POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering

The role of doppler artifacts in color score assessment. Algorithms to remove artifacts from doppler signal and to evaluate vascularization of adnexal lesions in diagnostic ultrasound.

Supervisors

Prof. Filippo MOLINARI

Prof. Massimo SALVI

Dr. Rosilari BELLACOSA MAROTTI

Candidate

Chiara NOLI

December 2022

Acknowledgements

I would like to thank all the people who contributed with their collaboration and support to the development of this study.
I would like to express my gratitude to my supervisor Prof. Filippo Molinari and my co-supervisor Prof. Massimo Salvi from the Department of Electronics and Telecommunications of Politecnico di Torino for their assistance during the project and their insightful comments and suggestions.
Eventually, I wish to thank all the members of Syndiag srl, in particular my co-supervisor Dr. Rosilari Bellacosa Marotti and Eng. Flavia De Simone, who supervised all the stages of the project, guiding and helping me to overcome the many challenges that came with it, always showing great availability.

Table of Contents

Chapter 1

Introduction

Ovarian cancer accounts for 3.3% of all cancers in women worldwide, but it is a very aggressive type of tumor responsible for approximately half of the deaths related to gynecological cancer [1, 2, 3]. This high mortality is principally due to the difficult diagnosis and differentiation at an early stage [1]. Therefore, early detection and characterization of ovarian lesions is of utmost importance for an adequate management of the patient.

Nowadays, the method of choice to detect ovarian neoplasms is the ultrasound technique that is particularly convenient since it is non-invasive and at low cost. Specifically, Color Doppler and Power Doppler imaging represent useful tools to differentiate between benign and malignant neoplasms because the vascularization of a malignant mass may differ from that of a benign neoplasm [4, 5]. However, the main disadvantage of these techniques is that artifacts are particularly frequent, due to either inappropriate settings, anatomic factors or physical and technical limitations [6].

Since the ultrasound imaging technique is a complex diagnostic tool subject to the examiner interpretation, in literature several diagnostic standards were proposed with the aim of making the evaluation of ultrasound videos more objective and independent from the examiner. In this context, a group of researchers founded, in 1999, the International Ovarian Tumor Analysis (IOTA) with the aim of generating a standardized terminology to describe the sonographic features of adnexal lesions. Regarding vascularization, among the definitions introduced by IOTA, the color score is particularly interesting since it is a scoring system that indicates the degree of vascularization within ovarian masses. The color score is assigned by clinicians to the tumor as a whole and it is equal to 1 when no blood flow is detected within the lesion, 2 when the flow is minimal, 3 if the flow is moderate and 4 if the lesion is highly vascular [4].

The color score has been proved to be a good predictor of malignancy and, as a proof of it, has been included in several models (also developed by IOTA) that are

able to assess the probability that a lesion is malignant or benign.
However, the problem is that the estimation of the color content within a lesion is based on a subjective evaluation performed by clinicians. This means that different clinicians, with different levels of experience, may evaluate the same lesion differently, possibly resulting in a wrong interpretation of the adnexal mass and in a not satisfactory agreement among clinicians in assigning color score [7].
Moreover, the artifacts present in the doppler videos can influence the assignment of the color score making more difficult for clinicians to correctly interpret the flow information.
Therefore, developing an algorithm able to reduce the number of artifacts and to track, at the same time, the real signal within the mass during the whole doppler video, may be useful to simplify the clinicians' evaluation.

1.1 Ovarian Cancer

1.1.1 Epidemiology

Nowadays, ovarian cancer is the eighth of the most frequent cancers in women worldwide and represents the eighth most common cause of cancer death, with a mortality of 4.3% [1, 8]. In Europe, it is the main cause of death among gynecologic malignancies, ranking fifth in incidence (exceeded only by breast, colorectum, lung and corpus uteri), and it is the sixth bigger killer among all women's neoplasms (exceeded by breast, colorectum, lung, pancreas and stomach) [1].
The main issue related to this disease is the difficulty of its diagnosis at an early stage [9]. As a matter of fact, most of the times patients present ovarian cancer in advanced stages, mostly because the disease in the early stages is asymptomatic or associated with nonspecific symptoms. For this reason, the mortality related to ovarian cancer is high [1, 2].
This type of tumor is diagnosed at an advanced stage in approximately 70% of cases, otherwise it is very frequent in clinical practice to detect it incidentally [9]. Once diagnosed, the survival rate after 5 years is <30%; meanwhile if the ovarian cancer is detected at earlier stage when localized to the ovary, the survival is longer than 5 years for more than 90% of patients [2, 9]. This means that an early detection of any adnexal mass[1] is of immense importance.
Since the most significant factor for survival is stage at diagnosis, different screening methods have been developed over the years with the aim of detecting the ovarian cancer as soon as possible in order to reduce its high mortality [3].

[1]A mass in tissue near the uterus, usually placed in the ovary or fallopian tube. Adnexal masses include ovarian cysts, ectopic pregnancies, and benign or malignant lesions [10].

To treat ovarian malignancies properly, the type of tumor is a factor to be considered. Indeed, invasive tumors are commonly treated more aggressively than borderline tumors, especially if it is important to preserve the fertility. In selected cases, stage I ovarian cancer may be managed more conservatively than disease at advanced stages, whereas the treatment of cancers metastasized to the ovary depends on the nature of the primary tumor [3]. An accurate diagnosis of adnexal tumors before surgery will increase the probability that patients will receive the appropriate treatment.

1.1.2 Etiopathogenesis

Etiopathogenesis[2] of ovarian tumors seems to be multifactorial [1].
The principal risk factor is familiarity; however, only 5–10% of cases are due to hereditary syndromes, the main being the breast-ovarian cancer. The remaining cases are sporadic; nulliparity, early menarche and late menopause are risk factors in this case, while pregnancy, lactation, early menopause and use of oral contraceptives appear to be protective factors. Hereditary cases occur mainly in premenopausal age, while the sporadic ones affect mostly older women [1].

1.1.3 Classification of adnexal masses

Characterizing ovarian lesions is a diagnostic challenge of extreme importance because it allows to plan adequate therapeutic procedures and may influence patient's treatment. A multidisciplinary team is required to assess any adnexal mass properly through physical exams, laboratory tests and imaging techniques [1].
Ovarian neoplasms are distinguished in benign, borderline, or malignant tumors. An important issue to consider is that they are very common, but they are mostly benign and only a small part is borderline or malignant: $>90\%$ and $\leq 60\%$ of all cases of ovarian masses detected in premenopausal and postmenopausal women respectively are benign [9]. Moreover, borderline lesions have both benign and malignant features, so they are very difficult to detect [1].
Two staging systems have been developed to describe the spread of ovarian tumors: the TNM (tumor, node, metastasis) and the International Federation of Gynecology and Obstetrics (FIGO). According to these systems, stage I describes tumors limited to ovaries, stage II reflects pelvic extension or primary peritoneal cancer, stage III indicates spread to the peritoneum outside the pelvis and/or metastasis to the

[2]The cause and development of a disease or abnormal condition [11].

retroperitoneal lymph nodes, while stage IV refers to distant metastasis [1].
Primary ovarian tumors can be divided into three main groups based on tumor nature: epithelial, germ cell and sex cord-stromal tumors [1, 12].

Epithelial ovarian cancer arises from the surface of the ovary (the so-called epithelium), it accounts for 65% of all ovarian tumors and represents approximately 85% of ovarian malignancies [1, 13]. In this type of ovarian tumor, cancer cells form in the tissue that covers the ovary or lines the fallopian tube or peritoneum, defined as the serous membrane lining the cavity of the abdomen and pelvis and covering the abdominal organs. These diseases are referred as carcinomas [13].
Surface epithelial-stromal tumors occur more commonly in middle-age or older women and are rare in young adults, particularly before puberty [12].
These tumors are considered *benign* if they show a non-invasive behavior and low cellular proliferation, they are classified as *borderline* if there is exuberant cellular proliferation but no invasive behavior; and as *malignant* if they behave invasively.
The most part of borderline tumors behave like benign tumors and have a good prognosis, but some of them can show up again after the surgery, others, instead, may seed extensive implants within the abdominal cavity [12].
The surface epithelial-stromal cancers can be further divided into the following five major subtypes: serous, mucinous, endometrioid, clear cell, and transitional cell (or Brenner type) tumors [12].

Germ cell tumors originate from the reproductive cells of the ovaries. Ovarian germ cell malignant tumors are rare, they represent about one-fourth of all ovarian tumors and account for approximately 5% of all cases of ovarian cancer [13].
These tumors frequently affect only one ovary and are curable in about 95% of cases if they are diagnosed and treated at early stages. They can develop at any age but more often in young women or adolescent girls [13]. As a matter of fact, more than half of the ovarian neoplasms that occur in children and adolescents are germ cell tumors, and one-third of these is malignant. Conversely, germ cell tumors are relatively infrequent in adults, and the great majority of them are benign, with most being mature cystic teratomas (dermoid cysts) [12].

Sex cord-stromal neoplasms arise from connective tissue cells. These tumors account for approximately 8% of all ovarian tumors. Malignant ovarian stromal tumors are rare and represent approximately 1.2% of all primary malignant ovarian tumors [12, 13]. In contrast to the other two categories, sex cord-stromal tumors are frequently characterized by hormonal production, menstrual changes, or early puberty as well as symptoms of a pelvic mass; they are often found in adolescents and young adults [13]. Ovarian stromal tumors are often detected early and have a 75% survival rate. Some of these tumors, namely fibromas and thecomas, have a fibrous appearance, and some derive from the granulosa cells or their testicular sex cord counterparts, the Leydig and Sertoli cells [12, 13].

