deep Learning techniques.

Chapter 4

Neural Networks

Neural Networks constitute the first attempt of artificially mimicking the flow of input signals within the nervous system. In particular, artificial neural networks model the spiking activity of neurons in the brain through a *weighted sum* of the signals coming from the neighbouring neurons. At the same time, *neuronal plasticity* is also replicated, i.e. the ability of neurons to strengthen or weaken their connections with the surrounding elements, changing their own activity.

4.1 Artificial Neural Networks

The fundamental unit of Artificial Neural Networks is the **Perceptron**, the first example of an artificial neuron. This model was presented in 1958 by the psychologist Frank Rosenblatt [58], as a continuation of the work of neurophysiologist Warren S. McCulloch and mathematician Walter Pitts.

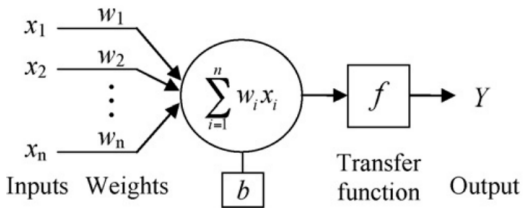


Figure 4.1: The model of an artificial neuron.

As its biological equivalent, the perceptron works as a processor, combining the inputs to produce an output signal. Each link with the input units has its own weight and contributes differently to the Perceptron output.

In particular, the output is produced as follows: each input x_i is multiplied by the weight of its connection w_i; a bias b is then added to the output to mimic the effect of the spiking thresholds in biological neurons. This sum goes through a transfer function f, determining the Perceptron output. As a result, the relationship between input and output is described as:

$$Y = f(\sum_{i=1}^{n} w_i x_i + b) \tag{4.1}$$

The Perceptron is a single-layer network that implements a linear type classification. Its capabilities are therefore extremely limited, being able to distinguish only linearly separable elements. In order to achieve more complex tasks, multiple Perceptrons are assembled to compose a network (ANN).
The minimal architecture of any ANN consists essentially of three types of layers: the **input layer**, the **hidden layer** and the **output layer**. The input layers receive the examples that are intended to be processed within the network, and their dimensionality depends on the shape of the input data to the algorithm, while the output layer yield the classification for the input example, and has a number of neurons equal to the number of possible categories to which the input can be assigned. Intermediate layers, also called hidden layers, connect the input layer to the output layer through a series of connections characterized by weights. Hidden layers serve to extract increasingly abstract and complex representations of the input, allowing the identification of the salient features from the input data needed to solve a given task.

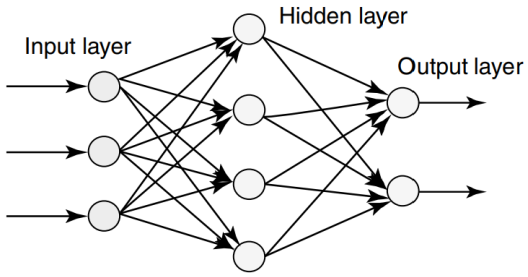


Figure 4.2: Artificial Neural Network.

The Perceptron solved classification problems with a single hyperplane; thus, a multilayer network uses multiple hyperplanes to solve elaborate multi-class classification tasks. As the number of hidden units increases, so does the number of connections, and the overall model complexity rises as well. If that is the case, there is a transition from Artificial Neural Networks to Deep Neural Networks, and the traditional Machine Learning field is left in favour of Deep Learning.

4.2 Learning and Neural Network Optimization

Neural Networks are provided with different training modes depending on the task to be implemented. **Classification** and **regression tasks**, the most popular in the medical field for diagnostic and prognostic purposes respectively, employ supervised learning. In contrast to *unsupervised learning*, where examples are the only input required for an algorithm to return an internally generated categorization (Clustering), *supervised learning* needs both examples and their correct labels to learn externally provided classifications. According to Rosenblatt mechanism [59] the NN training process requires the weights of connections to be modified in order to bring the output label closer to the correct one. Therefore, the training model requires a **loss function** $R(\hat{y}_i, y_i)$, sometimes called **cost** or **fitness function**, to assess how close the model prediction $\hat{y}_i$ fit the desired value y_i. The training process aims at minimizing the loss. Among the most popular loss functions there is the Mean Squared Error (MSE), defined as:

$$R_{MSE}(\hat{y}_i, y_i) = \sum_{k=1}^{K} (y_{i,k} - \hat{y}_{i,k})^2 \tag{4.2}$$

Also, Cross Entropy (CE) is frequently employed in DL approaches:

$$R_{CE}(\hat{y}_i, y_i) = -\sum_{k=1}^{K} (y_{i,k} \log(\hat{y}_{i,k})) \tag{4.3}$$

Minimisation of the CE is optimal in situations which require accurate estimation of small probabilities and is suited to predict class probabilities, whereas MSE is more suited to predicting values [60]
The goal of training is to find a set of parameters θ that minimize $R(\theta)$. Since this leads back to a minimum search problem, the most immediate strategy is to solve it through a derivative operation. Therefore, to obtain the optimal parameters, a system based on derivatives of loss functions for each parameter theta has to be solved. The algorithm which implements the search of the loss function minimum is commonly known as **Gradient Descent**. The gradient of the loss function, computed each time during the training process, provides the direction in which the function has the steepest rate of increase. Each parameter θ is thus updated in the negative direction of the gradient, with an arbitrary step size, identified by the

learning rate hyper-parameter. Generally, the space of the network parameters, i.e. the domain where the search for the minimum takes place, is multi-dimensional, and depends on the number of parameters defining the system.

Mathematically, the correction of the parameters at each update is formulated as

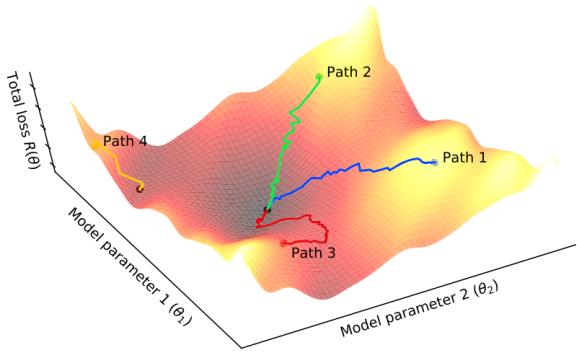


Figure 4.3: Gradient Descent representation.

follows:

$$\theta = \theta - \alpha \cdot \frac{\partial R}{\partial \theta} \tag{4.4}$$

where θ stands for each learnable parameter, α stands for the learning rate, and R stands for the loss function. It is of note that, in practice, the learning rate is one of the most important hyper-parameters to be set before the training starts.

The way the correction extends across the network from the last layers back to the first one is also known as **backpropagation** [61].

Usually, for reasons such as memory limitations, the gradients of the loss function with regard to the parameters are computed by using a subset of the training dataset called mini-batch, and applied to the parameter updates. This method is thus called **Mini-batch Gradient Descent**, whose mini-batch size represent an additional hyper-parameter to tune. Many improvements to the gradient descent algorithm have been proposed and widely used, such as RMSprop and Adam. Often in simple models, the initial parameters θ are simply taken to be random numbers. In modern neural nets, the parameters are instead frequently initialized to those of other networks that have already been trained on standard and huge image dataset, like ImageNet. This practice is known as **transfer learning**. One of the most frequent problems occurring at the end of the training is **overfitting**, which consists of an excessive adaptation to the training dataset. Among the strategies to avoid overfitting, such as the application of penalties during the training phase or the

imposition of the process to stop when the loss on the validation set increases, Data augmentation is also frequent. Using this technique, low variability is overcome by adding mock but similar examples to the train dataset, preventing the model from overfitting small amounts of data. Cross-validation technique is also helpful, as it allows the identification of any biases present within the possible combinations of data composing the training set.

4.3 Convolutional Neural Networks

Convolutional Neural Networks are deep learning architectures which make use of **convolution** as feature extraction operation. These models have become dominant in various computer vision tasks and are attracting interest across various domains, including **radiology**. Their main advantage is the ability to automatically and adaptively learn spatial hierarchies of features, from low- to high-level patterns. Moreover, the convolution algorithm captures prominent features preserving spatial relations within the image. The degree of feature abstraction derives from the number of convolutional layers: first layers identify simple patterns, such as lines or waves; moving towards deeper layers, the network recognizes more complex shapes, such as well-defined objects. Differently from standard ANN, where each neuron is connected to all the units of both the previous and following layers, CNN neurons have *sparse connections*, as the convolutional filter is normally set to be of smaller size than the input and consequently only a local patch is connected to one pixel of the next layer. This means that fewer parameters need to be calculated and stored, which improves computational efficiency. Moreover, kernel parameters may be shared by more than one input/connection, reducing the total amount of independent parameters (*weight sharing*). This leads to an overall reduction in the number of parameters defining the model; however, the deep structure of CNN usually implies that the overall parameter count will be higher for CNN.

A standard CNN architecture consists of several convolution layers and a pooling layer, followed by one or more fully connected layers (FC). The pooling and convolutional operations presented below apply to 2-D CNN (Equation 4.5); however, equivalent operations also apply to 3-D input volume data.

Convolutional layers perform feature extraction. They make use of a feature-specific filter (**kernel**) to investigate the different areas in the image in search of the defined pattern. Consequently, the elements involved in the convolution operation

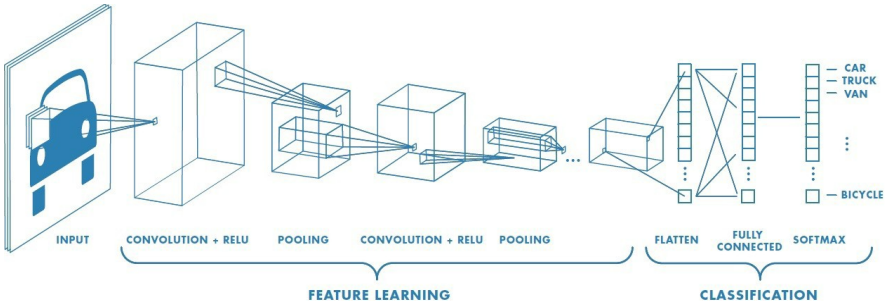


Figure 4.4: Example of Convolutional Neural Network architecture.

are the image pixel $I(x, y)$ and the filter $F(x, y)$ of K x K dimension:

$$H(x, y) = I(x, y) * F(x, y) = \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} I(i, j)F(x - i, y - j) \tag{4.5}$$

The output of each single convolution operation is a scalar value, as high as the investigated sub-section of the image presents the searched pattern. The convolution is repeated until the whole image has been covered: in order to achieve it, each time the kernel is moved to a new sub-area of the image. The result of the investigation across all the image is a **feature map** $H(x, y)$ (also called heatmap) for the searched pattern.

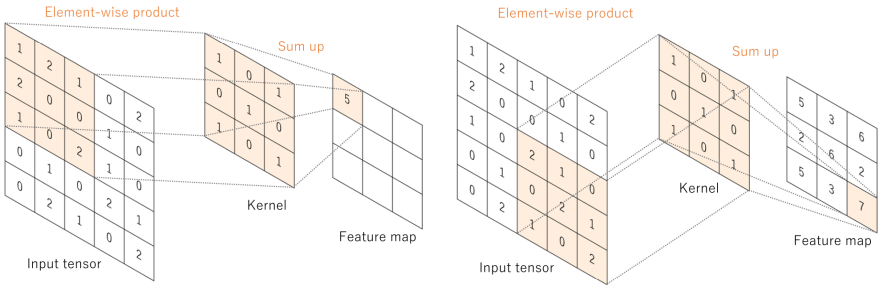


Figure 4.5: Convolution computation across the image

The resulting matrix has therefore reduced dimension if compared with the original image: the repetition of convolutional layers across the network leads to the reduction of the resolution in favour of more informative features.
Output dimensionality depends on three factors:

- **Kernel depth**: stands for the number of filters used for the convolution, so the numbers of researched patterns;
- **Stride**: stands for the number of pixels separating contiguous areas of the image to be investigated;

- **Zero-padding**: stands for the number of pixels added at the image border to preserve image input size.

The final feature maps dimensions (M_x,M_y,M_y) are computed as follows:

$$M_x = \frac{I_x - K_x + 2 * P_x}{S_x} + 1 \quad M_y = \frac{I_y - K_y + 2 * P_y}{S_y} + 1 \quad M_z = n^{\circ}\text{of filters} \tag{4.6}$$

where (I_x,I_y), (K_x,K_y) are the image input dimension and the kernel dimension respectively, while (S_x,S_y) stand for horizontal and vertical stride, and (P_x,P_y) for the padding.
Outputs of a linear operation such convolution are then passed through a **non-linear activation** function. In Deep Neural Networks, the most used activation function is **Rectified Linear Unit**, also known as **ReLU** (*f(x)=max(0,x)*), since it helps to avoid the vanishing gradient problem following multiple training iterations [62].

Pooling operation is placed immediately after the activation layer. It provides features with displacement-invariance and reduces their dimensionality. Among the possible reduction functions (Average-pooling, L2-normalization-pooling), **Max pooling** is the most commonly used: it preserves only one of the features related to a specific sub-region of the activation map. Similarly to convolutional layers, also for pooling layers it is possible to define a pooling filter characterized by a certain size, stride and padding.

The output feature maps from the last pooling layer are usually flattened and given as input to a **Fully Connected layer** (FC), which returns the final predicted value. Each layer of FC employs ReLU as activation function, exception made for the last layer. The activation function of the last layer is usually a **softmax**, which transforms the output values into the probabilities for each example to belong to the corresponding target class.

4.4 Fully Convolutional Neural Network

The typical use of Convolutional Neural Networks is on classification tasks, where the output of an input image is a single class label. However, in many visual tasks, especially in **biomedical image processing**, the desired output should include localization, i.e., a class label is supposed to be assigned to each pixel.
The main reason that prevents a standard CNN network from being used to directly

achieve segmentation is the usage of the final FC layer. This layer requires the input to be flattened, thus losing any spatial information preserved by the convolution operations.
In order to overcome this issue, Long et al.[63] proposed to replace the FC layer with a convolutional layer, thus obtaining a **Fully Convolutional Neural Network (FCNN)**. However, the function of this last layer differs from those used in the rest of the network, as it is not intended for the recognition of a specific pattern, but it aims at condensing together the feature maps extracted before to generate an activation map representative of the whole extraction process. The kernel of the last layer will therefore have unit dimensions and depth equal to that of the input feature maps tensor. The **heatmap** resulting at the end of the network is at the origin of the segmentation process.

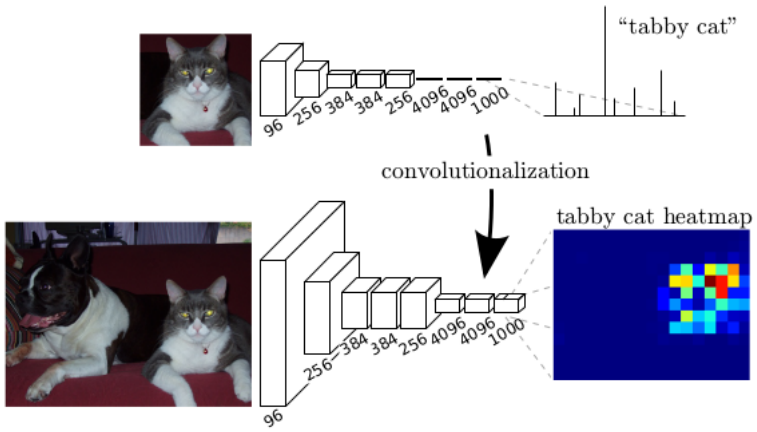


Figure 4.6: Fully Convolutional Neural Network (FCNN).

As shown in the **Figure 4.6**, although it is not yet able to correctly identify the contours of the image, it can be seen that the greater activation of the pixels is recorded in correspondence with the object of interest (the cat in the original picture).

The main issue of this feature map is its size: the convolution process inevitably leads to a resizing of the original image, condensing its information as it proceeds through the network, a process known as "**downsampling**". To obtain a segmentation mask to be superimposed on the original image, it is therefore necessary to restore the original dimensions of the image, and, at the same time, to retain the information content extracted from the previous layers: this is achievable thanks to layers performing "**up-sampling**" operations. Among the possible techniques

performing "up-sampling" operations, there are those involving layers which implements the inverse operations of pooling and convolution.

In particular, **Unpooling** restores the original image dimensions by mapping the content of the input map to a larger one. This will result into an enlarged but distributed activation map, where activations occur in enlarged approximations of the areas where the desired pattern is actually found. Examples of upsampling methods are Nearest-Neighbor, Bed of Nails, Max Unpooling. Instead, the **deconvolution** process (also called **transposed convolution**) is meant to densify these scattered activations. Through deconvolutions, activations strictly related to the classes defined by the filters are amplified, while noisy ones from other regions are effectively suppressed. Thanks to the combination of unpooling and deconvolution, the network generates accurate segmentation maps.

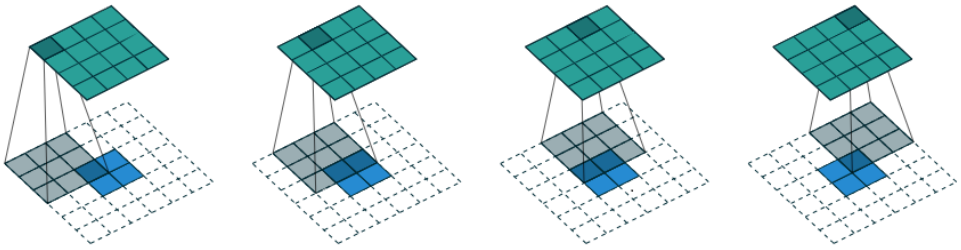


Figure 4.7: Example of a transpose convolution with 3x3 kernel and unit stride over a 2x2 input padded with a 2x2 border of zeros.

However, there are two major problems still affecting the FCNN's performance: firstly, the proposed architecture cannot preserve the finer details of the image: the final mask is still a coarse representation of the desired output; secondly, not all the generated masks are correct. In order to overcome the first issue, Long et al. [63] proposed the use of **skip connections** to combine the upsampled activations with the downsampling activations, improving the accuracy of the border segmentation. The second problem was solved subsequently by the work of Ronneberger et al. [53]. He realized that the upsampling steps provided by FCNNs architectures were too approximate, as they made use of a minimum number of layers to obtain the segmented mask from the decoding path output , resulting in loss of segmentation accuracy and insufficient integration of context information.. Therefore he proposed a new architecture, called **U-Net**, whose main feature was the usage of the same number of convolutional layers in upsampling and downsampling. The skip connections were then installed between each level of the upsampling layer and the symmetrical downsampling layer. The above-mentioned improvements make U-Net

more accurate in the aspect of pixel positioning and segmentation. The structure of U-Net is described in detail in the following section.

4.4.1 U-Net

U-Net entered the world of Fully Convolutional Neural Networks only recently, in 2015. Olaf Ronneberger, Philipp Fischer, and Thomas Brox [53] resumed the work on Long et al's FCNNs and developed the first FCNN intended for a segmentation task in the biomedical field. The architecture is shown in **Figure 4.8** and it consists of a **contractive path** (encoding path) and a symmetrical **expansive path** (decoding path), whose similarity to the letter "U" is an inspiration for the network name.

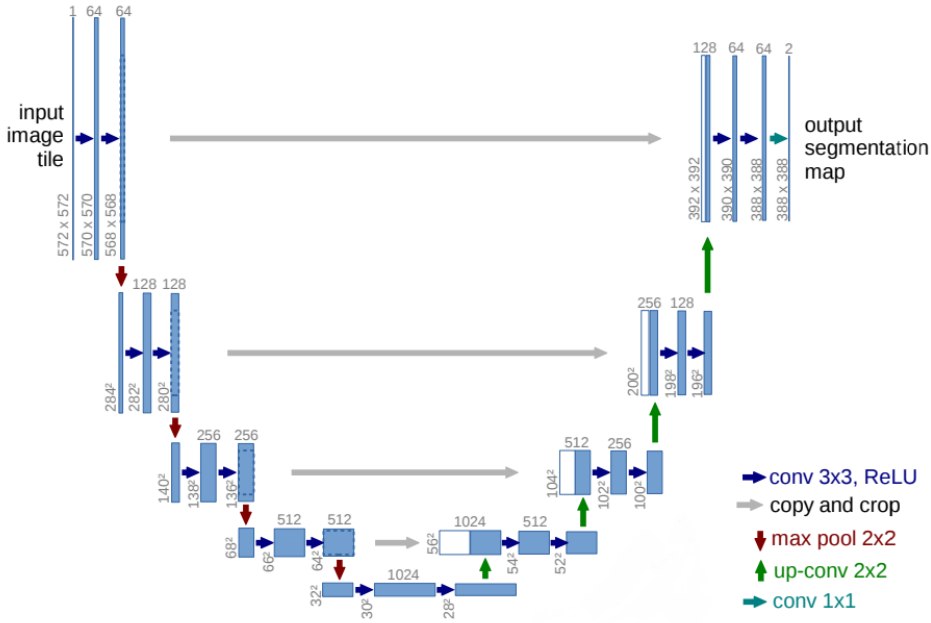


Figure 4.8: U-Net architecture

The **contracting path** follows the typical structure of a convolutional network and consists of repetitions of 2 convolutional blocks (of 3 x 3 kernels), each followed by a rectified linear unit (ReLU) and a max pooling with a 2x2 filter and stride 2. As downsampling proceeds, the number of feature maps increases while the map size decreases, resulting in a reduction of spatial resolution.
At the end of the encoding path, which is the **bottleneck of U-Net**, there are two

consecutive convolutions, which however do not foresee any max-pooling operation. Instead, the **expansive path** consists of an up-convolution step (of 2 x 2 kernel) which halves the number of feature maps, condensing their activations; a concatenation step realized through skip connections, which take up a part of the spatial information contained in the respective contraction step; 2 convolutions (of 3 x 3 kernel), each having a ReLU activation function. Being a full-fledged FCN, the last layer consists of a 1 x 1 convolution used to map the vector of features in each of the classes to which it belongs.
The performances of U-Net were evaluated in two different biomedical applications[53]: the identification of neuronal structures in electron microscopy images and the segmentation of cells derived from contrast microscopy. In both situations, the performances were found to be comparable, if not better than the techniques normally used to perform the same task.
[53]

4.4.2 Segmentation performances

The performance of a segmentation algorithm is evaluated by comparing the mask produced by the network, which may or may not be followed by a general post-processing operation, and the correct mask, usually drawn manually by experienced radiologists. Metrics computation usually demands a binary mask. It implies that in the case of multi-class segmentation, it is necessary to isolate the class of interest each time, estimate its metrics, and finally average over the total number of classes, to obtain the averaged trend on the image. By comparing two binary masks it is possible to refer to the case of binary classification and draw a confusion matrix as follows:

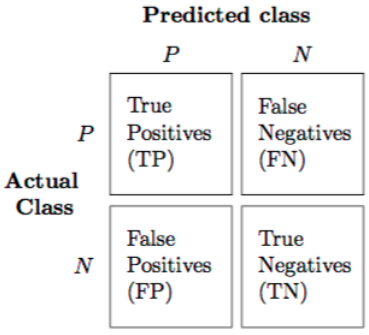


Figure 4.9: Confusion matrix.

where:

- Class 1 will be the positive class, belonging to the ROI
- Class 0 will be the negative class, usually the background

Similarly to how binary classification is evaluated, it is then possible to define Accuracy, Specificity, and Sensitivity. The metrics are computed on the basis of the correspondence of the single pixel, as the segmentation can be interpreted as a classification of the point units that make up the input image.

The first analyzed metric is **Pixel Accuracy**. It provides the percentage of pixels correctly classified within the image and can therefore be expressed as:

$$\text{Pixel Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.7}$$

In the context of segmentation, however, accuracy does not have a great discriminating power in distinguishing an optimal segmentation from a poor one, especially if there is a large disproportion between the ROI and the background. For this reason, the metric that is usually used in place of Pixel Accuracy is the **Intersection Over Union** (IOU), also called **Jaccard Score Coefficient** (JSC). This metric is computed as the ratio between the intersection of the masks and their union:

$$\text{IOU} = \frac{\text{Area of the overlap}}{\text{Area of the union}} = \frac{TP}{TP + FP + FN} \tag{4.8}$$

Together with the IOU coefficient, another widely-used segmentation metric is the **Dice Score Coefficient** (DSC):

$$\text{DSC} = \frac{2 \cdot \text{Area of the overlap}}{\text{Area of the union} + \text{Area of the intersection}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{4.9}$$

This metric is almost equivalent to the previous one, however, it tends to penalize segmentation errors to a lesser extent and provides a representation of an average performance trend of the algorithm.
Concerning Sensitivity and Specificity, e.g. the two most common binary classification metrics, they are usually not defined in segmentation tasks. In their place, Machine Learning practitioners typically use the metrics **Precision** and **Recall**. These are indicators of the tendency of an algorithm to over-segment or under-segment, respectively. It can be noticed, however, that the definition of Sensitivity is equivalent to that of Recall.

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN} \tag{4.10}$$

In the event that an elevated area of the image is segmented when it should not, a high number of FP is recorded and the Precision is lowered. Conversely, in the event that the algorithm fails to segment the area of interest, the FN rate rises and the Recall drops.

Chapter 5

State of the Art

5.1 Automatic Segmentation in Medical Imaging

Image segmentation is a classic problem in computer vision research and has become a hotspot in the field of image understanding.
Segmentation consists in the division of an image into several areas according to features such as grayscale, colour, spatial texture, and geometric shapes, so that each area shares similar features. The field of image segmentation includes semantic segmentation, instance segmentation and panoramic segmentation, each one requiring a different level of coarse-graining. However, the segmentation of medical images is regarded as a **semantic segmentation task**, where each pixel is assigned to a defined label, aiming at discriminating different anatomical structures (e.g. blood vessels, valves, organs, tumours). There is a high demand for automatic segmentation in the medical industry, since many medical procedures require imaging examinations (CT, MRI, PET, ultrasound) on daily basis. The isolation of the areas of interest from the background is frequently required in the medical field: it is performed by radiologists to estimate the dimensionality of anatomical structures within images; it is mandatory in radiotherapy, for the definition of patients' treatment plans; but it is also necessary for the reconstruction of three-dimensional volumes in medical computer vision applications. In addition, the elimination of the background from the image allows for a more accurate analysis of what is contained within the ROI. Newer applications agree in identifying **Fully Convolutional Neural Networks** as the best choice for segmentation operations. Depending on the technology behind the image generation and the anatomical district of interest, several solutions have been developed. The starting point for many of these applications was U-Net, but many others made later use of more complex algorithms. These models find a wide applicability in healthcare:

from the segmentation of whole organs, to the isolation of vascular trees, or the segmentation of tumour masses.
Whatever the case, it should be pointed out that DL-based segmentation applications that make use of CT and RMI images are much more numerous than those employing ultrasonic images. Due to low contrast, point noise, low signal-to-noise ratio, and artifacts typically associated with ultrasonic images, automatic segmentation of US images poses a considerable challenge for AI algorithms, compared to other acquisition modes [64]. Being the gynaecological ultrasound sector the one of interest in the present work, the next section will provide a brief review of automatic algorithms employed for the segmentation of gynaecological structures in ultrasound images.

5.1.1 Automatic Segmentation of Ovarian Follicles

The following is a brief overview of approaches used for the automatic segmentation of ovarian follicles, sharing with the aim of the thesis project both the use of ultrasound images and the proximity to the anatomical district of interest. The reason for this choice is mainly due to the scarcity of direct approaches to the segmentation of adnexal masses.

In 1993 Muzzolini et al. [65] used a **split and merge segmentation** method to isolate ovarian follicles from 2-D ultrasound images. In split and merge segmentation technique, the entire image is initially iteratively split into quadrants, following a parent-child relationship; the sub-blocks derived from the final split are subsequently condensed together to form the segmentation according to homogeneity criteria. In Muzzolini experiment, a simulated annealing algorithm was responsible for the control over the split and merge operation (Metropolis), aimed at the minimization of an energy function depending on the resolution of the size of the texture blocks within the image.
Sarty et al. [66] proposed in 1998 a semi-automatic approach to detect both internal and external walls of ovarian follicles in US images. Their technique consisted of three steps: the interactive definition of an A-ROI (anular region of interest) placed around the follicle, the minimization of a loss function based on edge prominence and direction to find the inner wall; the detection of the outer border based on the predicted inner wall.
Potocnik and Zazula [67] made use of a **region growing** algorithm for the automatic segmentation of follicles. This algorithm requires prior identification of the points that belong for sure to the area of interest; surrounding pixels are gradually added to the first identified according to affinity criteria defined in advance. The process stops when none of the surrounding pixel can be further added to the

growing segmentation. In this experiment, watershed technique and thresholding methods were initially used to automatically detect seed points on pre-filtered images. The region-growing algorithm was then used as second step to accurately define the boundaries of the follicles. The final step was aimed at eliminating non-follicle detected regions, based on a priori knowledge of ovarian follicle characteristics (shape, size, localization).
In most recent years **Active Countours** have also been employed for detection of ovarian follicles, even without the use of any previous edge based method [68].This models aims to achieve segmentation through the minimization of an energy functional that depends on two components: one used to control the deformability of the model, the other dependent on the force of attraction to the desidered contour. Among other attempted approaches was **K-means Clustering**, which was efficiently applied to the ultrasound image for the detection of the follicles after a pre-processing phase based on Colour-space and Wavelet transformations [69].
Deep learning techniques have been employed in follicular segmentation tasks just in recent years. Two studies in 2019 employed End-to-End DL approaches to ovarian structures segmentation. D. S. Wanderley experiment [70] was based on the use of a **U-Net** for the segmentation of the follicles and the surrounding ovary. The study resulted in 95% and 85% DSC for the two components, respectively. The study of Marquez Sonia et al.[71] in 2019 made still use of a U-Net, but was aimed at assessing the importance of defining the hyper-parameters of the network and the presence of a post-processing phase on the performance of an entirely Deep architecture. In 2020 Haoming Li at al. [72] proposed a **CR-Unet** model destined to follicle segmentation whose backbone is a standard U-Net, but with spatial RNN modules embedded between the encoder and decoder path. The ovary and follicles segmented masks achieved a Dice Similarity Coefficient (DSC) of 91.2% and 85.8%, respectively.

Nevertheless, although the follicle segmentation still shares similarities with adnexal masses, the comparison cannot be ever fair, both because of the possible presence of solid or semi-solid components within the cyst, and for the difficulty of segmentation of an object that does not have a well-defined pattern (as it instead happens in the case of the ovarian follicles, which are easy to recognize because consist of anechogenic segmentations grouped in close proximity to each other).

5.1.2 Automatic Segmentation of Ovarian Adnexal Masses

With regard to ovarian cancer identified by ultrasound examination, the interest is currently focused on their classification in diagnostic terms rather than on their segmentation [73, 74]. However, a wide-spreading idea is that the preventive isolation of the adnexal mass may favour a subsequent diagnostic classification step,

no matter if based on a CNN network or on a radiomic analysis followed by a ML classifier.

One of the first and few approaches to the segmentation of ovarian masses in ultrasound images was the work of Zimmer et al. in 1996 [75]. This study used a bivariate extension of an entropy-based thresholding method known as **"Minimum Cross Entropy thresholding"**(MCE), in which the variable for segmentation (grey level) is replaced by a linear combination of the grey level and the local entropy. The method was tested on ovarian cysts and was used as segmentation step for a malignancy detection of ovarian masses in a later work [76]. Over the years, few other approaches have followed, mainly based on pre-processing pipelines and thresholding operations [77] or the use of Active Contours[78].
Among further non-Deep approaches for the isolation of ovarian masses there is the **OvAi Focus** segmentation module,developed in SynDiag and representing the current baseline reference within the company. The product is commercially available and is already being used by physicians to extract information from ultrasound videos. As for the possibility of achieving the segmentation of the adnexal masses with a **Deep model**, the above-mentioned studies of Wanderley [70] and Marquez[71] can already be considered as valid starting points. Their works demonstrated that these techniques can be exploited for the segmentation of the gynaecological structures and could therefore also be tested for ovarian masses.
A unique example of a Deep Learning implementation study focused on the segmentation of ovarian masses is the experiment of Juebin Jin [79]. The study is aimed at comparing the performance of different Deep Fully Convolutional Networks for the segmentation of ovarian neoplasms. Among the tested architectures are the classic **U-Net**, **U-Net**$^{+++}$, **U-Net** with **Res-Net** as encoding network and **CE-NET**. The U-Net^{+++} is a U-Net characterized by a greater depth, with a dense network of connections between layer and layer.The use of Res-Net as a backbone for U-Net means that for this model, in addition to the long skip connection between each level of contraction and expansion paths, also local skip connections are introduced between the convolutions of each level. These additional skip connections help to achieve a smooth loss curve and help to avoid gradient vanishing and explosion. Lastly, CE-NET differs from the standard U-Net architecture for the introduction of a context extraction module at the bottle-neck of the NN.
The performance of the different algorithms has not only been evaluated on IOU and DSC metrics, but also on the basis of the comparison of radiomic features extracted from the correct mask. The average values of IOU and DSC for all networks were above 70% and 80% respectively. The highest values were recorded for CE-NET and U-net with Res-Net as backbone.
This study, performed on ovarian tumours of different histotypes, constitutes a

starting point for the application of fully convolutional networks for the segmentation of ovarian tumours, suggesting that they can be effectively used for the development of more complex diagnostic pipelines.

Chapter 6

Operative Context

6.1 SynDiag

As part of the **Polytechnic of Turin Incubator i3P**, **SynDiag** was founded in 2019 by Daniele Conti (CEO), Rosilari Bellacosa (Research and Developer Director) and Federica Gerace (Artificial Intelligence Director). The mission of the company is the development of AI-based software to support radiologists and gynaecologists in the diagnosis of ovarian cancer. These tools aim, on the one hand, at the minimization of the inter-operator dependency errors, frequent in sonography. On the other hand, they can favour early diagnosis, providing the physician with accurate information extracted from the image. Not only the patients could thus benefit from these products, but also healthcare facilities, since the optimization of diagnostic tests would allow more efficient targeting of hospital resources. Physicians' opinion is considered of fundamental importance for the development of the product, hence the relationship with the doctors is frequent, both in the evaluation of the performance of algorithms and in the design of the software interface. Currently, several services are already made available on the market:

- Education: SynDiag hosts an **Archive** composed of clinical cases from several centers of excellence; the huge collection of ultrasound videos and their histological reports allows the development of a learning algorithm, helpful for less experienced doctors and called **OvAi Tutor**;
- Management: the organization of the data flow of reports between doctors and patients is handled by the **DiGyn - Management** platform; while the **DiGyn - Telemedicine** platform has been conceived in order to favour the exchange of opinions among physicians on various clinical cases;

- Diagnosis: **OvAi Focus** is aimed at identifying the attached masses and their irregularities, it is currently well performing in the analysis of serous cysts; **OvAi X** is a proper diagnostic tool of ovarian lesions.

6.2 Baseline reference: Focus Algorithm

The segmentation algorithm currently employed in SynDiag for ovarian cyst segmentation is a module within OvAi Focus, i.e. the above-mentioned tool designed in the company for the processing of ultrasound videos.

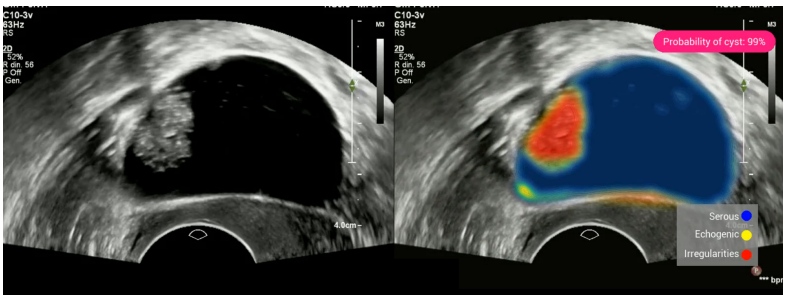


Figure 6.1: OvAi Focus interface.

Aside from isolating ovarian masses, the software can also perform cleaning, abstraction, and representation of the information content in order to assist the physician in analyzing ultrasound videos. A basic explanatory workflow is shown in Figure 6.2.
The software is capable of processing both Doppler and simple B-mode videos, but exclusively in the latter case it addresses cyst segmentation, which is performed after a dedicated AI-based module attesting the probability of the presence of the lesion. In particular, in the event that the probability of the mass is higher than a certain threshold, the internal mass components are analyzed and a colorimetric map is created to enhance their different localization within the lesion. At the end of this analysis on the individual frames, the video is reconstructed with the extracted information and their visual representation. The true reference for segmentation is thus the OvAi CV module (highlighted in red in Figure 6.2), which is based upon a non-DL algorithm. One of the main difficulty this system encounters more frequently is the segmentation of acoustic shadows, which, being anechogenic areas of the image, can frequently be likened to serous material within the cysts. Hence the need for the development of a new algorithm capable of minimizing errors relative to this issue.