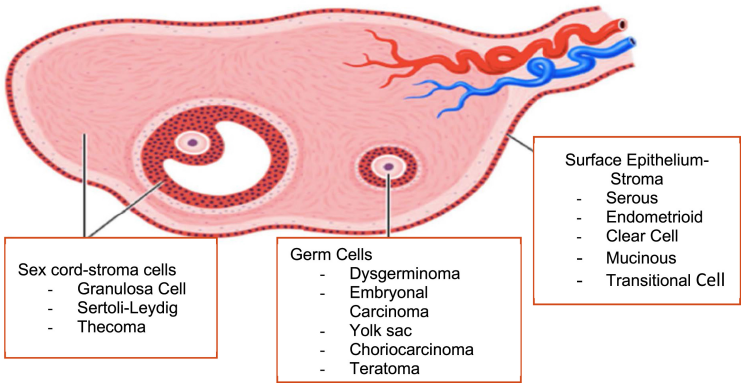


Figure 1.1: Ovarian cancer subtypes and its origin in the ovary [14]

Adnexal masses can be classified, from the morphological point of view, as unilocular cystic (i.e., a cyst with only one cavity), multilocular cystic (where at least two cavities are present), complex (cystic and solid that suggests the presence of tissue in the mass) and predominantly solid lesions [1, 15, 4]. These definitions have been introduced by the International Ovarian Tumor Analysis (IOTA) group with the objective of characterizing and describing ovarian neoplasms (see paragraph 1.2.1 for more details).

Unilocular cystic masses are mostly benign and can have non-ovarian or ovarian origin. The common extraovarian lesions are paraovarian cysts, hydrosalpinx, pyosalpinx and hematosalpinx, whereas ovarian lesions are usually represented by functional cysts and serous cystadenomas; finally, cystadenofibromas and mucinous cystadenomas are less common unilocular ovarian cystic masses.

- *Functional cysts* are the most frequent cystic masses in women of reproductive age as a normal part of the menstrual cycle, including follicles, follicular cysts, and corpus luteum cysts that results from a failure of the corpus luteum to regress. The corpus luteum cysts may become bigger because of an internal bleeding: follow-up is able to distinguish between hemorrhagic corpus luteum cysts and endometrioma.

- *Serous cystadenoma* is a benign ovarian cancer consisting of an unilocular cyst with a thin regular wall (less than 3 mm); the lining is flat without internal septations, papillary protrusions or solid components.

- *Cystadenofibroma* is an uncommon benign epithelial ovarian tumor with both epithelial and fibrous stromal components. It may be a purely cystic lesion or, more often, it presents as a complex cystic mass with thick septa and solid components [1].

Multilocular cystic adnexal masses can be benign or borderline; they are represented by endometriomas, mucinous cystadenomas and borderline tumors.

- *Endometriosis* is characterized by ectopic implantation of endometrial tissue outside the uterus, often involving the ovaries. Endometriotic cysts (or endometriomas) are often multifocal and bilateral. Patients suffering of endometriosis may develop ovarian malignancy (estimated risk about 2.5%).

- *Mucinous cystadenoma* is a benign mucin-containing tumor, often bigger than serous cystadenoma and monolateral. It usually presents as a multilocular cystic lesion with a thin regular wall and several septations, with no solid components.

- *Borderline tumors* are ovarian tumors characterized by epithelial anaplasia. They occur in younger patients with respect to malignant ovarian neoplasm, and they are more frequently serous and mucinous cystadenoma. They usually show non-invasive behavior, but there can be lymph nodes and peritoneal implants; however, they are characterized by a better prognosis than cystoadenocarcinomas. Borderline tumors have morphological features in between of benign and malignant ones. Borderline serous cystadenoma usually manifests as a complex cystic lesion with some septa and papillary projections [1].

If an adnexal mass has a mixed cystic and solid appearance, there is the risk, or at least the suspicion, of malignancy. However, some benign lesions like mature cystic teratoma also appear as complex.

- *Mature cystic teratoma* is the most common ovarian neoplasm, it is a benign germ cell tumor and affects mostly young patients. It consists of at least two of the three embryogenic germ cell layers, and usually contains ectodermal (skin, brain), mesodermal (fat, bone) and/or endodermal (thyroid tissue, gastrointestinal and bronchial epithelium) mature tissue. If these components are all present, the mass appears complex and heterogeneous; however, detecting fat-tissue inside the mass it is possible to produce a correct diagnosis. Malignant mature cystic teratomas are rare (1–2% of cases), and usually occur in postmenopausal women due to a squamous cell carcinoma originating from the cyst wall.

- *Struma ovarii* is one of the main subtypes of benign ovarian monodermal teratomas (defined as a type of ovarian teratoma in which one of the three germ cell layers is predominant). It is mainly composed of thyroid tissue, with no fat tissue.

- *Ovarian metastasis* represents about 5% of malignant ovarian tumors, and there is a potential risk of misjudgment if the primary tumor is not known. The

most frequent neoplasms that metastasize to the ovaries are stomach, colon, breast, lung, and contralateral ovary tumors. Ovarian metastases manifest more commonly as bilateral with a cystic and solid or a predominant solid morphological appearance.

- *Serous cystadenocarcinoma* is the most common type of surface epithelial neoplasms and accounts for about 40% of malignant ovarian tumors, whereas mucinous cystadenocarcinoma is less common and represents about 10% of ovarian malignancies. These tumors have a complex multilocular morphology, usually with thick and irregular walls, septations, solid components and papillary projections. They can be very large, even greater than 12–15 cm.

- *Endometrioid and clear cell tumors* are commonly associated with endometriosis. These tumors are usually seen as complex masses with solid and cystic components, but they can also be predominantly cystic. The rapid growth of an endometrioma, the multilocularity and the presence of mural nodules inside the hemorrhagic cyst should raise the suspicion of malignancy.

- *Granulosa cell tumors* are usually benign neoplasms; but they may also be malignant. Clinically, they can manifest by producing hormones; they consist of cystic and solid masses, but they can also appear as multilocular cystic or predominantly solid. These tumors can be further divided into two subgroups: adult and juvenile. The adult type is responsible for about 95% of cases and affects preferentially perimenopausal and postmenopausal women; whereas the juvenile type is less frequent and occurs in prepuberal children [1].

Predominantly solid adnexal masses can be benign, borderline, or malignant lesions. They include tumors of different origin: epithelial, germ cell, sex cord and metastatic lesions.

- *Fibromas, thecomas* and *fibrothecomas* are sex cord-stromal tumors and they represent the most common benign solid lesion of the ovaries. These tumors usually present with no symptoms and there may be an association with ascites and pleural effusion. Fibrothecomas consist of both fibrous tissue and theca cells with lipidic content. Fibromas are rare tumors generated from the spindled stromal cells that form collagen; in almost all cases they are benign and curable by surgical excision. Thecomas, instead, are formed by stromal cells resembling the theca cells that normally surround the ovarian follicles; most of them are unilateral and affect postmenopausal women [12].

- *Sclerosing stromal tumour* is a type of sex cord-stromal tumor into the thecoma-fibroma group. It occurs mostly in young women, manifesting as menstrual irregularities. This tumor has a predominantly solid appearance [1].

1.1.4 Diagnosis of ovarian cancer

The imaging technique that is more frequently applied to evaluate a suspected ovarian lesion is the ultrasound (US) technique because it has the advantages of being widely available, well accepted by patients, non-invasive and relatively cheap. Combining the conventional grayscale ultrasonography and Color Doppler features, obtained with transabdominal and/or endovaginal scanning, is possible to investigate both morphological structure and vascular organization of the ovarian mass, in order to characterize and differentiate ovarian tumors and provide an early diagnosis of malignancy by means of quantitative blood flow measurements obtained from tumor vessels [1, 2].
Sonographic evaluation of ovarian masses is based on multiple features - size, external contour, internal consistency, and signs of malignancy as ascites and peritoneal implants - and it correlates morphologic images with macroscopic pathologic features of tumor such as nonfatty solid tissue, thick (>2–3 mm) and irregular walls and septa, and papillary projections [1, 2].
Currently, there is no specific ultrasound criteria to distinguish between benign and malignant tumors. Several scoring systems based on the morphologic features have been proposed and developed; however, they concluded that it is not possible to differentiate benign malignant masses in a reliable way only basing on morphologic criteria [2].
Regarding the vascularization, thanks to the Color Doppler study, the presence and, eventually, the localization of new tumor blood vessel are shown. Therefore, knowing the blood flow characteristics, it is possible to predict if the tumor is benign or malignant: a mostly central blood flow is more often related to malignancy, while a peripheral vascularization is more characteristic of a benign lesion. Generally, the majority of malignant tumors show blood flow; conversely the absence of blood flow is a sign of tumor's benignity [1, 2].

1.2 IOTA group

Given the lack of standardized terms and procedures to derive categorical and continuous variables in gynecological sonography, a group of researchers from different centers gathered together to address this problem of standardization, giving rise to the International Ovarian Tumor Analysis (IOTA) group [4].
The group was founded in 1999 by Dirk Timmerman, Lil Valentin, and Tom Bourne aiming at developing a standardized terminology to obtain morphologic end-points by B-mode imaging and end-points of vascularity and blood flow by color Doppler imaging [4].
That is why, in 2000, IOTA published a consensus statement containing terms, definitions, and measurements useful to describe the sonographic features of adnexal

lesions, which today is widely used [16].
Afterwards, IOTA has focused its research and activity on the development of diagnostic methods in order to characterize the adnexal pathology and distinguish between benign and malignant ovarian tumors. Specifically, IOTA developed the Easy Descriptors, the Simple Rules, mathematical models based on logistic regression (LR1 and LR2) and the ADNEX model, which are employed by users with different levels of experience in the clinical practice to assess the risk of malignancy since they are very easy to use. These models have been validated and they showed optimal performances, comparable to the ones of an assessment performed by an expert sonographer [16, 17].