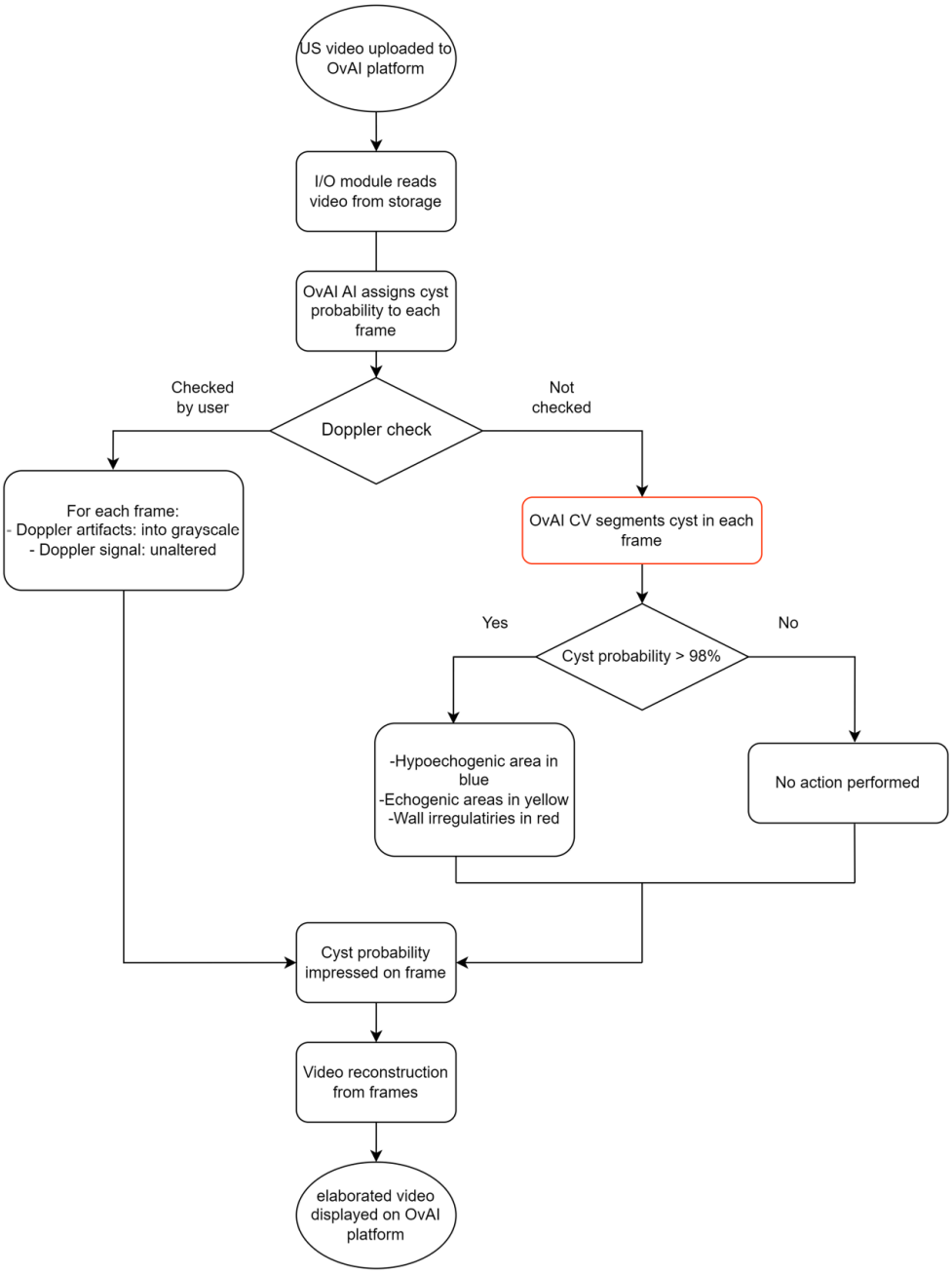


Figure 6.2: OvAi Focus workflow.

6.3 Project Hypothesis and Goals

Considering the rapid spread of Deep Learning algorithms for the segmentation of anatomical structures in radiological images and their satisfying performances, this project of thesis proposes an alternative algorithm for ovarian masses segmentation, based on a **Fully Convolutional Neural Network** (FCNN), with the aim of overcoming some of the limitations of Ovai Focus. In light of a fair comparison, since OvAi Focus does not include pre-processing steps for noise filtering, these steps will be equivalently omitted in the DL model. Furthermore, it was decided to limit the case of interest to unilocular serous cysts just as a proof of concept. Indeed, they represent the easiest cyst morphology type which OvAi Focus still have some troubles to detect in presence of acoustic shadows or sharp curvature of the cyst elliptic shape 7.3.2. Considering the literature review, comparable or better performance is expected from the DL algorithm. The ability of Convolutional Neural Networks lies precisely in the automatic identification of the features needed to correctly isolate the ROI. Nevertheless, it is not possible to foresee the behaviour of the network regarding image artifacts, in particular the ones related to acoustic shadowing. Being generally anechogenic shapes with defined contours it cannot be excluded a-priori that the model will efficiently avoid their segmentation. Indeed, a great influence on final performance will have the dataset numerosity and the variability of the cases included in the study, as it frequently highly affects Deep Learning algorithms.

Although primarily intended for the improvement of the state-of-the-art in image segmentation of the company, the developed **segmentation pipeline** is nevertheless independent from that of OvAi Focus. Thus, the proposed approach involves the interaction with the data sources autonomously and the conduction of ad hoc experiments.
Indeed, the DL segmentation approach could be also employed to improve the detection of the ROIs, in such a way that the other modules in the detection pipeline, like OvAi Focus or Radiomics, can enhance their performances.

Chapter 7

Methods

7.1 Equipment and tools

The equipment used for the development of the thesis project has the following specifications:

- **Windows Subsystem for Linux (WSL)**: Ubuntu®20.04 and Miniconda3 vv. 4.12
- **Processor**: 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz 2.80 GHz;
- **Ram**: 8GB

The development of the model and the corresponding simulations requires instead the following softwares:

- **Microsoft Visual Studio Code (MVSC) 1.73.0**: main code editor;
- **VcXsrv 1.20.14** : open-source display server for Microsoft Windows, used to plot images from MVSC;
- **Gitlab**: open-source web platform, employed to maintain the code updated and accessible to other team members;
- **Amazon Web Services (AWS)**: **S3** storage cloud service served as clinical data storage, as well as result repository; **EC2** platform was exploited to run model simulations.
- **RedBrick AI**: a purpose-built application to support healthcare AI teams in the annotation of medical data; used to perform labelling.

7.2 Data Ingestion Workflow

The data provided for this thesis project come from a retrospective study conducted with some of the hospitals with which SynDiag is currently cooperating. In particular, clinical cases included in this study derive from Mauriziano Umberto Hospital (Torino,Piemonte).
These data are uploaded on the SynDiag main platform OvAi at the hands of physicians. Each clinical case usually consists of several ultrasound videos (B-mode or Doppler) and a series of other metadata (ultrasound evaluation, subjective assessment, histological report and others). The agreed **acquisition protocol** requires to provide at least two B-mode TVS scans of frontal and sagittal planes; some of the uploaded videos can also include power-doppler or eco-doppler signals. The echographic evaluation consists of a straightforward description of the adnexal lesion and includes some of the first clinician's hypotheses. Together with the subjective assessment, which is the first provided diagnosis, it contributes to the clinical picture of the disease. While it is rare for this type of information to be missing, it is not uncommon for the histological report to be absent, since not all of the patients undergoing a TVS exam require further investigation, such as a biopsy or a surgical operation. Regardless of the data completeness, at the time of data upload on OvAi platform, they undergo an anonymization process which ensures patients' privacy. As soon as the data loading request from a clinician is approved, three events take place:

- **Unique ID generation** : A univocal identifier is associated with the clinical case;
- **Metadata storage**: The metadata associated with the case are stored as objects in Mongo DB;
- **Video storage**: Each video is saved as an object within the AWS storage platform;

In parallel to the upload of new clinical cases, a *data curation* process also takes place; this procedure will subsequently increase the *data usability* for the different applications under development. First, the information about the new clinical cases is reported in a file, which collects metadata related to all the clinical cases from the different hospitals. This file is constantly updated upon the arrival of new data on the platform. Until now, as the data flow was quite limited, this was considered as the optimal solution to the internal usability problem; however, further improvements regarding the use of Database software for data management, will be taken into account in the next future.

The main features identifying a clinical case are reported as follows :

- **Case ID**: the unique ID associated with the clinical case;
- **External ID**: external ID of clinical case;
- **Item ID**: the unique ID associated with the object belonging to the clinical case;
- **Item Type**: the data type of object associated with the clinical case (it was introduced to allow doctors to upload also images);
- **Doppler**: flag identifying Doppler acquisition mode;
- **Histological Report**: the confirmed taxonomy (histotype) for the clinical case in question;
- **Subjective assessment**: first diagnostic hypothesis provided by the physician immediately after the TVS examination;
- **B /BOT/M**: the final diagnosis (in accordance with the histological report);
- **Overall morphology**: corresponds to the general morphology of the lesion(s).

Case_ID	External_ID	Link web	Item ID	Item Type	Doppler	Hystological Report	Subjective Assessment	B / BOT / M	Overall Morpho
zom9zvjM4-38GmJ8-3rRU	AL10_1952	https://ovai.syndiag.ai/projects/OoBgXI7K_O2en8sFnYRH6/clinical-cases/zom9zvjM4-	TodVhUPJkBk5BfSqu4OW5	video	1	angiosarcoma	M	M	multiloc_solid
			cg1QAzZ97LxsOKRDfZ4wz	video	0				multiloc_solid
			9MH98zyKY-8133DSFsPbg	video	1				multiloc_solid
SR6Z1oyA8dGwad7B98gec	DRN 02_1958	https://ovai.syndiag.ai/projects/OoBgXI7K_O2en8sFnYRH6/clinical-cases/SR6Z1oyA8d	jLsnv3r_JhbHhU9Tzqoa8	video	0	high_grade_serous_adenocarcinoma	M	M	multiloc_solid
			__yXi2ciMMB0jsGxvWcl2	video	0				multiloc_solid
			EK9m4iaF4bx7Py9qRQ6nu	video	0				uniloc_solid
F9mtRM6NoGBlYFto8RK1r	CP 03_1975	https://ovai.syndiag.ai/projects/OoBgXI7	n59tKn_9JI7sJEbcSMbyX	video	0	high_grade_serous_adenocarcinoma	M	M	multiloc_solid
CXOZvjJCzZ3BDDICNuHJ-	SC 08_1968	https://ovai.syndiag.ai/projects/OoBgXI7K_O2en8sFnYRH6/clinical-	6BC5m_sQtFi85ltLwERji	video	0	clear_cell_carcinoma	M	M	solid
			XLvrg46mK92ForRaFWJZS	video	1				solid
1vAjqRbwxf4H5r33Te25G	CGE 02_1997	https://ovai.syndiag.ai/projects/OoBgXI7K_O2en8sFnYRH6/clinical-	5V2g6m8o51VeGrYJS0EUj	video	1	borderline_serous	B	BOT	uniloc_solid
			nrNwGzdpRy0zHeU_GVPDI	video	0				uniloc_solid
b9izlNR5rximq7g0VSiSV	LG 03_1967	https://ovai.syndiag.ai/projects/OoBgXI7	mu0JT5LKl1TanqfXurz02	video	0	endometrioid_adenocarcinoma	M	M	multiloc_solid
Q086cgHGKGFYJZIqk2F3n	BL 03_1986	https://ovai.syndiag.ai/projects/OoBgXI7K_O2en8sFnYRH6/clinical-cases/Q086cgHGKG	kT6Ow72Iq2Ta1BimxbC1X	video	0	metastasis_group_1	M	M	multiloc_solid
			fMOPlw7v3UC-RPrWMRZkN	video	1				multiloc_solid
			BOzv-Ww9EQdVOpAYrxDQ1	video	0				multiloc_solid
ur1LOSCF83AQmvsYQJxea	RA 05_1967	https://ovai.syndiag.ai/projects/OoBgXI7	uguale a BL 03_1986	video	0	clear_cell_carcinoma			
XZU5obqwaNddvvkTZg4D(	CJ 07_1977	https://ovai.syndiag.ai/projects/OoBgXI7	HKvqVvJwYyf5sJe_7P_Qh	video	1	granulosa	M	M	solid
GM_WyQKVTMseoqIMIu9X	BIMI 06_1975	https://ovai.syndiag.ai/projects/OoBgXI7	l4wm_miHHCqZxGFpzDhQz	video	0	borderline_mucinous	BOT	BOT	multiloc

Figure 7.1: Data management of information related to clinical cases.

This file allows, on the one hand, to provide an overview of the data situation within the OvAi platform, on the other hand, to efficiently filter clinical cases for the development of a specific application.
However, along with the pure data, supervised learning models, such as Fully Convolutional Neural Networks, require appropriate labels to accomplish a given task. Labels are assigned to individual frames extracted from the video, since

the frames are usually exploited in the development of algorithms that abstract information from ultrasound videos. Generally, the labelling activity should be carried out by expert radiologists and gynaecologists. In the context of this thesis project, a considerable part of the labelling procedure, as will be clarified later, was carried out by myself.

There are different types of labels that can be assigned to frames extracted from the video; those relevant for the current project are listed below:

- **Morphological label**: The activity of morphological labelling consists in the analysis of the video ultrasound and in the subdivision of the different time intervals (and consequently of the frames) corresponding to the different morphology of the mass. Five categories have been identified, equivalent to those addressed by the IOTA group: uniloc, uniloc-solid, multiloc, multiloc-solid, and solid. The morphological label of a video comes in the form of a CSV file, one-hot encoded.
  The entire video is then assigned an Overall Morphologycal label, based on the following hierarchy:

  $$uniloc < multiloc < uniloc - solid < multiloc - solid \tag{7.1}$$

  The hierarchy is established in function of the risk of malignancy associated with the cyst morphology.

- **Segmentation label**: Segmentation labelling consists in the isolation of the mass components within the single frame. The lesion may contain serum, mucin, hemorrhagic content, papillary projections or solid components. In accordance with the purpose of the segmentation task, the label has different granularity. In the simplest case of cyst identification, as it is the one of the thesis, the output is a simple full-cyst perimeter binary mask.

Generated labels are then transferred back into the S3 storage, within a label-intended bucket, to be ready-to-use for the different AI projects under development. Below is a summary of the flow of data from their arrival in the platform to their use (Figure 7.2).

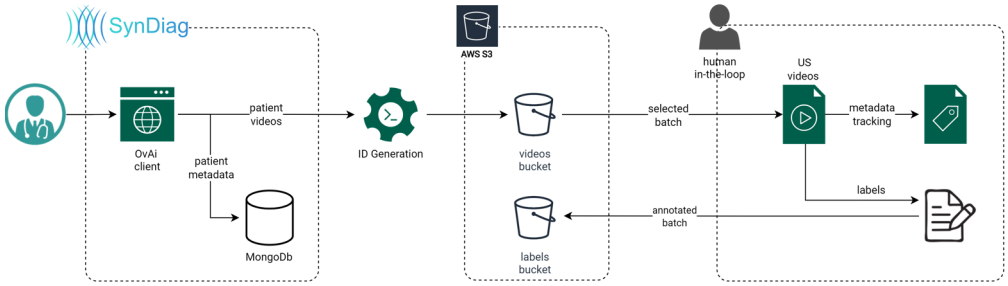


Figure 7.2: Data Ingestion Workflow.

7.3 Data preparation

The data preparation phase involves the retrieval of relevant data for the development of the DL model, i.e., the videos related to the chosen morphology and the ground-truth masks for segmentation necessary for NN training.

7.3.1 Data selection

At the time of the beginning of my internship, the available clinical cases within the OvAi platform amounted to 316.
As stated in **Section 6.3**, the cases object of the thesis project are unilocular serous cysts. Along with all other mass morphologies, also Doppler images were excluded. The Doppler signal, even if turned into greyscale, could prevent the correct cyst recognition. Moreover, Doppler is usually exploited to determine the vascularization level within solid components of adnexal masses. It is thus not so relevant for the detection of serous cysts.
A filter was therefore applied to the clinical cases in the metadata sheet in order to isolate only cases marked with "uniloc" as Overall Morphological label and without a Doppler flag. After the filtering, the number of available cases dropped to 35, corresponding to a total number of 70 videos. Subsequently, the extracted cases followed a more accurate control, aimed at verifying that the characteristics required for their use in the study were actually met. As a result of this evaluation, 9 cases and 5 videos not matching the required features were removed, in particular, due to the following issues:

- the presence of anechogenic structures attached to the main cyst, which could be assimilated to a multilocular cyst type;

- the presence of solid, mucinous or haemorrhagic components within the cyst.

Some examples are shown in **Figure 7.4**.

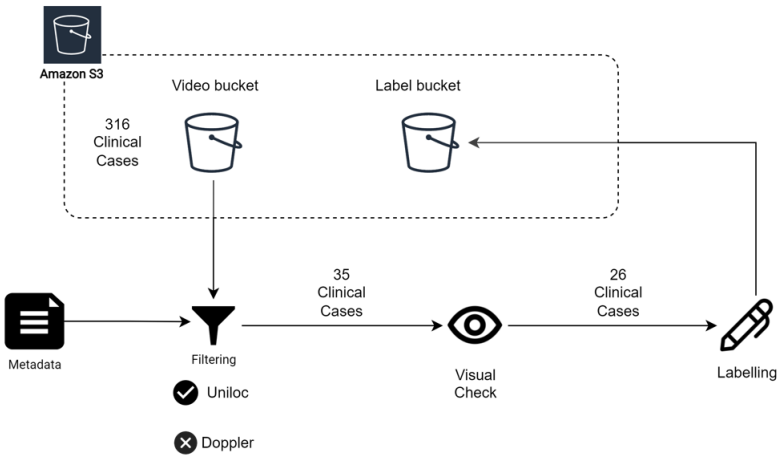
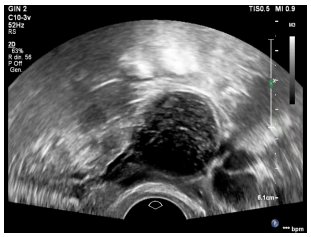
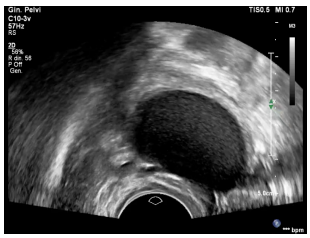


Figure 7.3: Data preparation pipeline

(a) Multiple anechogenic zones



(b) Hemorragic cyst



(c) Endometroid cyst

(d) Semisolid cyst

Figure 7.4: Ultrasound modes.

The final dataset consisted of 26 clinical cases and a total of 50 ultrasound videos.

7.3.2 Data labelling

The segmentation labelling activity focused on the generation of the ground truth masks needed for model training. At the time of my arrival in SynDiag, a dataset of masks of unilocular cysts serous was still to be built. Indeed, until that time, the priority had been given to the examination of the different tissue components present within the mass, and therefore only cysts with at least one solid component had been segmented.
Since OvAi Focus is a Syndiag tool already available on the market for the detection of ovarian cysts in real-time, it was decided to make use of its segmentation algorithm to have masks immediately available.

Automatic labelling

The videos belonging to the cases of interest were therefore given as inputs to the automatic segmentation algorithm for the extraction of the masks (OvAi CV module - Section 6.2). From the video, one mask every three frames was selected, within the interval in which the cyst appeared in its entirety within the video.

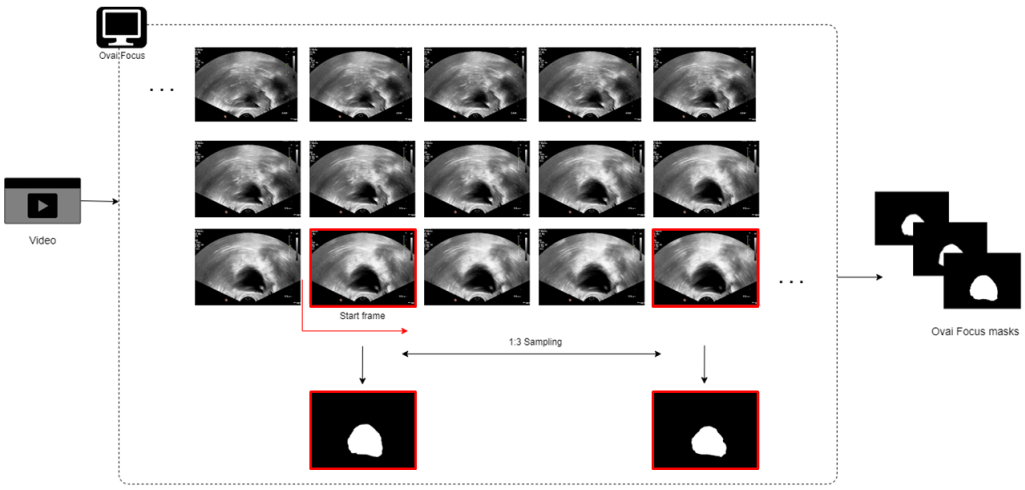


Figure 7.5: OvAi Focus automatic mask generation.

The total amount of extracted masks was 4887. The result of OvAi Focus segmentation was however only partially satisfactory: in several of the proposed cases, the algorithm committed multiple errors during the segmentation process. This is indeed partly the motivation behind the inception of this thesis project 6.3. The main detected flaws are presented in Table 7.1 according to the following classification:

- **Segmentation of acoustic shadows and nearby anechogenic zones**: the algorithm correctly identifies the cyst but extends the binary segmentation also to acoustic shadows and anechogenic zones that are in the vicinity of the cyst;

- **Inhomogeneity errors**: even when the cyst is serous, it may not appear completely anechogenic within a single frame; the causes are the transient nature of the images, extracted from ultrasound videos, and the artifacts that characterize the ultrasound examination;

- **Cut errors**: occasionally, the algorithm does not identify the full cyst but only part of it, separating it with well-defined cuts;

- **Inferior border errors**: the algorithm often struggles to recognize the lower curved edge of the cyst, while correctly detecting the entire upper part;

- **General detection errors**: in some frames, the algorithm fails to recognize the cyst and thus detects a different object.

The distribution of errors within the masks to be corrected in the source dataset is shown in Figure 7.6.

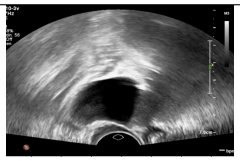
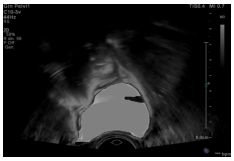
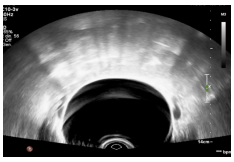
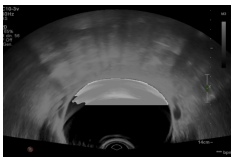
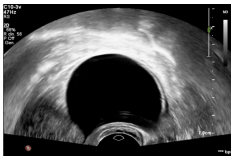
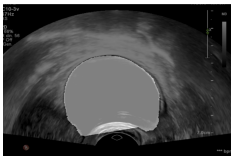
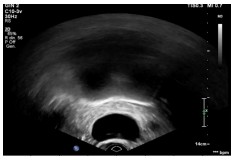
Error type	Original Image	Automatic Mask
Shadows segmentation		
Inhomogeneity errors		
Cut errors		
Inferior border errors		
General detection errors		

Table 7.1: Example of common segmentation errors made by OvAi Focus algorithm.

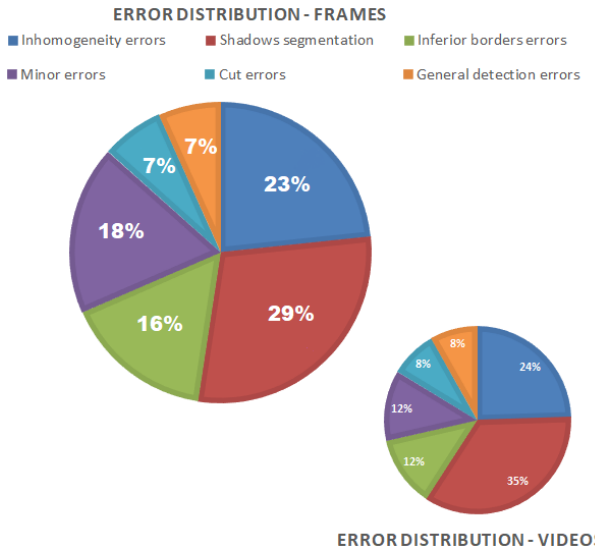


Figure 7.6: Errors distribution at frame and video level.

Manual labelling

The segmentation performance of the Focus algorithm for the task was therefore below expectations, preventing it from moving directly to model development. It was therefore required a step of **manual selection** aimed at the elimination of masks whose quality was not sufficient for the training of a neural network.

After the manual filtering operation, the 3312 labels (67,7% of total) needing a manual correction were isolated. Until that moment, Redbrick AI software had been used for the creation of masks starting from the image, so the possibility of uploading images with a pre-segmentation had not yet been considered. However, thanks to the *create_datapoint_from_masks* Redbrick method, it was possible to load into the software also masks derived from a previous segmentation pipeline, with the only care of transforming the binary mask into an RGB mask, replicating the information content of the single channel on the three channels.

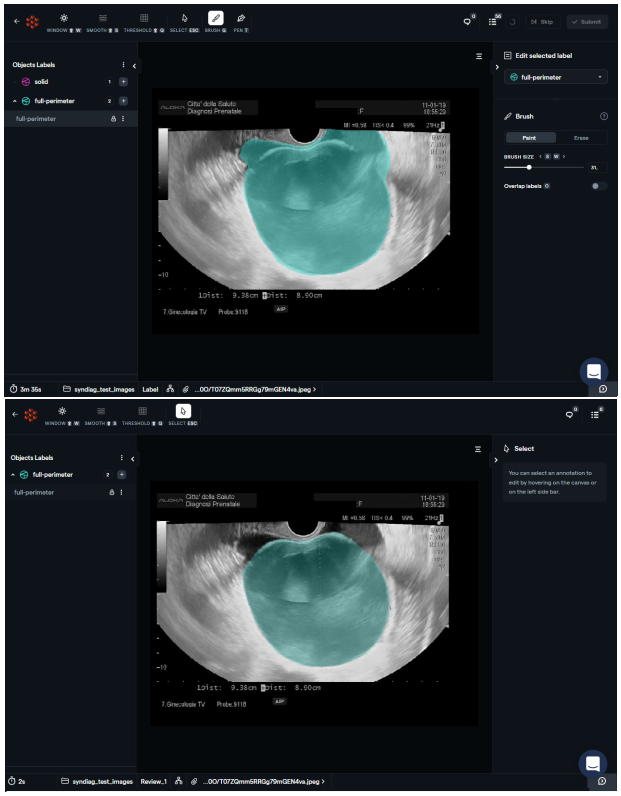


Figure 7.7: Redbrick AI software for labelling.

The activity of segmentation labelling for the thesis project was carried out by myself in an autonomous way. In the first months I spent in SynDiag I was trained on the labelling of cysts with solid and non-solid components: every week a meeting was held with a member of the team together with a gynaecology specialist, with the aim of learning the basics of this type of labelling. This training period allowed me to replicate the activity on a simpler task, such as the segmentation of serous cysts. Once the process of labelling serous cysts was completed, the resulting masks were downloaded from RedbcrickAI software, converted to single-channel grayscale images and uploaded to S3 together with the original right masks obtained from OvAi Focus.

7.4 Segmentation Pipeline

The whole proposed pipeline from the raw data to a fully-trained DL model is presented below.

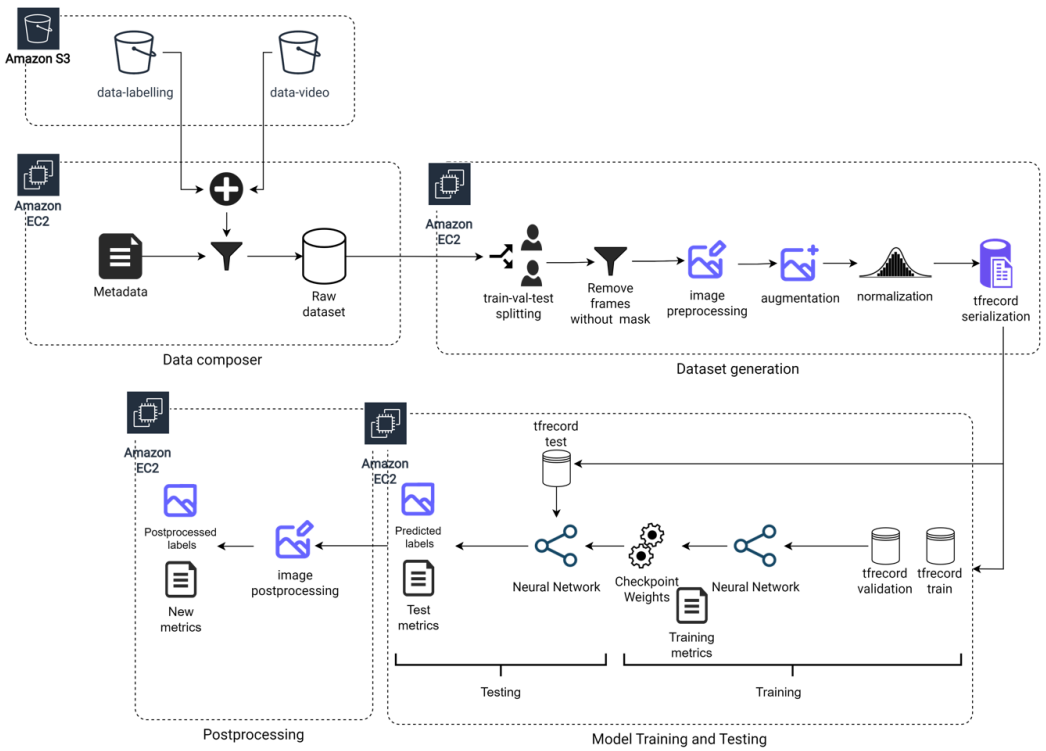


Figure 7.8: Redbrick AI software for labelling.

As shown in the figure, the first part of the pipeline is involved in data retrieval from the storage S3. The raw data are then assembled and pre-processed to build the datasets needed to train the network and assess its generalization performances. Only afterwards the NN training and testing take place. Finally, a post-processing step is added to refine the predictions computed by the network.

7.4.1 Dataset Composer

Ideally, the pipeline should be generic enough to be used for other tasks, similar to the one addressed by this study. For this purpose, the first part of the pipeline must allow the gathering of the data useful for a given experiment. In this case, it follows that the data that contributed to the final dataset were those discussed in **Section 7.3**. The class responsible for the data retrieval and their consequent organization is *RawDatasetComposer*, employed within *DataComposer* module. The input files are the metadata sheet containing the general information of the cases in OvAi platform and a JSON file, in which are specified the data to be filtered in light of the cases of interest. Filtering is performed before the actual data download from S3; this allows the minimization of the time associated with the information retrieval as well as the amount of data occupying the memory. *DataComposer* script also deals with the extraction of frames from the videos and the organization of all data locally, according to the hierarchical structure shown in Figure 7.9.

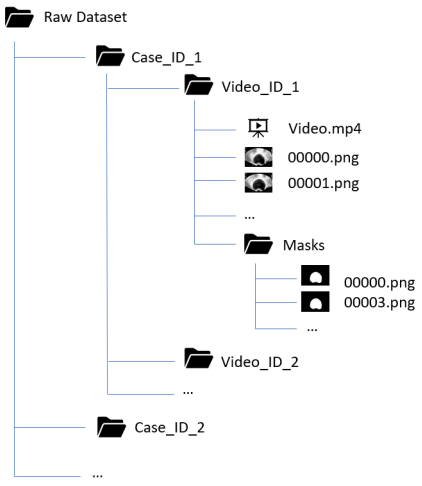


Figure 7.9: Local Data Organization.

7.4.2 Dataset Generation

Once the data collection and organization have been completed, the central phase of dataset generation can start. In order to minimize the computational time, the images and the related masks are loaded as array only at the pre-processing level; until that, images and masks paths are used as *data pointers* instead.

The operations performed on the raw data to generate a suitable dataset for DL models are the followings:

- **Train-Validation-Test Splitting**: as usually expected in the development of DL algorithms, the entire dataset is divided into training, validation and test sets. The split is made with respect to cases and not with respect to individual videos, to prevent information relative to the same clinical case from being distributed in different splits. The proportions followed for the division are 80%, 10%, 10% respectively for train,validation and test;
- **Filtering frames without masks**: the frames in the folders derive from the extraction operation performed at the previous level. For subsequent steps, only the frames for which a mask was actually generated must be carried forward. The output of this step is a tuple containing the path to the image and the path to the corresponding mask;
- **Pre-processing**: starting from this point the array of masks and images are actually loaded and a series of pre-processing operations, which will be detailed in the next section, is performed;
- **Date Augmentation**: Data Augmentation step helps in increasing data variability and can be selected through a proper flag. It will be covered in the next section;
- **Normalization**: normalization step is mandatory to correctly provide input images to the network. Image values are scaled between [-1,1], the range required by the employed encoding network;
- **Tfrecord Generation**: the pre-processed data is serialized according to a specific example and saved inside a proper folder. The advantage of choosing tfrecord format to save datasets is its efficient storage, which minimises occupied memory, and the ability to enclose different types of data together (images and masks arrays, as well as their paths). Uploading and downloading operations are also more rapid thanks to the lower memory weight. Each run of the DatasetGeneration script is associated with an alphanumeric code that uniquely identifies the dataset. Together with the output datasets, a CSV file

is updated to keep track of the datasets generated across the experiment: it matches each identification code with the set of parameters (related to the pre-processing steps and Data Augmentation) used to generate that dataset.

To ensure the validity of the algorithm, independently from the choice of a particular train/test/validation split of the original dataset, a cross-validation approach has been implemented. The dataset generation is thus repeated a number of times, defined by the *seed* parameter.

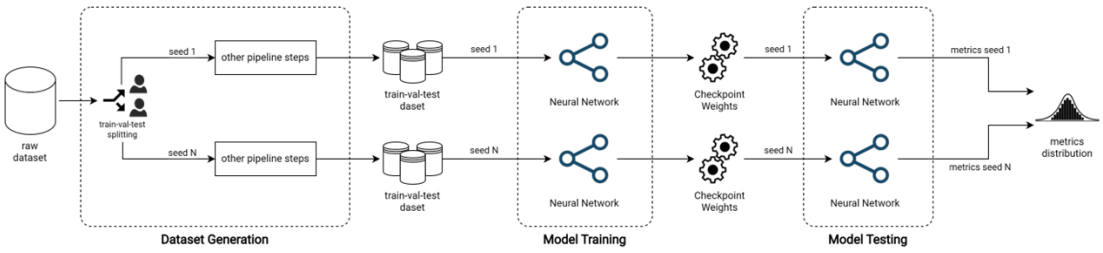


Figure 7.10: Cross validation pipeline.

Data preprocessing

The pre-processing step involves cropping and resizing operations. These two stages are necessary in order to remove useless information from the original image and bring it back to the input size expected from the pre-trained U-Net backbone. Ideally, a proper cropping operation would require the use of an **object detection** algorithm, able to automatically detect the area where the lesion is approximately located. Instead, in this approach, the crop is performed manually in two steps: the first makes use of a fixed-size window and is aimed at the exclusion, from the image, of the majority of the lateral annotations characteristic of the ultrasound image; in the second step, the image is made square by centered cropping of a size equal to the smaller side of the image. These steps aim to preserve as much as possible the cyst area within the image.
From the original size of 566 x 800, the cropping operation returns 498 x 498 images, and the final resize operation adjusts the height and width to the input size

chosen for the network, which in the present work has been set to 224 x 224. Depth remains altered in the case of both images (3 channels) and masks (1 channel). Since mask values are altered by the resizing operator, they undergo an extra step that consists of binarization (Otsu).

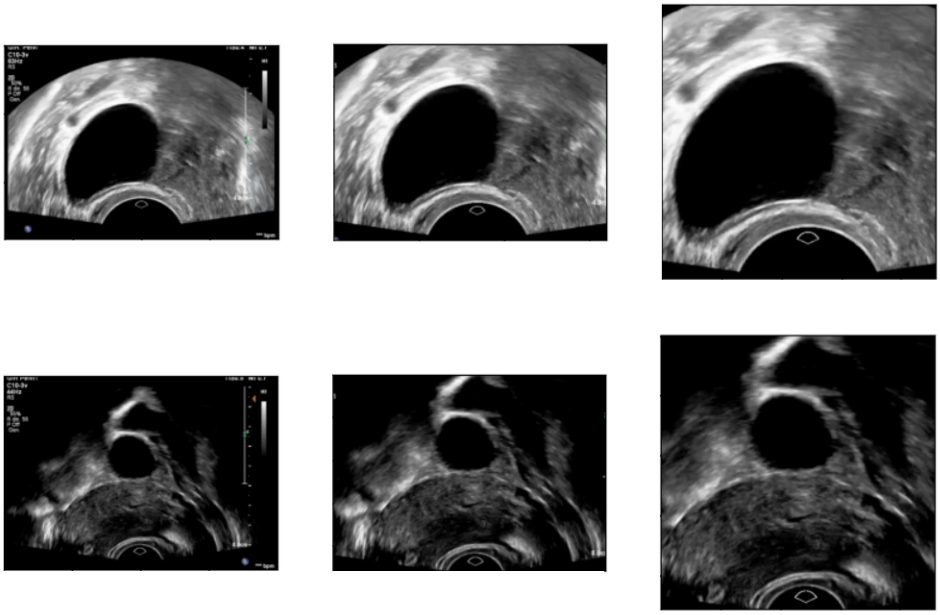


Figure 7.11: Result of preprocessing.