1.2.1 Standardized Terminology

The IOTA group produced a standardized terminology related to adnexal masses with the aim of homogenizing and setting a standard of quality, description, and evaluation of ultrasonography across different centers resulting in an increased diagnostic accuracy [17]. In this section, the most significant definitions are introduced.

An *adnexal lesion* is the portion of an ovary or of an adnexal mass that, according to the ultrasonographic results, is not consistent with normal physiology [4].

A *septum* indicates a thin echogenic strand of tissue running across the cyst cavity from one internal surface to the contralateral side. The septum is not considered as a solid component [15, 4].

An *incomplete septum* is a septum that is not complete in some scanning planes. If a cyst contains only incomplete septa, it is classified as unilocular, even though in certain sections the cyst appears multilocular [15, 4].

An adnexal mass is referred as *solid* when it shows echogenicity that suggests the presence of tissue (e.g. the myometrium, the ovarian stroma, myomas, fibromas). In many situations, it is not easy to distinguish between blood clots and solid tissue, but solid tumors can be detected and identified in the 3 following ways: when there is no internal movement in the adnexal mass while moving the US transducer; by the presence of the typical internal texture; by analyzing the vascularization of the tumor with color Doppler imaging. In particular, if blood flow is detected, the tissue is regarded as solid, whereas, if there is no flow, the diagnosis is not definitive [4].

Solid papillary projections are intended as any solid projections protruding into the cyst cavity from the cyst wall greater than or equal to 3 mm in height. They can be smooth or irregular (e.g., cauliflower-like). The hyperechogenic not vascularized

area that is present in a dermoid cyst or a sludge placed on the internal walls of endometriotic cysts are not considered papillary projections [15, 4].

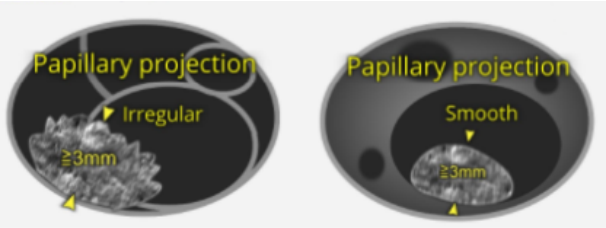


Figure 1.2: Pictorials of papillary projections with irregular walls and smooth walls [15].

The internal cyst wall is described as *smooth* or *irregular*. The wall is irregular if there is a solid papillary projection, a sludge or any irregularity in either the inner wall of the cyst or in the outer wall of a solid tumor or on the surface. In the contrary, if no papillary projections are present and the wall lining is flat, cystic walls are regarded as smooth [15, 4].

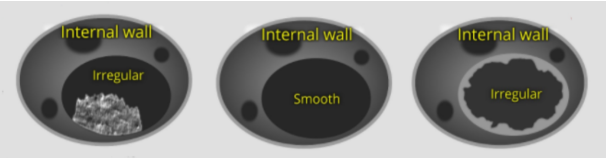


Figure 1.3: Pictorials of irregular and smooth cystic walls [15].

The cystic content is described as anechoic (meaning that it is black), low-level echogenic (as in mucinous tumors), ground glass appearance (where the cystic contents are homogeneously dispersed and echogenic, as it often happens in endometriotic cysts), hemorrhagic (in which strands show thread-like structures), or mixed echogenic (as seen in teratomas) [15, 4].

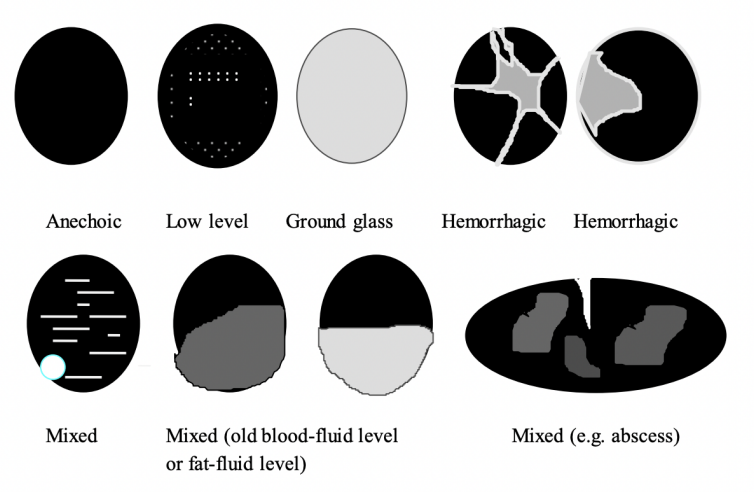


Figure 1.4: Pictorials of cystic contents' dominant feature [4].

Acoustic shadows are intended as loss of acoustic echo behind a structure that absorbs the sound [4].

Ascites is the fluid outside the pouch of Douglas, it can be present or absent [4].

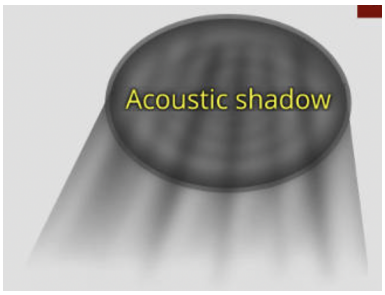


Figure 1.5: Pictorials of acoustic shadow [15].

The IOTA group has then introduced the following six categories in which adnexal lesions can be divided from the quantitative point of view, basing on their morphological features:

1. Unilocular cyst: a cyst having only one cavity without the presence of septa, solid parts or papillary structures [4].

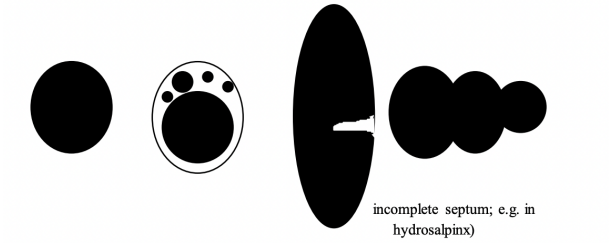


Figure 1.6: Pictorials of unilocular cysts [4].

2. Unilocular cyst with solid component: a unilocular cyst containing either measurable solid components, one or more papillary structures, or both [15, 4].

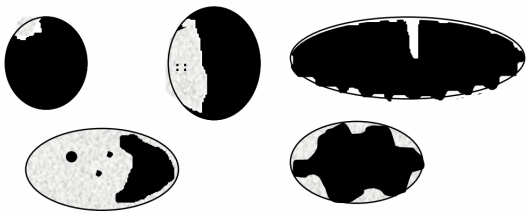


Figure 1.7: Pictorials of unilocular solid cysts [4].

3. Multilocular cyst: a cyst having at least one septum (so there are at least two cystic cavities), but without measurable solid components and papillary projections [15, 4].

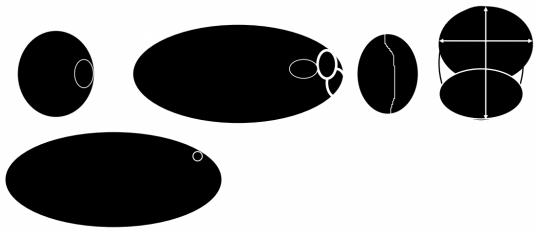


Figure 1.8: Pictorials of multilocular cysts [4].

4. Multilocular solid cyst: a multilocular cyst where there is a solid component or at least one papillary projection [15, 4].

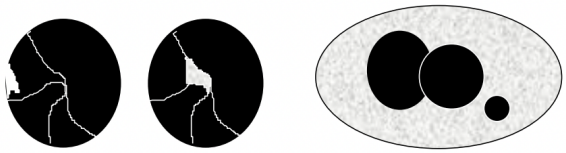


Figure 1.9: Pictorials of multilocular solid cysts [4].

5. Solid cyst: tumor in which at least the 80% of the mass is solid when assessed in a two-dimensional section. A solid tumor can be unilocular or multilocular and it may contain papillary projections [15, 4].

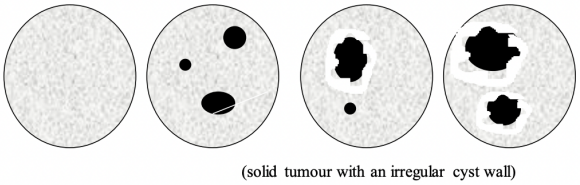


Figure 1.10: Pictorials of solid cysts [4].

6. Not classificable: a lesion falls into this category when the visualization is poor [4].

Color Score

As mentioned in section 1.1.4, doppler imaging techniques improve the diagnostic accuracy of gray-scale imaging and are particularly helpful in distinguishing between benign and malignant ovarian tumors because the vascularization of a malignant mass may differ from that of a benign neoplasm [4, 5].
In particular, in malignant lesions usually a higher amount of blood flow is shown due to angiogenesis. The color content of the lesion is strictly associated to its vascularity that, in turn, reflects the number and dimension of tumor vessels and their functional capacity [5].
For this reason, the IOTA group introduced a subjective semiquantitative assessment of the blood flow with the aim of assessing the degree of vascularization within ovarian masses. This scoring system is called color score and is a score between one and four that indicates the amount of blood flow within the septa, cyst walls or solid tumor area. The color score is assigned by the clinicians for the tumor as a whole and it refers only to the doppler image [4, 5]. It can assume the following values:

- 1 is given when no blood flow is present in the lesion.

- 2 is given when the mass contains a minimal flow.
- 3 is given when there is a moderate flow.
- 4 is given when the lesion is highly vascular with marked blood flow.

[4, 17, 15, 5]
Despite color score being a good predictor of malignancy and being employed in several predictive models also introduced by IOTA (as described in the next section), assigning the color score to an adnexal mass is complicated since the estimation of the doppler content within a lesion is influenced by the subjective evaluation of clinicians. This subjectivity leads to a not satisfactory agreement among clinicians in color score assignment, as illustrated in section 1.3.4.