As stated in **Section 6.3**, since the objective of the present project is to compare a DL approach with Ovai-Focus and since Ovai-Focus does not include any noise-filtering step, no noisy-cleaning stage is provided within this pre-processing pipeline as well. In addition, it is believed that for the solely unilocular serous cyst identification, speckle noise filtering has a limited effect on the model performance. On the other hand, in the case that the same algorithm is used for the segmentation of cysts with solid components, the implementation of noise-filtering is believed to be mandatory.

Data Augmentation

The problem of little variability of the data could constitute one of the reasons for the low performance of a DL algorithm. The state-of-the-art study of J. Jin et al.[79] achieved high performances with a number of images 10 times lower but a number of cases definitely higher. **Data variability** is linked to a variety of factors, such as the morphological characteristics of the cyst (shape, position, content) but also the device used to carry out the radiological examination. Data

Augmentation is a widespread technique used in the preparation of medical datasets consisting of few data [80]. The goal is to apply transformations to images in order to generate new data; however, caution must always be placed for such operations, since the generated image must always be plausible and respectful of the true data statistics. Here, the implementation of Data Augmentation is a slight variant of the Deep Stacked Transformation illustrated by Ling Zhang et al. in [81]. The list of implanted plausible transformations is reported below.

Group	**Tranformation**	**Variability Source**
Appearance	Gamma Contrast	Gain and dynamic range knobs
	Brightness Shift	Gain and dynamic range knobs
Quality	Gaussian Blur	Focus knobs
	Sharpen	Focus knobs
	Spekle Noise	US equipment electromagnetic interferences
	Additive Noise	US equipment electromagnetic interferences
Spatial	Rescaling	Magnifying factor knob
	Rotation	Probe orientation
	Horizontal flip	Probe orientation
	Horizontal shift	Probe movements
	Vertical shift	Probe movements

Table 7.2: Data augmentation transformation and their variability source.

Each transformation is applied to the image according to the following law:

$$(\hat{x}_s, \hat{y}_s) = \tau^n_{m_n}(\tau^{n-1}_{m_{n-1}}(...\tau^1_{m_1}(x_s, y_s))) \tag{7.2}$$

where x_s and y_s are the training data and the associated label respectively, τ is the applied transformation and m is its magnitude.
As for the number of possible transformations applicable to each image and the corresponding intensities, both are defined randomly within a predefined range. In the present case, it was decided to limit the number of applicable transformations between 2 and 5, while the intensities are defined in the Appendix (A.2). Only the training dataset undergoes Data Augmentation, since variability improvement is needed during the learning phase. The transformations have been applied to both images and masks; however, some of the alterations (such as noise addition or saturation changes) did not affect the masks, as they do not involve any spatial transformation, e.g. translations or rotations. A proper sub-function of the augmenter is responsible for applying to the masks only the transformations necessary to maintain consistency with the altered image.
The augmentation produces a number of images depending on the chosen **ratio**

of increase K; considering an input tensor of dimension $B \times M \times N$, where B stands for the batch number and M and N for the image height and width, the output tensor dimension after augmentation will be:

$$B' = (B + K \cdot B) \times \ M \times \ N \tag{7.3}$$

In experiments involving Data Augmentation, K was set equal to 3. Some examples of the result of the transformations applied to the images are shown in Figure 7.3.

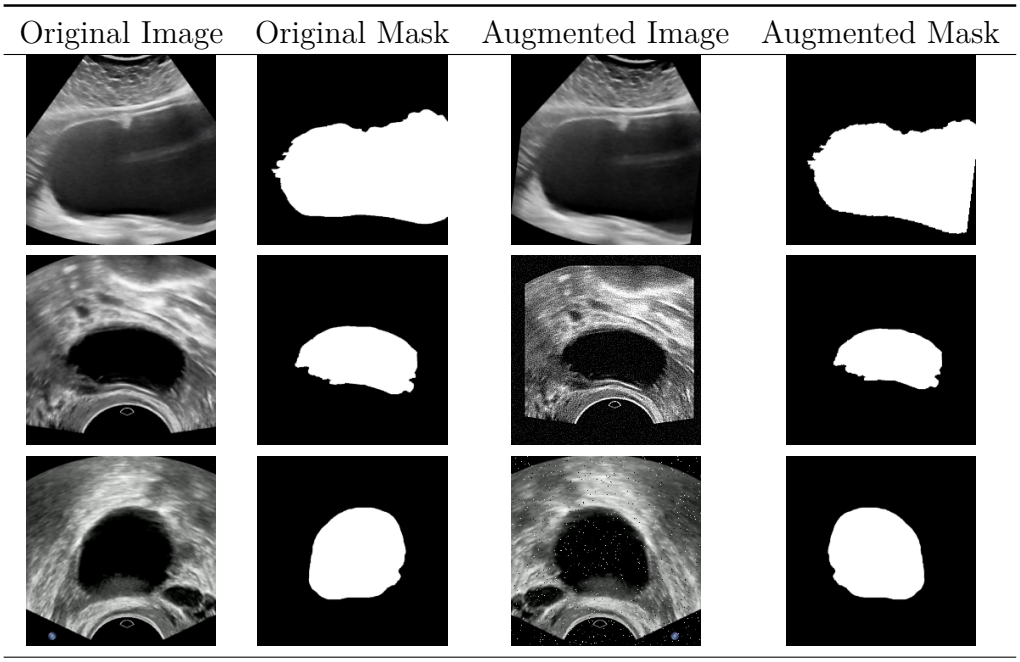


Table 7.3: Data Augmentation: result on random images and the corresponding binary masks.

7.4.3 Model Architecture

The model employed in this thesis project to tackle the segmentation of serous cysts is a variant of the U-Net architecture, proposed by Ronneberg et al. [53]. As previously stated, other few architectures have been tested for this same task (Section 5.1.2). However, since this was the first SynDiag attempt at segmentation of ovarian adnexal masses using a Neural Network approach, it has been decided to start with the simplest and reliable network architecture. Indeed, the selected learning model had already been successfully tested in the company when dealing with segmentation of liver in CT scans. The chosen network architecture therefore

represents the baseline on which more complex learning model will be designed in the next future.
The encoding path of this neural network consists of a **MobileNetV2**, while the decoding path is made up of up-sampling blocks composed of deconvolutional ($Conv2DTranspose$), batch normalization ($BatchNormalization$) and activation ($Relu$) layers. MobileNetV2 is a Convolutional Neural Network characterized by an inverted residual structure, whose residual connections are between the bottleneck layers. It outperforms MobileNetV1 in both latency and accuracy, and it also benefits from a lighter structure. Among the others CNNs implemented by Tensorflow, MobileNetV2 ranks second for the time needed for an inference step. This was essential for reducing both the time associated with simulation and their computational cost.

7.4.4 Model Training

Despite training a neural network from scratch on the target dataset would be the preferable choice, the dataset numerosity and variability available for this project are believed to be insufficient for efficiently train the network from random initial conditions. In order to speed up the simulation time, it is therefore decided to take advantage of pre-trained weights for the encoding network. This is a widely used technique in DL applications that is known under the name of $transfer\ learning$. The weights are transferred from a MobileNetV2, previously trained on the data-abundant **ImageNet** dataset. Given the Imagenet numerosity, this dataset is often chosen for transfer learning purposes even in the medical field, despite the poor similiraty between its natural images and the medical images in healthcare datasets. The hyper-parameter setting is reported in Table 7.4.

Hyper-parameter	**Value**
Loss function	Sparse Categorical Cross-Entropy
Optimizer and Learning rate	Adam [10^{-2}, 10^{-3}, 10^{-4}] SGD with CosineDecay [10^{-2} - 10^{-5}]
Batch size	64
Epochs	100
Shuffling	Yes
Seeds	10

Table 7.4: Hyper-parameters setting.

In order to optimize the computation time related to the data flow across the

pipeline, it was decided to make use of **dummy masks** instead of **one-hot encoded masks**. Therefore, each mask does not consists of a tensor where the number of channels corresponds to the number of classes (one binary mask for each class), but rather a unique channel contains information for multiple classes. That is, each pixel of the mask is assigned a numerical index related to the class to which it belongs (0 = background, 1 = cyst). This choice leads the usage of *Sparse Categorical Accuracy* as loss function. The decision also favours the future implementation of the model for multi-class segmentation, since the labels are always contained within one channel depth. Across the training epochs, a callback custom class takes care of the evaluation of the validation loss through a moving average computation. Monitoring the behaviour of the validation loss across the epochs is not only suitable for hyper-parameter tuning (e.g. batch-size, learning rate and many more) but also for implementing an early-stopping protocol. According to early-stopping, the network training should be stopped once the validation loss starts keeping on growing in a given time window (patience). The loss reference value computed at each iteration is given by the average of the last three epochs seen so far, while the number of patience epochs can be set within the configuration file (here equal to half of the total epochs). However, the *EarlyStopping* Tensorflow class does not allow to go on with training once the best epoch, corresponding to the smaller error on the validation set, is found. Since when training a network there are two important check-points, e.g. the one detected by early stopping and the one reached at convergence, a custom implementation of early-stopping was set-up in order to save the configuration of the network also at the last epoch.
Both **Stochastic Gradient Descendent** (SGD) and **Adam optimizer** (Adam) were tested for the model optimization. Adam optimizer is derived from the RMSProp and AdaGrad optimizers and is thus able to adapt the learning rate for each neural network weight based on the first and second moments of the gradient. SGD has been instead used with a variable learning rate, and in particular with a *CosineDecay* across the epochs. At the beginning of the training, the model must be able to move through the solution space with ease (thus with a high learning rate); as the training progresses, while approaching a global minimum, the "mobility" of the research must be restricted, in order to prevent the model from straying away from the solution. The decay steps are computed according to the batch size and the decay extends to all the training epochs. Both type of optimizer protocols are widely-used in neural network training, and because of that it was decided to test both of them.
The first trials of the algorithm have been run locally on a reduced-dimension dataset. As soon as the pipeline functionality was attested, the desired and complete simulation has been run to a remote AWS instance (EC2). Once the whole simulation was completed, the results were transferred back to the local machine to be displayed and analysed.

7.4.5 Post-processing

Simple post-processing steps were applied to the output image, with the aim of isolating, in case of multiple segmentations, the one corresponding to the cyst. The class that has been created for this purpose allows the composition of a specific post-processing pipeline. Currently, it only consists of the filling of the holes within the segmentations found within the output mask ($maxArea$ =250) and of the isolation of the biggest segmented area (cyst area). The first operation was implemented through the *remove_small_holes* function (skimage.morphology library); the second makes use of *findContours* and *contourArea* operators of the Open CV package. The post-processing phase is not aimed at modifying the output of the network but rather at its refinement.

7.5 Experiment Traceability and Observability

Each experiment performed within the proposed segmentation pipeline must be as modular and trackable as possible. To this end, the different operations within the segmentation pipeline take in and out a variety of information, which is exchanged between modules in different formats. The different data formats, together with their functionality, are listed below.

- **JSON**: JSON files are the key players in the information flow. They are used as **input files**, to set the configuration of a given operational block (for example the pre-processing and Data augmentation design of the hyper-parameters in the Dataset Generation module, or the ones associated to the learning model in the Training module) and as **outputs files**, to store the output metrics resulting from the model simulations. However, their content is not just limited to the settings of the individual experiment: since the code is constantly evolving and the network of users is expanding, JSON files also contain metadata related to the boundary conditions of the experiment, such as the simulation user, the versioning information of the environments, the Gitlab branch referenced by the codes on which the simulations are actually being run, and a series of others useful metadata.

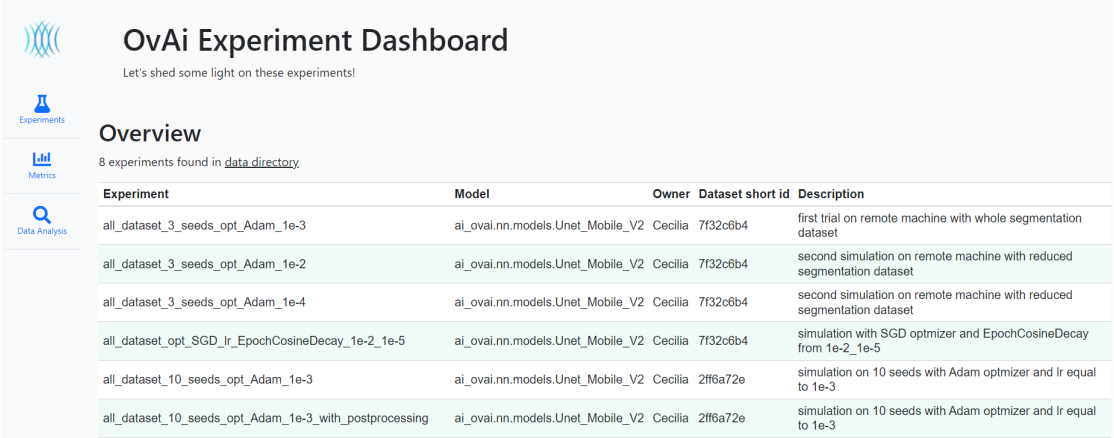
- **JSONL**: JSONL files store information printed within the editor's terminal during the execution of a single module. Logs are exploited to keep track of the outcome of single operations carried out by the executed block (whether they were successful, or whether problems that prevented their completion arose). The use of logs is particularly advantageous during simulations performed on

virtual EC2 instances, since no direct interface with the code is possible, and given the time required for a simulation to end, it would otherwise be hard to discover inconsistencies during the run of the simulation.

- **TFRECORD**: TFRECORD files are the data format used to store train/test and validation datasets, produced as a result by the Dataset Generation module.Tfrecords allow data of different format (strings, arrays...) to be saved through appropriate serialization; corresponding deserialization examples are also saved, allowing the datasets to be correctly unpacked when needed. Each dataset generation assigns a unique alphanumeric code to the produced tfrecord file types. Thus, for each run of the GenerateDataset script at least 3 tfrecord files (Training, Test, Validation) sharing the same id are generated. If cross-validation is also scheduled, the unique id is also followed by the seed index.
- **CSV**: CSV files are used both to store information inherent to ultrasound data (e.g. morphological labelling) and to keep track of the type of datasets created during the course of the experiments. In the latter case they associate the unique identification code of the tfrecord with the dataset configuration declared in the JSON file and used for its generation.
- **INDEX**: INDEX files are simply the weights generated during training for a simulation.

The combination of this information enables the complete reproducibility of an experiment. At the same time, being able to keep track of the results of the different experiments completed on a virtual EC2 instance, contextually with the information that led to their generation, allows for the avoidance of repeated experiments, thereby saving money and time.

In order to allow direct observability of the simulation results, JSON output files of the training and validation processes are then given as input to a custom-made OvAi Experiment Dashboard (Figure 7.12) for performance analysis. This interface was built in such a way as to expect a precise data format for the automatic plotting of the simulation output metrics.
Thanks to the completeness of the data enclosed in the JSON files, it is possible to provide each experiment with name and description in the dashboard, thus enabling the ease of its tracking within the interface. Two main plots are automatically computed: one displays the trend of the metrics of interest during the training, e.g. train and validation loss and accuracy, the second shows the final performance on the test set. The former represents the metrics of interest as a function of epoch trends. If multiple seeds are identified within the folder containing the results, the resulting plot will be the average performance obtained over the different seeds, with



Experiment	Model	Owner	Dataset short id	Description
all_dataset_3_seeds_opt_Adam_1e-3	ai_ovai.nn.models.Unet_Mobile_V2	Cecilia	7f32c6b4	first trial on remote machine with whole segmentation dataset
all_dataset_3_seeds_opt_Adam_1e-2	ai_ovai.nn.models.Unet_Mobile_V2	Cecilia	7f32c6b4	second simulation on remote machine with reduced segmentation dataset
all_dataset_3_seeds_opt_Adam_1e-4	ai_ovai.nn.models.Unet_Mobile_V2	Cecilia	7f32c6b4	second simulation on remote machine with reduced segmentation dataset
all_dataset_opt_SGD_lr_EpochCosineDecay_1e-2_1e-5	ai_ovai.nn.models.Unet_Mobile_V2	Cecilia	7f32c6b4	simulation with SGD optmizer and EpochCosineDecay from 1e-2_1e-5
all_dataset_10_seeds_opt_Adam_1e-3	ai_ovai.nn.models.Unet_Mobile_V2	Cecilia	2ff6a72e	simulation on 10 seeds with Adam optmizer and lr equal to 1e-3
all_dataset_10_seeds_opt_Adam_1e-3_with_postprocessing	ai_ovai.nn.models.Unet_Mobile_V2	Cecilia	2ff6a72e	simulation on 10 seeds with Adam optmizer and lr equal to 1e-3

Figure 7.12: OvAi Experiment dashboard.

the possibility of also showing maximum and minimum values and the standard deviation for each epoch. The latter illustrates test performances via a boxplot chart and therefore displays the median, 25th and 75th percentiles for each metric, along with the performance of individual seeds.

Chapter 8

Results and Discussion

8.1 Design of Experiments

The final dataset, cleaned of cases not conforming to the chosen morphology consisted of **26 cases**, for a total of **50 videos** (Section 7.3.1).
OvAi Focus was employed to extract **4887 masks** from the source videos, taking one frame every three within the time window containing the cyst. After the visual inspection of the labels, **3312** of them were chosen to be manually corrected on **Redbrick AI** software.
The wrong masks have therefore been replaced with the fixed ones, and together with the unmodified masks produced by OvAi Focus, composed the final dataset. As soon as the dataset was ready, the first simulations started.