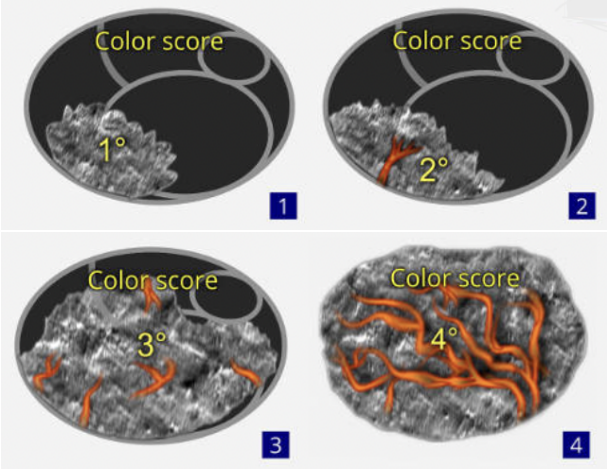


Figure 1.11: Pictorials of color score assignment [15].

1.2.2 Diagnostic models: Simple Descriptors, Logistic Regression models, Simple Rules and ADNEX

Simple Descriptors

Since many adnexal masses (like teratomas and endometriomas) are usually identified relatively easily because they show some typical features that are characteristic and not shared by other lesions, the IOTA group elaborated six rules named Easy descriptors (or Simple descriptors), that can be applied to detect these tumors immediately producing an instant diagnosis, without the necessity of employing more complex models [17].
There are four features typical of common benign lesions (the so-called benignity descriptors, BDs) and two descriptors suggestive of malignancy (malignant

descriptors, MDs). These rules take into account morphological characteristics assessed during ultrasound imaging, the detection of CA 125 tumor marker[3] and the patient's reproductive age.

The descriptors are:

- **BD1**: unilocular tumor with ground-glass echogenicity in a premenopausal woman (suggestive of endometrioma)
- **BD2**: unilocular tumor with mixed echogenicity and acoustic shadows in a premenopausal woman (suggestive of benign cystic teratoma)
- **BD3**: unilocular anechoic tumor with regular walls and maximum diameter of lesion < 10 cm (suggestive of simple cyst or cystadenoma)
- **BD4**: remaining unilocular tumors with regular walls
- **MD1**: tumor with ascites and at least moderate color Doppler blood flow in a postmenopausal woman
- **MD2**: age > 50 years and CA 125 > 100 U/mL

[15]

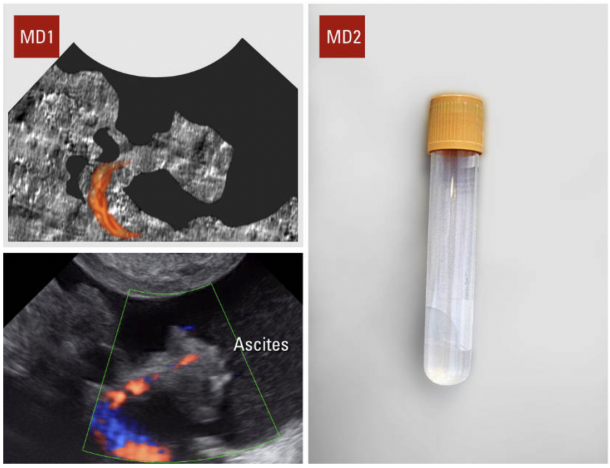


Figure 1.12: Examples of malignancy descriptors [15].

[3]Carbohydrate antigen-125 (CA-125) is the most commonly used serum tumor marker for epithelial tumors. It is extensively employed in clinical application for the monitoring of ovarian cancer, diagnosis, effective evaluation, and recurrence [9]. The CA 125 levels may be influenced by a high body mass index, ethnicity, the age of the patient, pregnancy, inflammatory processes and the presence of fibroids or endometriosis [17].

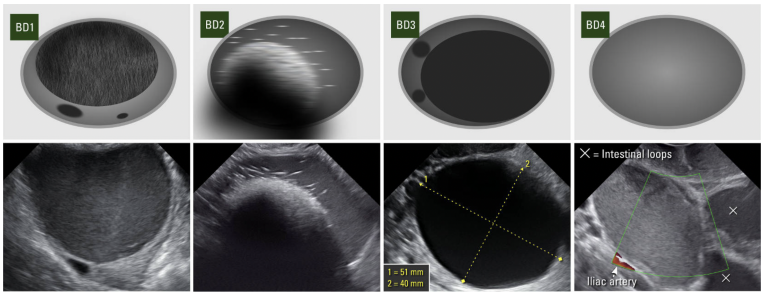


Figure 1.13: Examples of benignity descriptors [15].

Easy descriptors can be applied to about 43% of adnexal masses, in the remaining cases it is necessary to use more complex rules and diagnostic models [17].

Logistic Regression models: LR1 and LR2

The IOTA group has then developed and validated two predictive models, namely LR1 and LR2, based on logistic regression in order to evaluate the risk of malignancy in adnexal masses [18].
The LR1 model consists of 12 variables that are the patient's age, the presence of ascites, the presence of blood flow in a solid papillary projection, the maximal diameter of the solid component, the presence of irregular internal cyst walls, the presence of acoustic shadows, personal history of ovarian cancer, current hormonal therapy, the maximum diameter of the lesion, the presence of pain during the examination, the presence of a purely solid tumor and the color score of intratumoral blood flow [19, 20].
The LR2 model, instead, is simpler than the former since it is based only on the first six variables [19, 20].
It has been proved that both models perform well but LR1, containing all significant variables (including also the color score), performs better than LR2 [20].

Simple Rules

The Simple Rules represent a classification system for ovarian tumors, they have been formulated by clinicians and statisticians of IOTA, considering the clinical and ultrasound data from 1066 women recruited at different centers in Italy, Belgium, Sweden, France, and UK [21].
The Simple Rules consist of five features characteristic of benign tumors (B-features) and five rules applicable for malignant neoplasms (M-features). They can be applied to detect ovarian cancer in women who have at least one persistent adnexal (ovarian, para-ovarian, and tubal) tumor and require surgery [21].

The 10 rules on which the classification is based are:

- **B1**: unilocular cyst
- **B2**: presence of solid components with largest diameter < 7 mm
- **B3**: presence of acoustic shadows
- **B4**: smooth multilocular tumor with largest diameter < 100 mm
- **B5**: no blood flow (color score 1), no vascularization on color Doppler
- **M1**: irregular solid tumor
- **M2**: presence of ascites
- **M3**: at least 4 papillary structures
- **M4**: irregular multilocular-solid tumor with largest diameter ≥ 100 mm
- **M5**: very strong blood flow (color score 4)

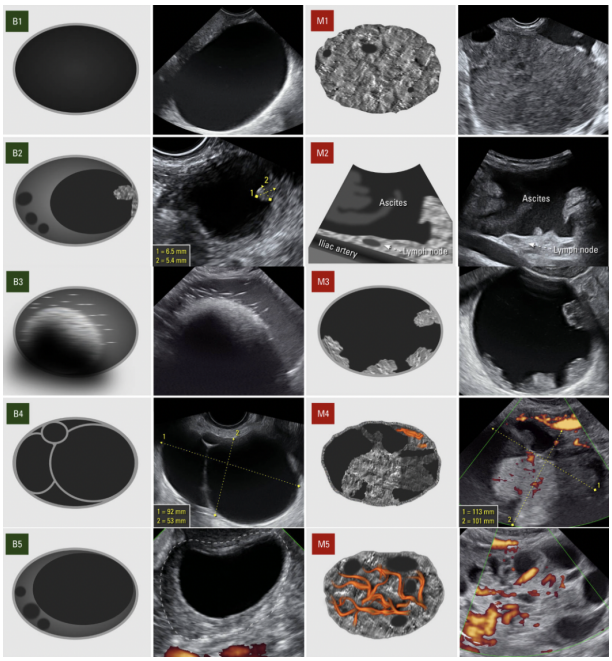


Figure 1.14: Examples of Simple Rules' applications [15].

If at least one B rule applies and no M rule is satisfied, the mass is immediately classified as benign, while if at least one M rule is present and no B rule is satisfied,

the tumor is classified as malignant. If both B- and M- features apply or if none of the ten rules is satisfied, the resulting diagnosis is inconclusive, and the patient needs second-stage diagnostic exams [21].
It is important to notice that, among the 10 Simple Rules, the ones that do not involve the evaluation of the lesion's vascularization employ quantitative parameters that can be measured or counted (for instance diameter, number of papillary structures, etc.), on the other hand, B5 and M5 features consider a variable - the amount of blood flow within the lesion, thus the color score - that cannot be quantitatively and objectively evaluated.

Nowadays, these rules are widely accepted and used on a daily basis in clinical practice as method of choice to assess the risk of malignancy. As a matter of fact, the simple rules are appliable to approximately 80% of adnexal masses, with a sensitivity of 95% and a specificity of 91% [17].
Nevertheless, the Simple Rules cannot replace the experience in ultrasonography and cannot compensate for poor quality ultrasound equipment [21]. Moreover, they provide a categorical output (either benign, malignant or inconclusive) based on dichotomized ultrasound features, not an actual estimation of the likelihood of malignancy, and about 20% - 25% of tested adnexal maxes result inconclusive, although it is possible to simply consider them malignant and refer them to a second line diagnostic exam.
The Simple Rules do not give a predicted risk, and so not a level of confidence in the classification. This limit was overcome by the introduction of a prediction model, namely ADNEX (see next paragraph), that is able to provide an individual estimated risk of malignancy for any type of lesion [17].

ADNEX

The IOTA group has also developed and published the Assessment of Different NEoplasias in the adneXa (ADNEX) model, a multiclass prediction model. This is the first risk model able to differentiate between benign and four types of malignant ovarian tumors: borderline, stage I cancer, stage II-IV cancer, and secondary metastatic cancer [22].
Moving from the previous logistic regression models to ADNEX, the IOTA group decided to select, as predictors of malignancy, only the more robust and less subjective variables, thus the color score was removed (see section 1.3.4 for details).
Therefore, the resulting indicators that constitute the ADNEX model are nine among which three are clinical and six are ultrasound variables that can be evaluated by examiners familiar with the IOTA terms and definitions.
The clinical predictors are age (evaluated in years), serum CA-125 (expressed in U/mL) and type of center (oncology center or other hospitals) to which the patient

has been referred.
On the other hand, the ultrasound predictors are: maximal diameter of the lesion (mm), proportion of solid tissue (defined as the ratio of the maximal diameter of the biggest solid component and the maximal diameter of the mass), number of papillary projections (that can be 0, 1, 2, 3 or larger than 3), presence of more than 10 cyst locules, presence of acoustic shadows, presence of ascites [22].
It has been demonstrated that this model can discriminate very well between benign neoplasms and the four types of malignancies reaching a sensitivity of 96.5% and specificity of 71.3% on the data employed in the validation studies. Moreover, its performances seem similar to, or even slightly better than, both LR2 and simple rules [22].

1.3 Color Doppler and Power Doppler Imaging

1.3.1 Color Doppler Imaging

Doppler evaluation hemodynamics can be employed to study the presence or absence of flow in a vessel, flow direction, pulsatility, and velocity [6]. Color doppler technique provides a visual image of the movement of blood through the heart, arteries, and veins, but it may be also used to image the motion of solid tissues such as the heart walls [23].
The most common target is blood for color flow imaging (CFI, also known as color Doppler imaging), while it is the solid tissue for Pulse-Echo (PE) imaging [23].
Standard PE imaging generates anatomical cross-sectional images of the body. Meanwhile, CFI is an imaging technique that combines anatomical information derived using ultrasonic pulse-echo techniques with velocity information derived using ultrasonic Doppler techniques to generate color-coded maps depicting movement superimposed on grey-scale images of tissue anatomy [23]. In this way, this imaging technique is able to image tissues and detect blood flow at the same time in order to immediately identify the direction of the different blood flows and circulation anomalies.
In principle, CFI techniques are similar to PE techniques in which the information about the position of each target in the body, corresponding to each pixel in the image, is obtained in the same way, namely by knowing the direction of the ultrasonic beam and the pulse round-trip transit time. The main difference between the two methods is that, in the case of CFI, the returning echoes are analyzed in terms of Doppler shift rather than amplitude [23].

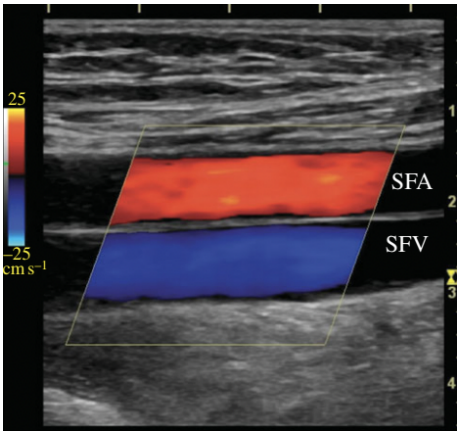


Figure 1.15: Color Doppler image showing the vascularization in the superficial femoral artery and the superficial femoral vein [23].

The Doppler Effect

Doppler shift or Doppler effect is defined as the change in frequency of a sound wave due to a reflector moving towards or away from an object that corresponds to the trasducer in the case of ultrasound. This phenomenon is named after the Austrian physicist Christian Doppler, who described it in 1842 [24].
When sound of a given frequency is discharged and then reflected from a source that is not moving, the frequency of the returning sound waves will be equal to the frequency at which they were emitted [24].
However, if the reflecting source is in motion either toward or away from the emitting source - in this case, the ultrasound transducer - the frequency of the sound waves received will be higher (positive Doppler shift) or lower (negative Doppler shift) than the frequency at which they were emitted, respectively [24].
In this context, the doppler equation is employed to calculate the magnitude of the frequency shift when reflectors are moving with respect to the ultrasound beam:

$$F = \frac{2 f_o v}{c} \cdot cos(\theta) \tag{1.1}$$

where:

- F is Doppler frequency shift,
- f_o is transmitted frequency from ultrasound probe,
- v is the velocity of moving reflector,
- c is the velocity of sound in the medium,

- θ is the angle between ultrasound beam and axis of flow [6, 24].

As shown from the equation, the magnitude of the Doppler shift is influenced by the angle at which the reflecting source is traveling with respect to the transmitting source. This factor is taken into account in the Doppler equation by the "cos(θ)" parameter; the maximum Doppler shift occurs when the Doppler angle is of 0 degrees (the cosine of $\theta = 1$) and no Doppler shift will be present when the motion of the reflecting source is perpendicular [24].
Therefore, in conventional CFI systems, the velocities measured and displayed are usually not the actual velocities but correspond to the components of the flow velocity of the target towards or away from the transducer [23].
The color doppler technique evaluates the velocity of the target towards the transducer from the phase shifts or time delays between echoes from the same sample volume during subsequent pulses. This changes in time or phase can be detected and electronically processed to produce a signal containing Doppler effect information. The frequency of this signal is the Doppler shift in the ultrasound frequency, hence the velocity of the target can be calculated from this frequency shift using the Doppler equation described above [23, 25]. In this way, significant information such as direction of flow relative to the transducer, flow velocity, and important flow conditions such as turbulence are inferred. Specifically, the direction of flow information extracted from the sequence of returning echoes is shown in the images by means of shades of red that normally denote flow toward the transducer and shades of blue representing the flow away from the transducer [25].

Characteristics of Color Doppler Imaging

Most color flow instruments produce a Color Doppler Velocity Image by applying signal processing of the Doppler signal, detecting the echo signal waveform and then estimating the mean velocity of the blood cells [25]. Even if, in principle, it should be possible to calculate velocity from just two pulse transmissions, in practice, the methods employed to estimate the velocity require multiple pulse echo sequences along each beam line in order to have a sufficient amount of data to estimate the mean Doppler frequency within each pixel [6, 23]. This is partly due to the stochastic nature of echoes from blood, but also because it is necessary to filter the returning signal from blood to reject larger signals from the surrounding solid tissues (which move, but with a much lower velocity) [23].
Nearly all modern CFI systems use array transducer technology, where the transducer consists of many single elements that transmit and receive ultrasound pulses. During the transmission, the beam-former is employed to apply the proper combination of signals to the individual transducer elements in order to generate an appropriate transmitted beam; whereas, during the reception, it is used to correctly combine the returning echoes to generate the appropriate receive beam. The output

from the beam-former is amplified in a time-dependent manner (to compensate for the additional attenuation experienced by echoes from deep within the body), and then demodulated to generate the components of the Doppler signal [23].

Limits of Color Doppler imaging

Color Doppler imaging is now provided on almost all commercial ultrasound machines and has been proved to be a powerful technique in assessing blood flow in many clinical conditions [23]. However, even if the mean frequency color Doppler is the most commonly used parameter, it also has a number of weaknesses.
The first main issue related with the use of the mean as the parameter of choice is that random noise can look like flow in any direction since it has a random frequency shift in the ultrasound imaging. As image noise increases, the more aberrant flow there appears to be, and the more the background seems to fill up with flow-related artifact. In the worst cases, this random noise totally dominates the image, and the identification of true flow becomes impossible [26].
Moreover, being a frequency-detection technique, color Doppler necessarily aliases (for more details see paragraph 1.3.2). In presence of aliasing, vessels look discontinuous and the directional and speed information are distorted. This problem is particularly relevant in slow-flow situations where low pulse repetition frequencies produce aliasing. Therefore, avoiding aliasing, the diagnostic detection of the presence or absence of flow would improve [26].
Another drawback of this technique is the difficulty to image deep vessels both because of sensitivity issues (due to the attenuation introduced by overlying tissues) and because of the increased inter-pulse interval necessary for the ultrasound pulses to make a round-trip from the transducer to the target (this results in a decreased pulse-repetition frequency that may lead to inadequate sampling of the Doppler signal resulting in misinterpretation of the Doppler shift frequency) [23].
Finally, color Doppler is angle dependent since the values of the displayed velocity components depend on the angle between the true velocity direction and the beam direction at each sample volume [25]. Doppler devices lose sensitivity to flows that are perpendicular to the sound field, and the frequency shift is greater when the object's surface is parallel to the receiving beam, thus requiring that the beam should be directed to detect these flows. Since the beam cannot be steered in all directions, lost segments not containing flow can occur. Consequently, Color Doppler can underestimate the presence of vessels because some vessels may not be identified when they course perpendicular to the sound field. This issue, added to the multiple colors within each vessel caused by varying insonifying angles along a vessel's path, can make vessel tracking and identification difficult [26].

1.3.2 Artifacts

As previously mentioned, Doppler Flow Imaging techniques allow to study and analyze blood flow features resulting in the assessment of tumor vascularity. However, the main downside of these methods is that artifacts are very frequent.
Artifacts in color Doppler imaging can be confusing and may lead to a wrong interpretation of flow information. That is why recognizing artifacts is of immense importance in order to arrive to a correct diagnosis. There are three main causes that may produce different types of artifacts: inappropriate equipment settings, anatomic factors, and physical and technical limitations of the modality [6].

Artifacts related to inappropriate settings

Doppler gain setting errors
Setting the gain properly is necessary in Doppler processing to depict the flow characteristics. As a matter of fact, if the gain is too low, valuable flow information can be lost since Doppler shifts are not displayed in vessels, especially those with relatively slow flow. In the contrary, if the gain setting is too high the image becomes cluttered with color noise in a random pattern [6].
In this context, the blooming artifact can be present in the Doppler image. It usually occurs in large vessels and it is characterized by the spread of the colored area beyond the vessel walls. In order to avoid the presence of this artifact, the gain must be lowered [27].