The goal of the first simulations was to find the best method of **optimization** and an adequate **learning rate**. For this reason, these first experiments were performed on a **limited number of seeds** (3 seeds). Once the setting of these hyper-parameters was established, the focus was directed at the possible improvement of the overall performance of the algorithm; for this purpose, taking into account the low variability of the original dataset, **Data Augmentation** was introduced. The **post-processing** step was then added at the end of the pipeline, as its first goal was not so much to bring a concrete increase in performance, but to visually improve the predictive output of the Neural Network.

Standard accuracy metrics for segmentation were used to assess the performance of the algorithm, i.e. **Intersection Over Union** (IOU) and **Dice Score Coefficient** (DSC) as general accuracy metrics, together with **Precision** and **Recall** as indices of over- and under- segmentation (defined in Section 4.4.2).

8.1.1 Overall Performances

The results of the first simulations immediately demonstrate the great ability of the model in tackling the segmentation of serous cyst type, as the averaged metrics stay above 90% (Figure 8.1)

Figure 8.1: Averaged performance comparison with different optimizer configurations.

Optimizer	DSC	IOU	Precision	Recall
Adam 10^{-3}	**95.9%±0.1%**	**92.3%±0.3%**	**95.5%±0.5%**	**96.4%±0.5%**
Adam 10^{-2}	95.7%±0.2%	92.0%±0.3%	95.4%±0.4%	96.1%±0.6%
Adam 10^{-4}	95.3%±0.5%	91.3%±0.8%	96.3%±0.5%	94.5%±1.4%
SGD 10^{-2}-10^{-5}	95.5%±0.2%	91.7%±0.4%	94.9%±0.4%	96.4%±0.2%

Table 8.1: Averaged performance comparison with different optimizer configurations.

As the intent is to observe the performance solely based on the selected optimizer configuration, the comparison is made using the *last _weights*, e.g. the weights of the model at convergence. The boxplot (Figure 8.1) shows that the performance of the two optimizer protocols taken into account is generally comparable with each other. Only the use of a learning rate equal to $1 \cdot 10^{-4}$ for the Adam optimizer seemed to negatively affect the variability between the seeds and the average performance in the accuracy metrics (IOU and DSC). From the Recall and Precision

metrics, it is clear that this is primarily due to a moderate under-segmentation error (low Recall but good Precision). The training loss and the learning rate trends across the epochs are reported in **Figure 8.2a** and **Figure 8.2b** respectively.

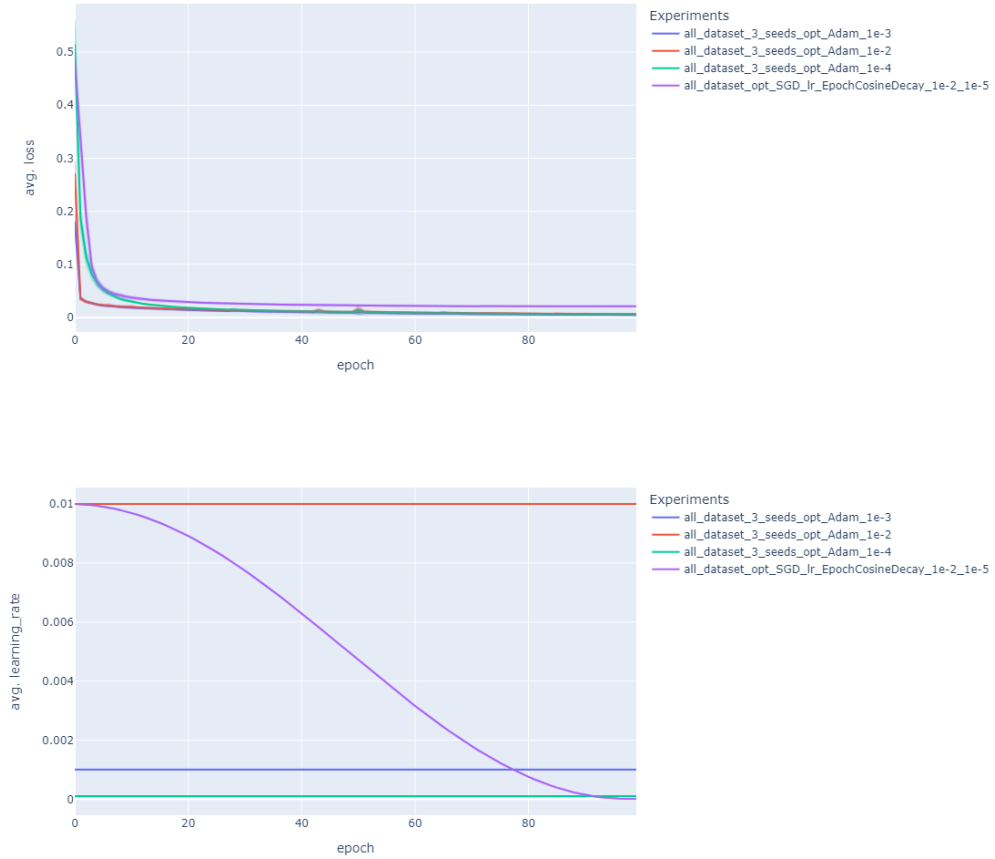


Figure 8.2: Average IOU (up) and Loss (down) across training epochs for training set.

As expected, the models trained with Adam and with the highest learning rate converge fastest in the learning phase. Instead, the SGD optimizer associated to the cosine learning rate scheduler seems to get stuck at higher values of the training loss.
From **Table 8.1**, it can be noticed that the combination of Adam optimizer with $1 \cdot 10^{-3}$ as LR turns out to be the optimal choice, having the highest average performances together with the smallest variability for IOU and DSC metrics

computed on the test set.

The winning optimizer configuration was then chosen for the simulations with 10 seeds. The increase in the number of seeds is meant to perform a more reliable application of $cross - validation$ and thus a better convergence towards the typical case. This technique allowed to truly evaluate the variability between seeds according to the different possible combinations of the dataset, and therefore to test the robustness of the model. Results look promising, as can be seen looking at the violet box plot in **Figure 8.3**. Indeed, the variability between seeds remained very low even if the distribution of clinical cases varied within the train/test and validation sets. Moreover, as shown in **Table 8.3**, the use of transformations on images introduced with Data Augmentation further improved cyst segmentation capabilities, leading to an increase of about 1% in IOU (red box plot). A final post-processing step was then added to the pipeline including data augmentation. As mentioned above, the post-processing phase was primarily aimed at visual improvement of the masks and only secondarily at actual performance improvement. However, it can be noticed that the introduction of cleaning operations contributed to a slight increase in performance too (green box plot). In particular, removal of minor segmentation area resulted in a small raise in Precision, indicative of a diminished over-segmentation error; the Recall, on the other hand, saw no appreciable improvement after the post-processing stage.

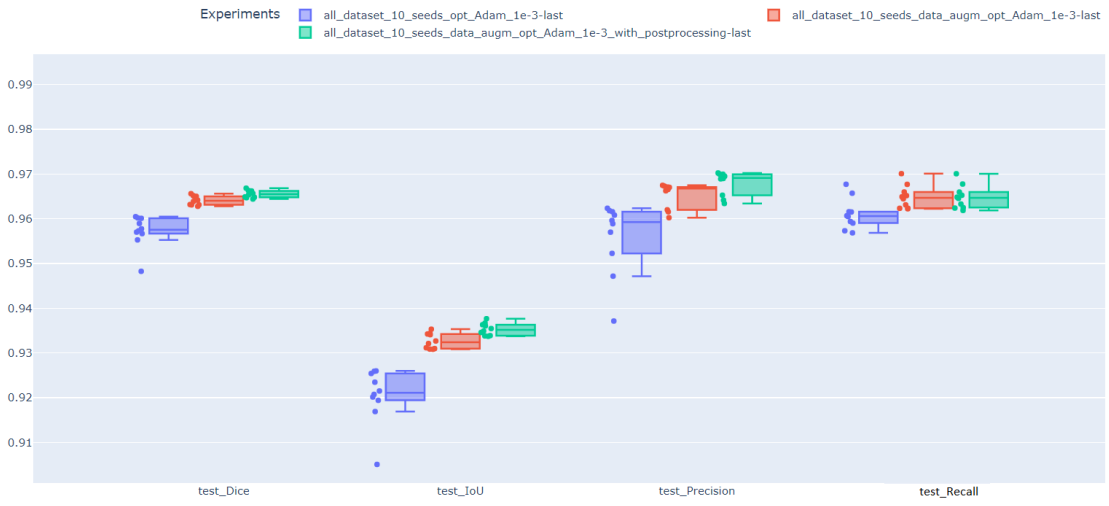


Figure 8.3: Performance comparison after the introduction of Data Augmentation (DA) and postprocessing.

Configuration	DSC	IOU	Precision	Recall
Baseline	95.7%±0.03%	92.0%±0.1%	95.9%±0.2%	95.8%±0.1%
DA	96.3%±0.04%	93.1%±0.1 %	96.7%±0.1%	96.2%±0.01%
DA & postpr	**96.5%±0.03%**	**93.4%±0.1%**	**97.0%±0.1%**	**96.2%±0.03%**

Table 8.2: Performance comparison after the introduction of Data Augmentation (DA) and postprocessing.

In addition to the above comparison, it is possible to observe the result relative to the best epoch detected by the custom early-stopping callback class.

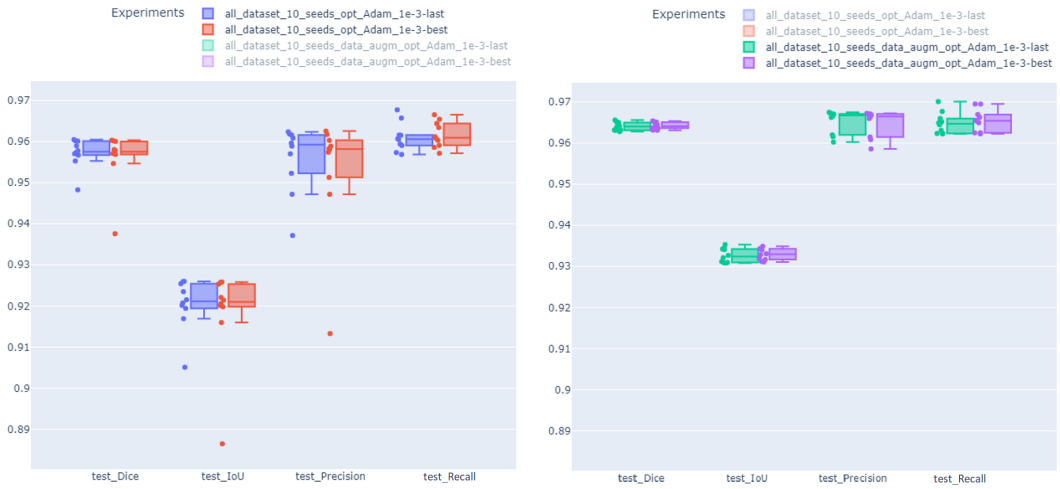


Figure 8.4: Last epoch and best epoch performance comparison: without DA (left), with DA (right).

The performances of the best epochs are computed as the average of the metrics corresponding to the minimum validation loss recorded for each seed. Consequently, they are the result of the average over different best epochs, i.e. one for each seed. It can be noticed that independently from the execution of data augmentation within the pipeline (the post-processing step does not influence the NN weights), the metrics recorded using last epoch weights and best epoch weights were almost comparable. As a result, the performance of the model was already satisfactory even before reaching the end of the training; this information could be used to slightly decrease the number of total training epochs, in accordance with the average number of epochs needed to achieve the best performances. In the pipeline achieving the highest performance, e.g. the one including data augmentation and post-processing, the metrics at the best and the last epochs can therefore be considered approximately equivalent.

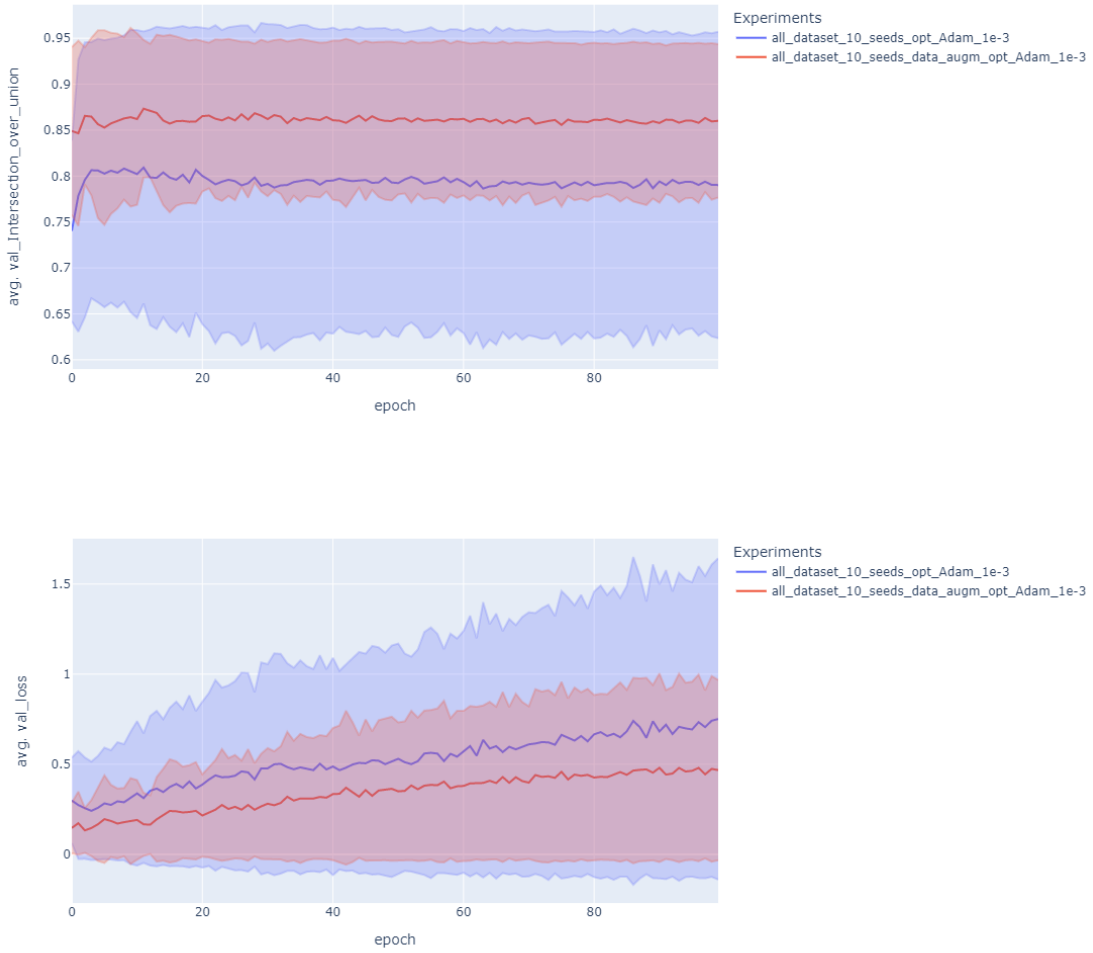


Figure 8.5: Average curves and standard deviation intervals for IOU (up) and Loss(down) across training epochs for validation set computed as mean ± std. deviation.

The loss and IOU trends calculated for the validation set over the training epochs confirmed the model's improvement. The introduction of data augmentation decreased the loss by a considerable factor and also reduced the seed variability with respect to the original dataset. Similarly occurred for the Intersection Over Union, where its averaged value increased more or less of a delta of 11% at the end of the training with respect to the non-augmented case.

8.2 Model comparison : U-Net-MobileNetV2 vs OvAi Focus

The comparison between the performances of the two automatic algorithms for segmentation was compulsory. For a fair comparison, however, the mask sizes generated by OvAI Focus and MobileNetV2 * had to be consistent. Initially, it was considered to restore the output dimension of the FCNN algorithm to the original image size. A simple resize, however, would not have been sufficient to overcome the issue, because of the cropping operation applied in the pre-processing step. Moreover, even attempting to restore the dimension through a combination of both resize and padding would have been equally unfair, since the information added by padding does not come from the Neural Network algorithm, and it would inevitably affect the metrics. For all these reasons, the most practical solution was to pre-process the images produced by OvAI Focus according to the same pre-processing in the DL segmentation pipeline.

In addition, not all the images could be included in the comparison. A portion of the images composing the final dataset directly derive from the OvAi Focus segmentation module (the masks that at Section 7.3.2 were considered correct enough to act as ground truth). Thus, these masks were excluded, as the performance metrics evaluated on them would have inevitably resulted in 100% accuracy.

To summarize, the steps necessary for the comparison are: of the total masks produced by the model under the best conditions (e.g. best optimizer with data augmentation and post-processing step) on 10 seeds, masks with OvAi Focus mask as reference ground truth were excluded. The remaining were pre-processed and then grouped according to the main errors highlighted in Section 7.3.2. Minor errors were not considered for the purpose of this analysis.

Regarding the overall performance trend evaluated by considering the average over the 10 seeds (Table 8.3), a substantial improvement is immediately observable. Not only the segmentation accuracy metrics registered a significant increase (almost 10% for both IOU and DSC), but also the Recall and Precision values suggest that the proposed segmentation pipeline generally commits fewer over- and under-segmentation errors with respect to the original algorithm.

*From now on we will refer to the U-Net-MobileNetV2 network simply as MobileNetV2, and any following comparison with the state-of-the-art will be made on the basis of the best model configuration (with Data Augmentation and postprocessing).

Metric	OvAI Focus	MobileNetV2
DSC	86.9%±8.1%	96.1%±0.8%
IOU	80.8%±8.0%	92.8%±1.1%
Precision	89.0%±7.5%	97.0%±0.8%
Recall	86.5%±8.6%	95.3%±1.0%

Table 8.3: Averaged model performance.

Table 8.4 shows the same metrics of Table 8.3 grouped by the main error types committed by Focus and highlighted in **Section 7.3.2**.
As can be seen, the DL approach described in the present project generally increased the overall performance in all identified error categories, though for some errors we observe a greater delta with respect to OvAi-Focus performances than others. Delta values exceeding 10 % are highlighted in the table.
As can be deduced from this analysis, first, the Deep model had a better ability to recognize the cyst (**general detection errors**), aligning the values of the metrics for these images with the others. The second error class lessened by the new model is the segmentation of **acoustic shadows**, which was significantly reduced. For this error, Precision (over-segmentation index) rose by 16.6%, leading to an increase in the corresponding Intersection Over Union as well. The algorithm also performed better in identifying the correct cyst edges, decreasing the **cut errors** and those related to the **inhomogeneity** of the cyst itself. Unlike the previous case, this time the metric showing an improvement is Recall, since these errors typically consist of segmentation defects within the ROI. The result of the improvement in the segmentation task can be observed in Figure 8.5.