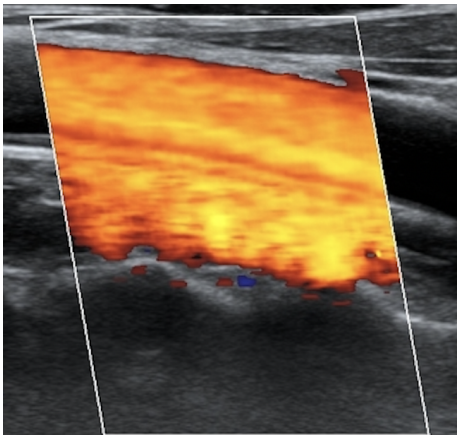


Figure 1.16: Typical appearance of the ultrasonic blooming artifact [28].

Velocity Scale errors and Aliasing
Another parameter that is crucial to correctly display doppler signals is the velocity scale. When the velocity range is too high, it is possible that low-flow information is

not shown. Conversely, if the velocity scale setting is too low for the flow conditions present, aliasing occurs. This phenomenon is related to the fact that color flow uses pulsed sound beams [6].
In particular, when blood velocities are high, the measurement of high Doppler shift frequencies requires that echo signals are collected at a high rate. However, there is a limit to the rate at which echoes are returned since there should be enough time must between each pulse transmission in order to collect the related echoes. There is therefore an upper limit on the Doppler shift (called Nyquist sampling rate) which can be measured and hence a corresponding upper limit on velocity [25]. When the frequency of the Doppler signal is higher than the Nyquist sampling rate (i.e., the half of the pulse repetition frequency), ambiguous or aliased signals are generated [6, 27]. The aliasing artefact appears as regions of wrongly colored pixels in a Color Doppler image [25].
Aliasing can be overcome applying the following expedients: increasing the Doppler angle (that, in turn, decreases the Doppler shift), increasing the velocity scale (which increases the pulse repetition rate), changing the baseline setting, or using a lower ultrasound frequency [6].

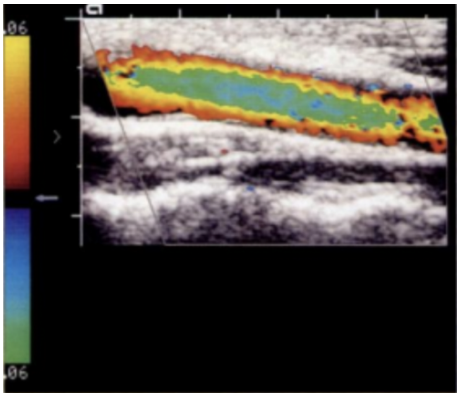


Figure 1.17: Aliasing artifact with flow reversal [6].

Incorrect wall-filter setting
Filtration is employed in order to remove unwanted Doppler signals at low frequency originating from soft-tissue reflectors moving slowly. If the filtration is too high, velocity information significant from the diagnostic point of view can be lost [6].
In the contrary, due to too low wall-filtering settings the edge artifact can appear in the doppler video. This artifact might also result from high velocity scale settings, and it occurs when the Doppler signal is identified at the margins of strong reflectors, therefore blood flow appears along their rim and may mimic vessels [27].
The twinkling artifact is another type of artifact produced by a narrow band of machine noise, called *phase jitter* that is usually excluded by wall filters. When in

presence of a rougher and strongly reflecting surface (such as renal calculi, bladder calcifications or cholesterol crystals), this noise is amplified, and the artifact results in a *mosaic* of different colors quickly changing. In order to identify this artifact and detect the stones, it is necessary to set color-write priority to a high value and grayscale gain to a minimum [27].

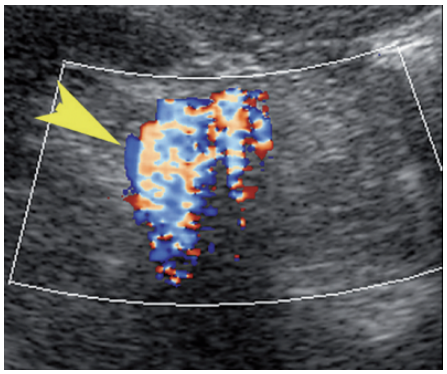


Figure 1.18: Edge artifact [29].

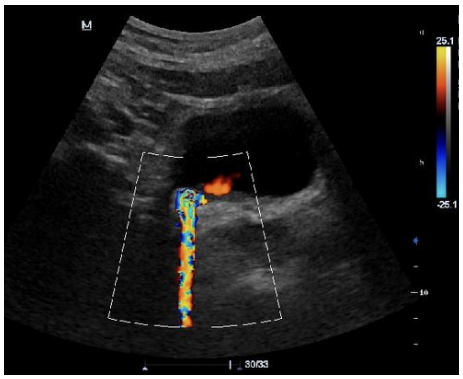


Figure 1.19: Twinkling artifact [30].

Anatomically related artifacts

Mirror image artifact
Mirror image artifacts occur with color Doppler imaging of any vessel close to a highly reflective surface, such as the lung. These artifacts are produced when echoes coming from a reflecting object are directed to another reflector before going back to the transducer. Because of this, blood flow is displayed identically on both sides of the reflector, even though the real signal is only on one side [6].

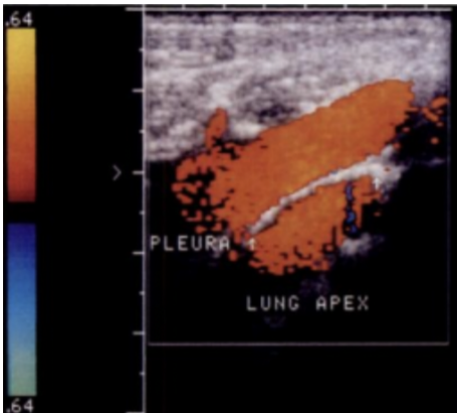


Figure 1.20: Mirror artifact in subclavian vein [6]

.

Vascular-Motion artifact
When a vessel moves with respect to the transducer, artifactual variation in velocity can be introduced into the spectral tracing as the sample volume passes through different velocities in a laminar flow state. This artifact can appear in the portal vein and its branches. It may be reduced by increasing the size of the Doppler gate so that the entire vessel is included, by imaging other portal branches, or varying the angle [6].

Color in nonvascular structures
Color flash artifacts are random bursts of color that fill large portions of the image, obscuring everything underneath, and they are caused by sudden internal or transducer motion (figure 1.21). These artifacts can be perceived in areas of low echogenicity such as cysts or ducts [6, 27].
Most color flow processors incorporate motion discriminators able to separate true flow from random motion of soft-tissue reflectors. However, the lower-level signals identified in the hypoechoic soft-tissue regions less effectively trigger the motion discriminator and so the color flash is not suppressed. This artifactual color signals can be erroneously considered as real flow, especially if the color sensitivity settings are high [6].

Pseudoflow artifact
Pseudoflow artifact occurs when any physiological fluid different than blood moves; the motion produces Doppler signal that can be detected as real blood flow, but, in reality, no vessel is present (figure 1.22). This type of artifact is frequently shown in presence of ascites, amniotic fluid (when imaging the uterus), urine (when imaging the bladder) and also within the walls of serous ovarian cysts [27].

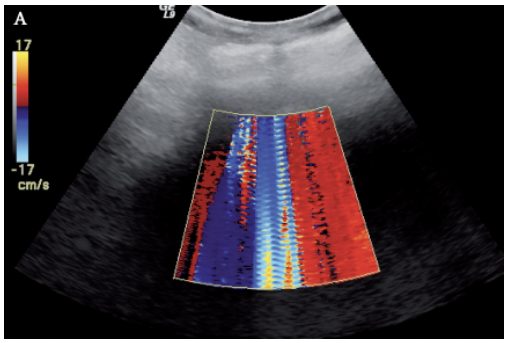


Figure 1.21: Example of color flash artifact [29].

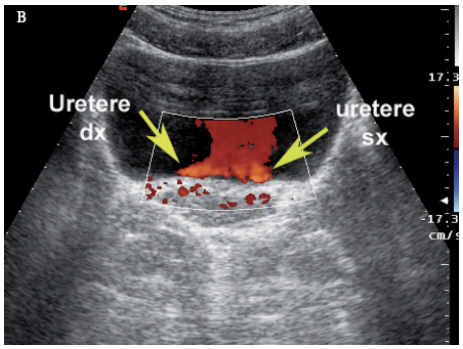


Figure 1.22: Example of pseudoflow artifact [29].

Instrument and processor related artifacts

Grating or side lobe artifact
Electronically focused, phased-array transducers focus the primary beam toward the Doppler sample volume. However, due to the distance between the array elements and the sequence of firing, weak secondary lobes of focused sound may interrogate areas/vessels that are not related to the primary beam. Side lobes occur in proximity to the primary beam, while grating lobes can be quite far removed from it [6].
An artifact that might be caused by grating or side lobes of the beam is the partial volume artifact. It is a transducer-related artifact that appears when the slice thickness is not infinitesimal, and the reflecting object is partially in the slice and partially not. Therefore, blood flow is extended to the whole slice and appears in areas that should be anechoic. This often happens when imaging the ovary and part of the iliac artery is shown as if there is vascularization inside the cystic walls but, considering a different plane of imaging, it is demonstrated that blood flow is outside of the cystic walls [27].

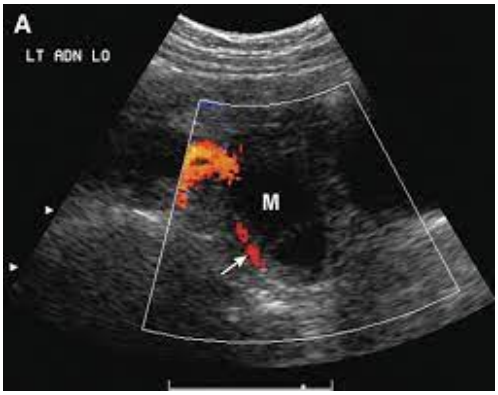


Figure 1.23: Partial volume artifact [27].

1.3.3 Power Doppler Imaging

The Power Doppler technique represents an alternative to mean-frequency Color Doppler, it encodes the power in the Doppler signal in color. With this imaging technique it is possible to explore all vessels, typical of growing tumors, allowing a complete analysis of the tumor vascularization [26].
The power of the Doppler signal is determined by the power of the echoes (derived from the amplitude of the echoes) related to the number of blood cells in the sample volume. Power is therefore a readily obtained measure of the number of moving cells in the sample volume [25].
The principal advantage of using power is the representation of random noise that,

in the power mode, is different from that in the mean frequency mode. This is because the noise always has uniformly low power, so encoding power in color and increasing the sensitivity of the color Doppler unit to image the noise floor, a uniformly colored background is imaged representing low power instead of a random distribution of colors representing any possible flow. Any true flow will have more power in the Doppler signal than the noise, and hence it will emerge from the noise background [26].

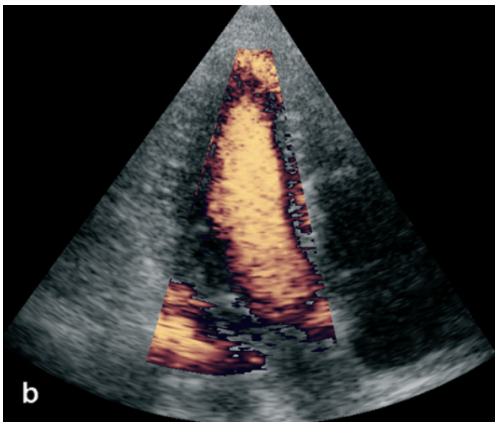


Figure 1.24: Power Doppler image of blood flow in the left ventricle [25].

Advantages of Power Doppler Imaging

The main attraction of Power Doppler Imaging is that it is a sensitive technique useful to study blood flow in small vessels or vessels that are deeper with respect to the skin surface; it therefore gives more complete images of vascularity than Color Doppler Imaging. As a matter of fact, in these cases, Color Doppler is less informative since the signal appears much weaker [26, 25].
Another significant advantage of this technique is that it is not prone to aliasing artifacts because it only indicates the presence of flow and does not attempt to measure velocity [25].
Moreover, compared with mean frequency Color Doppler, Power Doppler Imaging is relatively angle independent. The reason is that the total power in the signal is represented by the integral under Doppler power curve (i.e., the power spectrum). The power in the Doppler signal is related to the number of moving scatterers and red blood cells. So, changing the angle of insonification relative to the red blood cells, their mean Doppler shift will change, but the power remains the same. As a result, a power image will change very little with changing angle [26].
At perpendicular incidence the power will be lower, but it may not be zero. Hence, vessels look continuous, and, in many cases, there is no need to steer the beam.
Other advantages are better boundary detection (power mode depicts continuous

smooth representation of boundaries that, in this way, are easy to see), improved quantification of vascularity, better 3D depiction of blood vessel anatomy, and advantageous properties of blooming in power mode with contrast agents.
The first two advantages are both explainable with the fact that power mode is a continuous estimator of the amount of blood in a pixel compared with standard mean frequency color Doppler, which is a bistable estimator of pixel blood content. Bistable means that an arbitrary threshold, the color-write echo priority, is defined for color Doppler below which no flow is shown and above which a pixel is written as entirely containing flow [26].
The power level at each pixel is presented as a level of brightness and corresponds to the vascularization that is present: different amounts of blood in a pixel will appear with different powers giving the possibility to obtain a continuous map of vascularity. In addition, it is possible to compensate the Doppler power for depth and transducer affects. By comparing the power in a given pixel to the power in a neighboring blood vessel, one can normalize the tissue signal for depth and transducer affects.
Finally, power doppler employs a larger part of the dynamic range of the Doppler signal, increasing overall sensitivity, which is three times larger than that of Color Doppler, also because angular difference-based errors are not relevant [26].

Limitations of Power Doppler Imaging

Power Doppler technique has also some limits. The first disadvantage is that it does not provide any information about flow speed or direction.
Moreover, it is a high-sensitivity technique in which any motion is detected, thus it is extremely sensitive to motion artifacts, like flash artifacts. Therefore, soft tissue motion, that can be difficult to distinguish from blood flow, may seriously degrade the image. This limitation is explained by the high flow intensity of power imaging and the longer time necessary to build the image [26].
Edge artifacts are also more commonly seen in Power Doppler, due to an increase in the dynamic range.

1.3.4 Color Score

The color score can be a variable having strong diagnostic value, able to predict whether an ovarian tumor is benign or malignant by characterizing the amount of blood flow within the lesion.
For this reason, the color score was one of the possible variables that could be selected as predictors of the ADNEX model, described in the paragraph 1.2.2.
In particular, the selection of the ADNEX predictors was performed by clinicians with experience in characterizing adnexal masses, who, in the first phase, identified

a limited set of variables that could represent useful predictors. This first selection was based on the variables' perceived predictive value to distinguish between the four types of malignancies, subjectivity, dependency on the experience of the ultrasound examiner, and impact on the patient. In addition, the variables were selected or discarded basing on an analysis of consistency conducted across the IOTA study centers from which resulted that color doppler variables, including color score, and the presence of cyst wall irregularities should not be included as predictors in the model [22]. From this analysis conducted by Wynants et al. in 2013, the color score of intratumoral blood flow was proved to be a subjective variable with eight among the forty participant physicians that were detected as outliers for the color score, five with high and three with low values, possibly due to the use of Color or Power Doppler ultrasonography by different examiners [31]. After the analysis, ten variables remained, on which further statistical selection was conducted and, at the end, one variable, namely the family history of ovarian cancer, was omitted and the ADNEX predictors became nine [22].

Agreement among clinicians in color score assignment

Since the evaluation of the color content within the tumor is subjective, the values of color score are likely to vary both within and between examiners [32]. This variability has an influence on the calculation of the risk of adnexal mass malignancy.

A study by L. Zannoni et al. estimated the intra- and inter-observer variability in assessing the color scores to 103 adnexal masses examined through color/power doppler imaging. Four expert and three less experienced sonographers evaluated the lesions twice before and twice after a consensus meeting. During this meeting emerged that there were some differences in assignment between and within observers because they had different opinions or were uncertain on how estimating the color content of the lesion. [32]
No significant differences were obtained in the assignment between more experienced and less expert observers. In particular, the intra-observer repeatability of color score was considered good or very good, and the consensus meeting did not produce any relevant improvement. Instead, inter-observer concordance was estimated as moderate or good with a slight improvement registered thanks to the consensus meeting; however, neither intra- nor inter-observer agreement were associated to the level of experience of the sonographers [32].

In the study conducted by R. Massobrio, 70 ultrasound videos of adnexal masses were evaluated by 11 clinicians having different experience in the diagnosis of ovarian tumors. The clinicians were asked to assign a color score to each video and their agreement has been quantified. It was obtained that the concordance among

all the observers was discrete (54.5%); whereas the agreement of more expert sonographers was higher, reaching a moderate concordance of 60.9%. Instead, clinicians with less experience showed an average agreement of 52% [7].
Unlike the other study, in this case the experience of the examiners was proven to be related to the color score assignment and the agreement among clinicians.

These studies show how difficult is to quantitatively evaluate the color content of the adnexal masses and that the agreement of different observers in the assignment of color score is still not optimal mainly because it is a subjective assessment that can be also influenced by observer's experience.

Color score as a predictive variable in literature

Given the subjectivity of the color score and the difficult agreement among clinicians in its assignment, the usefulness of this parameter in the distinction between benign and malignant ovarian neoplasms has been strongly debated. As a matter of fact, in literature, there are several studies that support, also with quantitative results, the use of color score and others which disagree with that mainly because of the subjectivity and the lack of reliability of this index.

The study of Sharon et al. evaluates prospectively and compares the usefulness of color doppler, spectral doppler, and gray-scale sonography in distinguishing between benign and malignant adnexal masses. In this study, 170 adnexal masses in 161 patients were classified as benign or malignant basing on gray-scale morphology, internal flow versus peripheral or no flow, and spectral Doppler pulsatility. Among the 170 masses, 123 were benign and 46 were malignant as revealed by the surgical pathology, one malignant mass was confirmed by cytologic evaluation.
The grayscale analysis showed that 46 of the 47 malignant masses were suggestive of tumor and 76 of the 123 benign masses were not. It resulted that a gray-scale prediction of benignity was reliable (NPV = 99%), whereas a prediction of malignancy was unreliable (PPV = 50%).
It was, then, investigated the use of Color Doppler imaging as a possible tool for improving the specificity of gray-scale sonography.
Color Doppler imaging detected flow in 153 masses. Internal flow in a solid component or septum was present in 36 (77%) of the 47 malignant masses and in 38 (31%) of the 123 benign masses. Peripheral flow along or within the wall of the mass, without internal flow, was shown in 69 (56%) of the benign masses and in 21% of the malignant masses. No flow was identified in 16 (13%) of the benign masses. Flow was absent in only one malignant mass that had gray-scale characteristics suggestive of malignancy.
As a result, the presence of internal flow was not useful in distinguishing benign

from malignant masses (PPV = 49%), the sensitivity and specificity for internal flow versus either peripheral or no flow were 77% and 69%, respectively. The most useful information provided by color flow sonography involved patients who have no detectable flow because it suggested benignity (NPV = 94%), but only in 10% of the masses the absence of internal or peripheral color flow was verified.
Therefore, from this study it emerges that gray-scale findings, although imperfect, provide a more useful guide in evaluating these masses, especially in the prediction of benignity, and color Doppler imaging has limited usefulness in the evaluation of ovarian masses.
However, it must also be considered that, as proven by other studies (Fleischen et al., [33]), there is a statistically significant difference between vascularity in benign lesions (which tend to be peripheral) and that in malignant lesions (which tend to be internal). This difference is also shown in this study where it was obtained a higher percentage of malignant lesions with internal flow (77%) than the percentage of benign lesions with such flow (31%) and this result was highly statistically significant but, since the PPV was less than 50%, this information is not useful [34].