It is still to be questioned whether the proposed solution always outperforms Focus. In order to make this assessment, masks with lower performance than those of the original algorithm have been selected. Within the 10 possible test sets available for the 10 seeds, an average of 11.7% masks per seed registered higher IOU and DSC for the OvAi-Focus algorithm. However, this is a very rough estimation, since the number of images per seed is highly variable. For the purposes of this consideration only, all the masks that make up the 10 seeds test set are considered as a single set of different elements. Despite the possibility that some masks are repeated across seeds, it remains true that these were generated under different training conditions and are thus characterized by different performances. Of the total number of masks composing the 10 seeds, amounting to 4212, 415 were found to have higher IOU and DSC for OvAi Focus algorithm (around 10%). In most of these cases, the performance of OvAi Focus very slightly exceeds that of the proposed algorithm (for the 89% of the predicted masks outperformed by OvAi

Shadows Segmentation			
	OvAi Focus	MobileNetV2	Delta
DSC	85.9%	95.3%	+ 9.4%
IOU	76.7%	91.2%	+ **14.5%**
Precision	80.1%	96.7%	+ **16.6%**
Recall	94.8%	94.2%	- 0.6%
Inhomogeneity errors			
	OvAi Focus	MobileNetV2	Delta
DSC	92.3%	96.6%	+ 4.3%
IOU	86.1%	93.6%	+ 7.5%
Precision	97.6%	97.6%	-
Recall	88.1%	95.8%	+ 7.7%
Inferior border errors			
	OvAi Focus	MobileNetV2	Delta
DSC	94.3%	97.2%	+ 2.9%
IOU	89.4%	94.7%	+ 5.3%
Precision	94.3%	97.9	+3.6%
Recall	94.6%	96.6%	+ 2.0%
Cut errors			
	Ovai Focus	MobileNetV2	Delta
DSC	87.6%	97.0%	+ 9.4%
IOU	80.3%	94.4%	+ **14.1%**
Precision	98.7%	98.3%	- 0.4%
Recall	81.2%	96.0%	+ **14.8%**
General detection errors			
	OvAi Focus	MobileNetV2	Delta
DSC	4.7%	93.8%	+ **89.1%**
IOU	2.8%	89.7%	+ **86.9%**
Precision	12.7%	95,4%	+ **82.7%**
Recall	3.2%	92.6%	+ **89.4%**

Table 8.4: Performance improvements related to error class.

Focus algorithm the difference is less than 5%), thus remaining comparable.
What is of interest to identify, however, are the images for which the DL algorithm completely failed to identify the cyst, i.e., IOU=0. Among the masks belonging to the 10 seeds, this same condition occurred only in one clinical case, shown in Figure 8.6. For this same images, also OvAi Focus recorded a null IOU.
In this circumstance, the network correctly succeeded in detecting the cyst, but due to the post-processing operation, aimed at the isolation of the largest connected

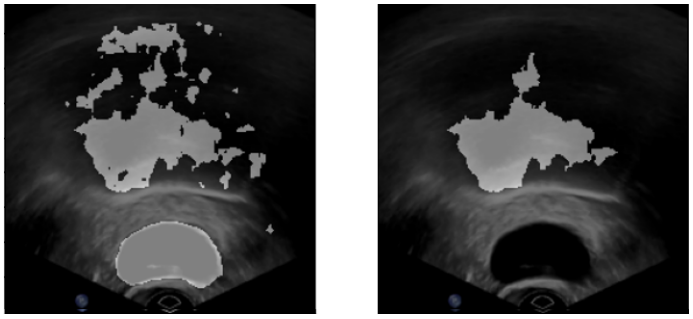


Figure 8.6: Example of the main adversarial effect of postprocessing on the segmentation label.

component, the shadow was finally returned instead. This was therefore one of the major limitations of the proposed segmentation pipeline.
This error, although very limited within the current dataset (only 4 images, all belonging to the same clinical case), could however represent a considerable problem for future usage of the model. In this regard, a possible solution would be the implementation of a function in the post-processing module, designed to identify the roughness of the connected components. This roughness metrics could be based on the computation of the average distance, or the area ratio, between a polygonal contour approximation of the segmentation and the segmentation itself. However, this implementation and its corresponding tests are left for future work.

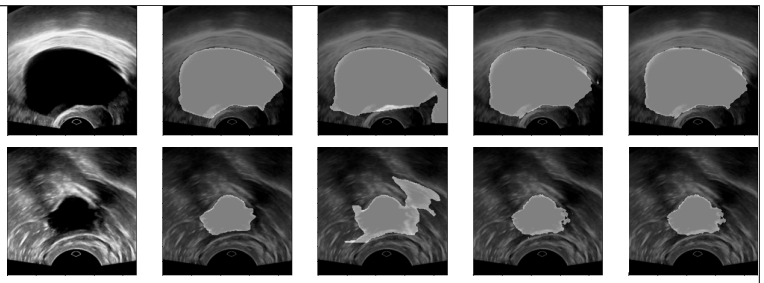
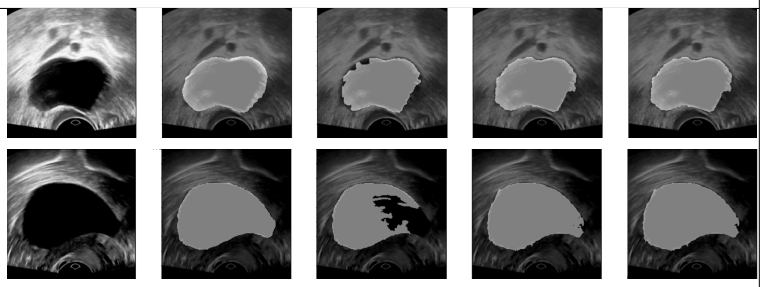
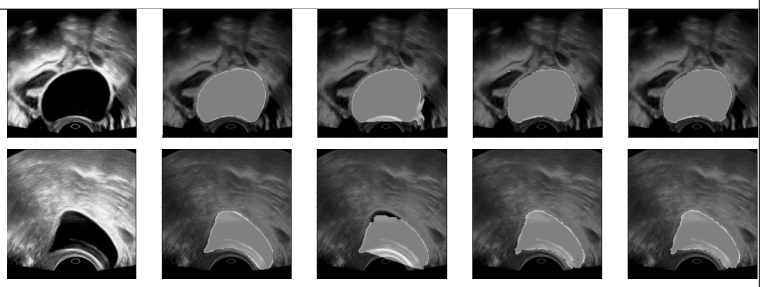
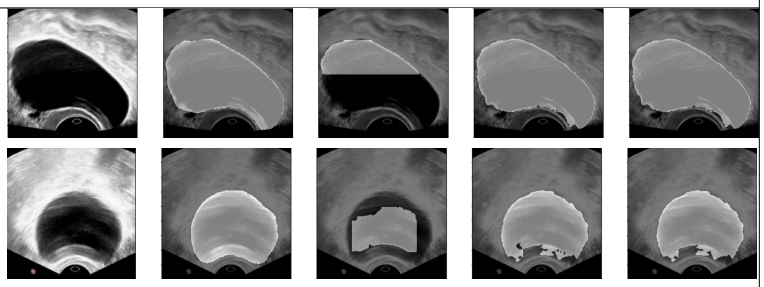
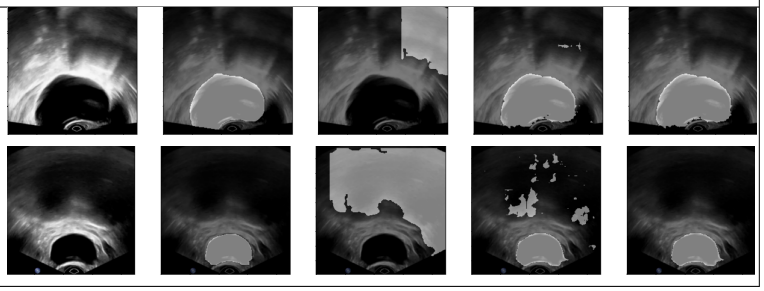
Shadows Segmentation	
Inhomogeneity errors	
Inferior border errors	
Cut errors	
General detection errors	

Table 8.5: Visual comparison of the original images (first column), the ground truth label (second column), the output of OvAi Focus Algorithm (third column), the output of the MobileNetV2 without postprocessing (fourth column), and after postprocessing (fifth column).

Chapter 9

Conclusion

The purpose of this thesis project was the realization of a dedicated pipeline for the segmentation of serous cysts from B-mode images, making use of a Deep Learning algorithm. Despite a reference standard already existing for this purpose in SynDiag, i.e. the OvAi-Focus segmentation module, the DL approach has been designed to try to overcome some of the main limitations of Ovai-Focus segmentation.
The development of the proposed system required the integration of data preparation and selection steps along with the programming of a fluid code interconnected with different data sources (local computers, AWS, OvAi platform). Traceability of the data characterizing the algorithm allows ease of reproduction of specific experiments, while the OvAi Experiment Dashboard takes care of their corresponding visualization. Coding was performed with the aim of minimizing computation costs and timing of data retrieval. Regarding the Deep model exploited for segmentation and its configuration, the use of a U-Net with MobileNetV2 as encoding network was found to perform well for the purpose. In addition, the use of transfer learning from ImageNet resulted still valid in the context of serous cyst segmentation. All of the tested optimizers generally proved capable of accomplishing the required task; the introduction of Data Augmentation as a palliative for data-scarcity led to a moderate increase in performance, followed by a slight increase brought by a preliminary post-processing step. The proposed new alternative pipeline surpasses the state-of-art model, by reducing the issues connected to missed cyst detection, acoustic shadows, and edge detection. The algorithm is therefore a likely candidate to be introduced within OvAi products, as well as to be used by other projects currently under development in SynDiag, that need to properly isolate the edge of serous cysts.

9.1 Future Improvements and Applications

It is possible to list several future improvements and applications that were not directly addressed in this study due to time constraints or that are going to be investigated in the next future.

- The first planned step for the proposed segmentation pipeline consists in its **validation**. To ensure that the algorithm works properly and it is not overfitting on the custom dataset, a re-training phase on the entire dataset would be needed. The resulting model would then be tested on completely new test cases, coming from different hospitals and ultrasound providers. Two test datasets are currently being prepared for this purpose. One contains unseen unilocular serous images, while the other consists also of cysts with solid components. The first serves to evaluate the performance of the proposed algorithm on an equivalent task but with a different set of test data; the second is meant to evaluate the generalization capabilities of the model for more complex mass morphologies.

- The segmentation pipeline could be improved in several aspects. The lack of adequate filtering operations was here intentional in order to achieve a fair comparison with the Ovai-Focus algorithm. However, for this type of image, an adequate filtering step would be necessary. **Speckle noise filtering** as well as **equalization** steps should be introduced within the pre-processing phase, since ultrasound images can not only be noisy but also very different in contrast and brightness, due to display options available in the ultrasound scanner. In addition, the cropping and resize operations, although succeeding in broadly isolating the location of the adnexal mass correctly, could also be improved. For instance, they could be replaced by a customized **object detection** algorithm aimed at finding automatically the bounding box around the cyst, preventing the loss of information caused by the fixed cropping. The post-processing step should also be revised, to avoid shadow segmentation in images containing small area cysts. To overcome this issue, a function estimating **roughness** of the edges could be advantageous.

- The Neural Network itself could be further optimized with a proper **hyper-parameter tuning**. In this case study, only two optimizers have been tested to evaluate model performances. However, other combinations could still be investigated. Moreover, the model currently employs transfer learning with the ImageNet database. The employment of pre-trained weights of learning models trained on medical datasets, preferably composed of ultrasound images,

could constitute another possible advance. Improvement in this sense would probably be major appreciated in more difficult segmentation tasks.

- **Data flow** across the whole pipeline is still to be optimized. The use of a proper Database software instead of a metadata sheet to store useful data related to clinical cases could greatly improve data filtering and retrieval steps, speeding up the whole data collection pipeline.

A future application of the proposed Deep approach proposed in this work would be its extension to **multi-object segmentation**. The use of dummy masks instead of one-hot encoded ones makes the current implementation easy to adjust to this more complex case and inexpensive in terms of required memory and code adaptation. In this regard, the main differences with respect to the experiments performed in the thesis project would be: the evaluation of the model performances, which would require a multi-class IOU and DSC instead of their binary counterpart, and the intrinsic difficulty of a multi-object segmentation, which would probably require much more efforts in the proper hyper-parameter fine-tuning.

Appendix A

Appendix

A.1 Dataset preprocessing options

Parameter	Value	Resultant dimensionality
Original dimensionality	-	566 x 800
Cropping window	[56, -11, 45, -75]	499 x 680
Squaring and Centering	-	498 x 498
Resize	[224, 224]	224 x 224

Table A.1: Cropping and resizing options.

Tranformation	Min	Max
Gamma Contrast	0.75	1.5
Shift	-20	20
Gaussian Blur	0.5	2.5
Sharpen	0.1	0.25
Speckle Noise	0.01	0.05
Additive Noise	8	15
Rescaling	0.85	1.25
Rotation	-10	10
Horizontal flip	1.0	1.0
Horizontal shift	-15	15
Vertical shift	-15	15
Number of transformations per image	2	5
Augmentation rate K	3	3

Table A.2: Data augmentation transformation and their range intensities.

A.2 Model Architecture

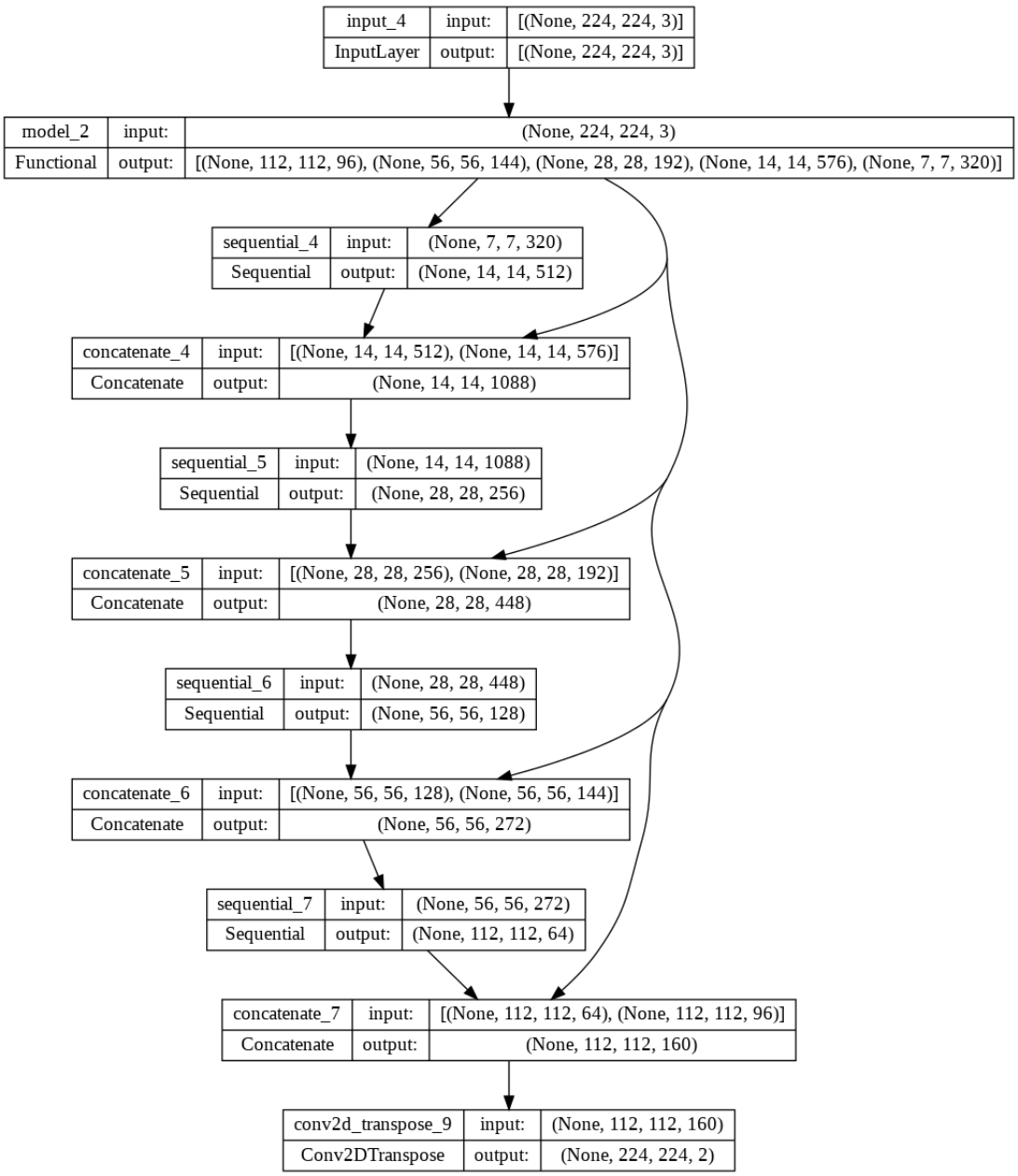


Figure A.1: MobileNetV2-UNet Architecture.

A.3 Results

	OvAI Focus				MobileNetV2			
	Dice	IoU	Precision	Recall	Dice	IoU	Precision	Recall
seed 0	64.0%	58.5%	68.7%	62.8%	96.6%	93.7%	97.2%	96.2%
seed 1	90.9%	85.4%	91.8%	91.5%	96.2%	92.8%	96.8%	95.7%
seed 2	91.8%	86.0%	89.9%	94.7%	96.5%	93.3%	97.6%	95.5%
seed 3	90.0%	83.7%	92.4%	89.3%	96.4%	93.1%	97.7%	95.3%
seed 4	92.7%	86.7%	94.5%	91.8%	96.6%	93.6%	97.1%	96.2%
seed 5	90.4%	83.8%	92.1%	90.3%	97.3%	94.9%	97.8%	97.0%
seed 6	86.7%	78.4%	94.8%	82.8%	95.4%	91.3%	98.1%	93.0%
seed 7	87.2%	81.6%	86.7%	89.3%	95.2%	91.0%	95.9%	94.6%
seed 8	91.6%	85.8%	94.7%	89.6%	95.8%	92.1%	96.6%	95.3%
seed 9	83.3%	78.5%	84.1%	83.5%	94.7%	92.2%	95.4%	94.2%

Table A.3: Comparison of per-seed performance of OvAi Focus with MobileNetv2*(*AGandpostprocessingincluded*)*limitedtomasksmissedbyOvAiFocus*.

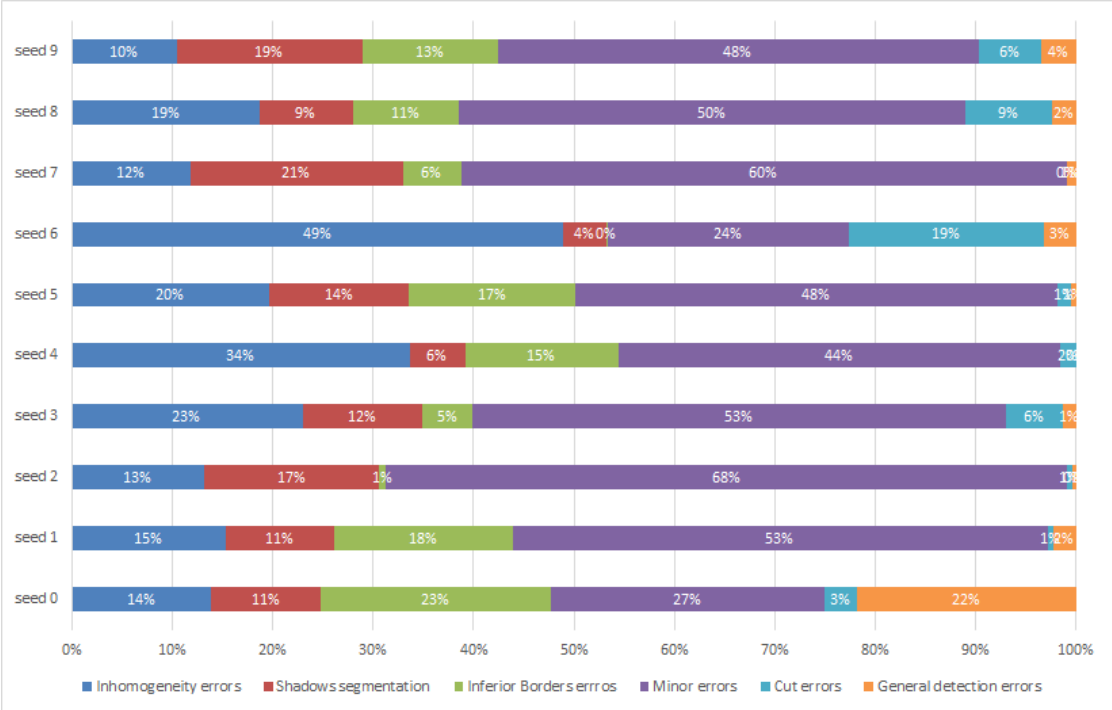


Figure A.2: Error distribution within each seed.

Bibliography

[1] World Cancer Research Fund International. «Ovarian Cancer Statistics». In: (2020) (cit. on p. 1).

[2] Associazione Italiana di Oncologia Medica (AIOM). «I numeri del cancro in Italia». In: (2020) (cit. on pp. 1, 4, 10).

[3] S. Chiara Cecere et al. «BRCA1 and BRCA2 in ovarian cancer: ESMO biomarker factsheet». In: (2016) (cit. on p. 2).

[4] C. Stewart et al. «Ovarian Cancer: An Integrated Review». In: *Seminars in Oncology Nursing* 35.2 (2019), pp. 151–156 (cit. on p. 2).

[5] C. Bulletti et al. «Aspetti morfo-funzionali dell'ovaio». In: *Caleidoscopio : rivista monografica di medicina* 8 (1984) (cit. on p. 3).

[6] C. Mimoun et al. «Masse ovariche: tumori benigni e maligni». In: *EMC - AKOS - Trattato di Medicina* 18.2 (2016), pp. 1–7 (cit. on p. 4).

[7] Associazione Italiana di Oncologia Medica (AIOM). «Linee guida CARCINOMA DELL'OVAIO». In: (2021) (cit. on p. 4).

[8] M. Horta et al. «Sex cord-stromal tumors of the ovary: a comprehensive review and update for radiologists». In: *Turkish Society of Radiology* 21 (4 July 2015), pp. 277–86 (cit. on p. 4).

[9] Abd Alkhalik Basha M. et al. «Comparison of O-RADS, GI-RADS, and IOTA simple rules regarding malignancy rate, validity, and reliability for diagnosis of adnexal masses». In: *European Radiology* 31 (2020), pp. 674–684 (cit. on p. 6).

[10] Lai H. et al. «Comparison of O-RADS, GI-RADS, and ADNEX for Diagnosis of Adnexal Masses: An External Validation Study Conducted by Junior Sonologists». In: *Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine* 41.6 (2022), pp. 1497–1507 (cit. on p. 6).

[11] D. Timmerman et al. «Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) group». In: *Ultrasound Obstet Gynecol* 16 (5 2000), pp. 500–505 (cit. on p. 6).

[12] D. Timmerman et al. «Simple ultrasound-based rules for the diagnosis of ovarian cancer». In: *Ultrasound Obstet Gynecol* 31 (6 2008), pp. 681–690 (cit. on p. 6).

[13] WT. Xie et al. «Efficacy of IOTA simple rules, O-RADS, and CA125 to distinguish benign and malignant adnexal masses». In: *Journal of Ovarian Research* 15 (1 2022) (cit. on p. 7).

[14] S.n Szubert et al. «External validation of the IOTA ADNEX model performed by two independent gynecologic centers». In: *Gynecologic Oncology* 142.3 (2016), pp. 490–495 (cit. on p. 7).

[15] D. Tinnangwattana et al. «IOTA Simple Rules in Differentiating between Benign and Malignant Adnexal Masses by Non-expert Examiners». In: *Asian Pacific journal of cancer prevention* 16.9 (2015), pp. 3835–8 (cit. on p. 7).

[16] R. Molina et al. «A Prospective Study of Tumor Markers CA 125 and CA 19.9 in Patients with Epithelial Ovarian Carcinomas». In: *Tumor Biology* 13 (1992), pp. 278–286 (cit. on p. 9).

[17] R. C. BAST. et al. «New tumor markers: CA125 and beyond». In: *International Journal of Gynecologic Cancer* 15.Suppl 3 (2005), pp. 274–281 (cit. on p. 9).

[18] P. Buamah. «Benign conditions associated with raised serum CA-125 concentration». In: *Journal of surgical oncology* 75.4 (2000), pp. 264–5 (cit. on p. 9).

[19] C. Miralles et al. «Cancer Antigen 125 Associated With Multiple Benign and Malignant Pathologies». In: *Ann Surg Oncol* 10.2 (2003), pp. 150–154 (cit. on p. 9).

[20] R. Molina et al. «Mucins CA 125, CA 19.9, CA 15.3 and TAG-72.3 as Tumor Markers in Patients with Lung Cancer: Comparison with CYFRA 21-1, CEA, SCC and NSE». In: *Tumor Biol* 29 (2008), pp. 371–380 (cit. on p. 9).

[21] G. Funston et al. «The diagnostic performance of CA125 for the detection of ovarian and non-ovarian cancer in primary care: A population-based cohort study». In: *PLoS Med* 17.10 (2020) (cit. on p. 9).

[22] K. Boyeon et al. «Diagnostic performance of CA 125, HE4, and risk of Ovarian Malignancy Algorithm for ovarian cancer». In: *Journal of Clinical Laboratory Analysis* 33.1 (2018) (cit. on p. 9).

[23] S. Pignata et al. «Follow-up with CA125 after primary therapy of advanced ovarian cancer: in favor of continuing to prescribe CA125 during follow-up». In: *Annals of Oncology* 22.8 (2011) (cit. on p. 9).

[24] A. Fawzy et al. «Tissue CA125 and HE4 Gene Expression Levels Offer Superior Accuracy in Discriminating Benign from Malignant Pelvic Masses». In: *European Journal of Cancer Prevention* 17.1 (2016), pp. 323–33 (cit. on p. 9).

[25] S. U. Wei et al. «The diagnostic value of serum HE4 and CA-125 and ROMA index in ovarian cancer». In: *Biomedical Reports* 5.1 (2016), pp. 41–44 (cit. on p. 9).

[26] A. A. Ahmed et al. «Diagnostic accuracy of CA125 and HE4 in ovarian carcinoma patients and the effect of confounders on their serum levels». In: *Current Problems in Cancer* 43.5 (2018), pp. 450–460 (cit. on p. 9).

[27] V. R Iyer et al. «MRI, CT, and PET/CT for ovarian cancer detection and adnexal lesion characterization». In: *American Journal of Roentgenology* 194.2 (2010), pp. 311–21 (cit. on p. 10).

[28] A. Sahdev. «CT in ovarian cancer staging: how to review and report with emphasis on abdominal and pelvic disease for surgical planning». In: *Cancer Imaging* 16.19 (2016) (cit. on p. 10).

[29] Signe R. et al. «The diagnostic value of PET/CT for primary ovarian cancer–a prospective study». In: *Ginecologic Oncology* 105.1 (2007), pp. 145–9 (cit. on p. 10).

[30] M.A.G ElHariri et al. «Usefulness of PET–CT in the evaluation of suspected recurrent ovarian carcinoma». In: *Egyptian Journal of Radiology and Nuclear Medicine* 50.2 (2019) (cit. on p. 10).

[31] U. Menon et al. «Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial». In: *Lancet 2021* 197.10290 (2021), pp. 2182–2193 (cit. on p. 10).

[32] I. A. Qureshi et al. «Transvaginal versus transabdominal sonography in the evaluation of pelvic pathology». In: *Journal of the College of Physicians and Surgeons– Pakistan : JCPSP* 14.7 (2004), pp. 390–393 (cit. on p. 12).

[33] D. S. Watson. «Clinical applications of machine learning algorithms: beyond the black box». In: *BMJ* 364 (2019) (cit. on p. 24).

[34] C. Rudin. «Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead». In: *Nat Mach Intell* 1 (2019), pp. 206–215 (cit. on p. 24).

[35] Y. LeCunn et al. «Deep learning». In: *Nature* 521.7553 (2015), pp. 436–444 (cit. on p. 24).

[36] Y. Zhang et al. «A strategy to apply machine learning to small datasets in materials science». In: *npj Computational Materials* 4.25 (2018) (cit. on p. 24).

[37] S. Tsimenidis. «Limitations of Deep Neural Networks: a discussion of G. Marcus' critical appraisal of deep learning». In: () (cit. on p. 24).

[38] S. Tsimenidis. «Preparing Medical Imaging Data for Machine Learning». In: *Radiology* 295.1 (2020), pp. 4–15 (cit. on p. 24).

[39] World Health Organization. *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization, 2021, xvi, 148 p. (Cit. on p. 25).

[40] I. El Naqa et al. «A support vector machine approach for detection of microcalcifications». In: *Medical Imaging* 21 (12 Jan. 2002), pp. 1552–1563 (cit. on p. 26).

[41] M. Abdar et al. «A new machine learning technique for an accurate diagnosis of coronary artery disease». In: *Computer methods and programs in biomedicine* 179 (2019), p. 104992 (cit. on p. 26).

[42] S. Wang et al. «Classification of Alzheimer's Disease Based on Eight-Layer Convolutional Neural Network with Leaky Rectified Linear Unit and Max Pooling». In: *Journal of Medical Systems* 42 (5 2018), pp. 1–11 (cit. on p. 26).

[43] D. Karimi et al. «Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images». In: *Medical image analysis* 57 (2019), pp. 186–196 (cit. on p. 26).

[44] Gillies J.R. et al. «Radiomics: Images Are More than Pictures, They Are Data». In: *Radiology* 278.21 (2015), pp. 563–77 (cit. on p. 27).

[45] V. F. van Ravesteijn et al. «Computer-Aided Detection of Polyps in CT Colonography Using Logistic Regression». In: *Transactions on Medical Imaging* 29 (1 2010), pp. 120–131 (cit. on p. 27).

[46] T. W. Cary et al. «Comparison of naïve Bayes and logistic regression for computer-aided diagnosis of breast masses using ultrasound imaging». In: *Medical Imaging*. 2012 (cit. on p. 27).

[47] B. He et al. «MRI-based radiomics signature for tumor grading of rectal carcinoma using random forest model». In: *Journal of cellular physiology* 234 (11 2019), pp. 20501–20509 (cit. on p. 27).

[48] «Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform». In: *Measurement* 146 (2019), pp. 800–805 (cit. on p. 27).

[49] J. Ren. «ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging». In: *Knowledge-Based Systems* 26 (2012), pp. 144–153 (cit. on p. 28).

[50] M. Ghaderzadeh et al. «Deep CNN-Based CAD System for COVID-19 Detection Using Multiple Lung CT Scans.» In: *Journal of medical Internet research* (2021) (cit. on p. 28).

[51] N. Antropova et al. «A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets». In: *Medical Physics* 44 (2017), pp. 5162–5171 (cit. on p. 28).

[52] Y. Bar et al. «Deep learning with non-medical training used for chest pathology identification». In: *Medical Imaging*. 2015 (cit. on p. 28).

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. «U-Net: Convolutional Networks for Biomedical Image Segmentation». In: vol. 9351. Oct. 2015, pp. 234–241 (cit. on pp. 28, 38–40, 68).

[54] N. A. Siddique et al. «U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications». In: *IEEE Access* 9 (2021), pp. 82031–82057 (cit. on p. 28).

[55] X. Yang et al. «Fine-Grained Recurrent Neural Networks for Automatic Prostate Segmentation in Ultrasound Images». In: *AAAI*. 2017 (cit. on p. 29).

[56] P. Luc et al. «Semantic Segmentation using Adversarial Networks». In: *ArXiv* abs/1611.08408 (2016) (cit. on p. 29).

[57] B. Christoph et al. «Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images». In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham: Springer International Publishing, 2019, pp. 161–169 (cit. on p. 29).

[58] Frank Rosenblatt. «The perceptron: a probabilistic model for information storage and organization in the brain.» In: *Psychological review* 65 6 (1958), pp. 386–408 (cit. on p. 30).

[59] F. Rosenblatt at al. *The perceptron : a theory of statistical separability in cognitive systems*. Buffalo, N.Y.: Cornell Aeronautical Laboratory, 1958 (cit. on p. 32).

[60] J. A. Nichols et al. «Machine learning: applications of artificial intelligence to imaging and diagnosis». In: *Biophysical Reviews* 11 (2019), pp. 111–118 (cit. on p. 32).

[61] Yann Lecun. «A Theoretical Framework for Back-Propagation». In: (Aug. 2001) (cit. on p. 33).

[62] H. Wang et al. «The Role of Activation Function in CNN». In: *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*. 2020 (cit. on p. 36).

[63] J. Long et al. «Fully convolutional networks for semantic segmentation». In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3431–3440 (cit. on pp. 37, 38).

[64] J.A. Noble and D. Boukerroui. «Ultrasound image segmentation: a survey». In: *IEEE Transactions on Medical Imaging* 25.8 (2006), pp. 987–1010. DOI: 10.1109/TMI.2006.877092 (cit. on p. 44).

[65] R. Muzzolini et al. «Multiresolution texture segmentation with application to diagnostic ultrasound images». In: *IEEE transactions on medical imaging* 12 (Feb. 1993), pp. 108–23. DOI: 10.1109/42.222674 (cit. on p. 44).

[66] E. Gordon Sarty et al. «Semiautomated segmentation of ovarian follicular ultrasound images using a knowledge-based algorithm.» In: *Ultrasound in medicine biology* 24 (1 1998), pp. 27–42 (cit. on p. 44).

[67] B. Potonik et al. «Automated analysis of a sequence of ovarian ultrasound images. Part I: segmentation of single 2D images». In: *Image Vis. Comput.* 20 (2002), pp. 217–225 (cit. on p. 44).

[68] J. R. Tegnoor et al. «Automatic Detection of Follicles in Ultrasound Images of Ovaries using Active Contours Method». In: 2010 (cit. on p. 45).

[69] V. Kiruthika et al. «Automatic Segmentation of Ovarian Follicle Using K-Means Clustering». In: *2014 Fifth International Conference on Signal and Image Processing* (2014), pp. 137–141 (cit. on p. 45).

[70] D. S. Wanderley et al. «End-to-End Ovarian Structures Segmentation». In: *CIARP 2018: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. 2018, pp. 681–689 (cit. on pp. 45, 46).

[71] S. Marques et al. «Segmentation of gynaecological ultrasound images using different U-Net based approaches». In: *2019 IEEE International Ultrasonics Symposium (IUS)*. 2019, pp. 1485–1488 (cit. on pp. 45, 46).

[72] H. Li et al. «CR-Unet: A Composite Network for Ovary and Follicle Segmentation in Ultrasound Images». In: *IEEE Journal of Biomedical and Health Informatics* 24 (2020), pp. 974–983 (cit. on p. 45).

[73] et al. F. Christiansen. «Ultrasound image analysis using deep neural networks for discriminating between benign and malignant ovarian tumors: comparison with expert subjective assessment». In: *Ultrasound in Obstetrics & Gynecology* 57.1 (), pp. 155–163 (cit. on p. 45).

[74] M. AKAZAWA et al. «Artificial Intelligence in Ovarian Cancer Diagnosis». In: 40.8 (2020), pp. 4795–4800 (cit. on p. 45).

[75] Y. Zimmer et al. «A two-dimensional extension of minimum cross entropy thresholding for the segmentation of ultrasound images». In: *Ultrasound in Medicine Biology* 22.9 (1996), pp. 1183–1190 (cit. on p. 46).

[76] Y. Zimmer et al. «An automatic approach for morphological analysis and malignancy evaluation of ovarian masses using B-scans.» In: *Ultrasound in Medicine Biology* 29.11 (2003), pp. 1561–70 (cit. on p. 46).

[77] S. Rihana et al. «Automated algorithm for ovarian cysts detection in ultrasonogram». In: *2013 2nd International Conference on Advances in Biomedical Engineering* (2013), pp. 219–222 (cit. on p. 46).

[78] I. J. Hussein et al. «Fully Automatic Segmentation of Gynaecological Abnormality Using a New Viola-Jones Model». In: *Computers, Materials & Continua* (2021) (cit. on p. 46).

[79] J. Jin et al. «Multiple U-Net-Based Automatic Segmentations and Radiomics Feature Stability on Ultrasound Images for Patients With Ovarian Cancer». In: *Frontiers in Oncology* 10 (2021) (cit. on pp. 46, 66).

[80] Ph. Chlap et al. «A review of medical image data augmentation techniques for deep learning applications». In: *Journal of Medical Imaging and Radiation Oncology* 65 (2021) (cit. on p. 67).

[81] L. Zhang et al. «Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation». In: *IEEE Transactions on Medical Imaging* 39.7 (2020), pp. 2531–2540 (cit. on p. 67).