In contrast, the study of Timmerman at al. suggests the importance of including color score and other variables in a logistic regression-based model to distinguish between malignant and benign neoplasms. In particular, the study describes a statistical model that estimates the probability of malignancy for individual patients with a high sensitivity for malignancy while maintaining a high specificity.
This study shows that the most useful parameters for the logistic regression analysis were the menopausal status, the serum CA 125 level, the presence of one or more papillary growth, and a color score, from 1 to 4, indicative of tumor vascularity and blood flow.
Moreover, the presence of measurable arterial blood flow within the adnexal lesion and a high color score (3 or 4) were proved to be highly discriminating factors for differentiation between benign and malignant adnexal masses.
The model had a specificity of 87.1% and an accuracy that exceeds the performance of the widely used Risk of Malignancy Index. Besides that, the most important feature of the model remains the high sensitivity (95.9%), because clinically it is worse to have a false-negative test result than a false-positive test result for a patient with an adnexal mass who might need referral to a gynecologic oncologist for appropriate surgical intervention [35].

In another study, Abbas et al. developed a multiparameter scoring model using four gray-scale ultrasound and two color Doppler features (among which there is the color score) that has shown a high sensitivity and specificity for prediction of malignancy in adnexal masses compared with other scoring systems thanks to

the chosen variables. The features that resulted best fitted to predict the adnexal mass status and allowed statistically significant discrimination of benignity from malignancy were: volume of mass, type of mass, presence and thickness of septae, presence and length of papillary projections, location of vessels at color Doppler and color score. These features were, then, combined to obtain the scoring model, called Assiut Scoring Model.
Patients were evaluated by 2D ultrasound for morphological features of the masses combined with color Doppler examination of their vessels.
Among 115 benign masses, Color Doppler study detected 107 masses as benign (having a color score of 1 or 2) but labelled eight masses as malignant that were actually benign, while out of 46 malignant masses, only 29 masses were correctly diagnosed as malignant, the color score assigned to these masses was 4.
Another aspect to be considered is that, also in this study, Color Doppler results showed predominantly peripheral localization of vessels in benign masses (69.6%) and predominantly central or septal vessel localization (39.1% and 34.8%) in malignant masses. The 17.4% of masses showed absence of blood flow, while nearly all the malignant masses showed vascularity (97.8%) [18].

As shown from the described articles and others - [36, 37] - and from the decision to not include this parameter as a predictor in the ADNEX model, the estimation of the color score, intended as the color content of the tumor scan, to assess tumor vascularity is based on a subjective evaluation. Moreover, this element of subjectivity might have a negative effect on the reliability of the ultrasound methods incorporating doppler variables making them difficult to reproduce.
However, at the same time, it also emerges that this parameter represents a useful indicator employed in several models to successfully differentiate benign ovarian tumors from malignant ones.
Therefore, given the importance of this measure and its subjectivity, it is necessary to find a method that makes the color score a more reliable and effective indicator of malignancy. In this study, two methods that try to reach this goal, by removing the several artifacts present in doppler videos that may influence the value of color score assigned to the lesions, are desccribed.

Chapter 2

Aim of the study

Ovarian cancer, despite its not so large incidence in women, has a significantly high mortality, primarily due to the difficulty in identifying and diagnosing this neoplasm at the early stages [1, 2, 9] (see section 1.1).
In order to overcome this issue, Color Doppler and Power Doppler imaging techniques are employed as powerful, non-invasive and relatively cheap tools to distinguish between benign and malignant tumors by evaluating the degree of vascularization within the adnexal mass.
In this context, the IOTA group introduced a scoring system, namely color score, that indicates the amount of blood in septa, cyst walls or solid areas of the tumor. The color score assigned by clinicians is 1 when there is no blood flow in the lesion, 2 if the flow is minimal, 3 when the mass contains a moderate flow, 4 if there is a marked blood flow within the mass [4, 17] (see section 1.2).
However, assigning the color score to a lesion is not always immediate and different clinicians, with different experience level, may interpret the same adnexal lesion differently, resulting in the assignment of different color score values and in an overall low agreement between clinicians.
Moreover, Doppler imaging techniques suffer from several types of artifacts that can make more difficult for clinicians to interpret the flow information and produce a correct diagnosis (see section 1.3).
Therefore, in this thesis I will describe two algorithms that have been developed with the aim of removing the artifacts and easing the clinicians' evaluation:

- Pixel-based denoising algorithm: algorithm that suppresses artifacts based on the temporal persistence of colored doppler pixels within the video, assuming that artifacts are less persistent than real vascularization.
- Connected components-based denoising with component-tracking: algorithm that takes into account both temporal and spatial persistence, it relies on connected components, rather than single pixels, to assess signal and artifact

persistence. It suppresses artifacts while tracking the activations' clusters – connected components of colored pixels – during the video.

In this study, a quantitative assessment of the performances of these two artifact-removal algorithms was performed in order to understand

- if their application can help clinicians in evaluating the vascularization within the adnexal lesions and in assigning the color score,
- which of the two algorithms performs better.

In order to perform this assessment, a decision tree was trained and tested to predict the color score based on the estimation of the doppler signal within the lesion, obtained on 106 videos of ovarian cancer cases. The quantity of doppler signal was estimated as the number of colored pixels present within the lesion, whereas the color scores, employed as labels, were assigned by six expert clinicians. The starting hypothesis was that, by applying the artifact-removal algorithms and, thus, considering only the doppler activations *survived* to these algorithms, the color score prediction is more accurate than the one obtained on the original videos where the artifacts are not removed.
Three experiments were conducted in which the decision tree was trained and tested on the original videos (Experiment 1), applying the pixel-based denoising (Experiment 2) and applying the component-tracking denoising (Experiment 3). Afterwards, the results obtained for the three experiments were compared with the scope of identifying the experiment and the denoising algorithm that led to better classification performances.

Chapter 3

Artifact removal and color score prediction: methods and experimental setup

3.1 Dataset

The data employed in this study derive from the acquisitions performed by the clinicians of two hospital facilities: A. O. Ordine Mauriziano in Turin and Policlinico di Sant'Orsola in Bologna.
The dataset consists of 182 clinical cases, and data was obtained by patients with ovarian adnexal masses enrolled in a clinical study approved by the Ethics Committee, after providing their informed consent (age 18 and 80 years, average = 51 years). I chose the most representative color doppler or power doppler video for each of the selected clinical cases that, in general, included one or more videos representing the flow information. Each video was unpacked into three dimensional frames having color information encoded in R, G, B.
From the selected videos, two datasets were built: one was used to develop the artifact-removal algorithms (pixel-based denoising and component-tracking denoising) and to tune their parameters, while the other dataset was employed to assess the performances of these algorithms based on their capacity to predict the color scores assigned by six expert clinicians.

3.1.1 Dataset to develop the artifact-removal algorithms

The videos that constitute this dataset have been selected in order to balance as much as possible the number of selected data acquired from the two hospitals (A. O. Ordine Mauriziano and Policlinico di Sant'Orsola).

In particular, the dataset is composed of 101 clinical cases, 55 (54.5%) coming from A. O. Ordine Mauriziano hospital and the remaining 46 (45.5%) from Sant'Orsola hospital as shown in figure 3.1.
Among these videos, 63 were acquired using the color doppler, instead 38 videos are power doppler videos (see figure 3.2).

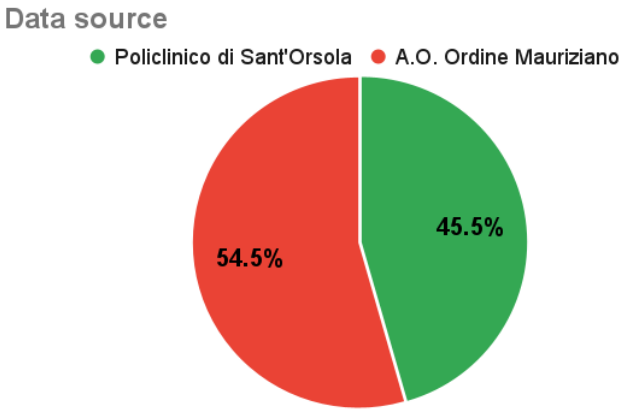


Figure 3.1: The figure shows the percentage of dataset's videos coming from the two hospitals.

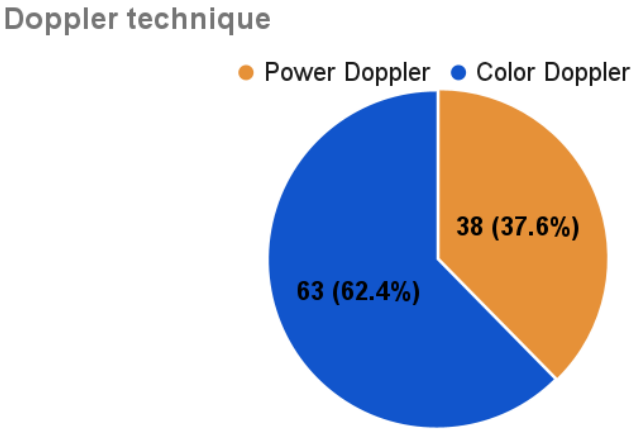
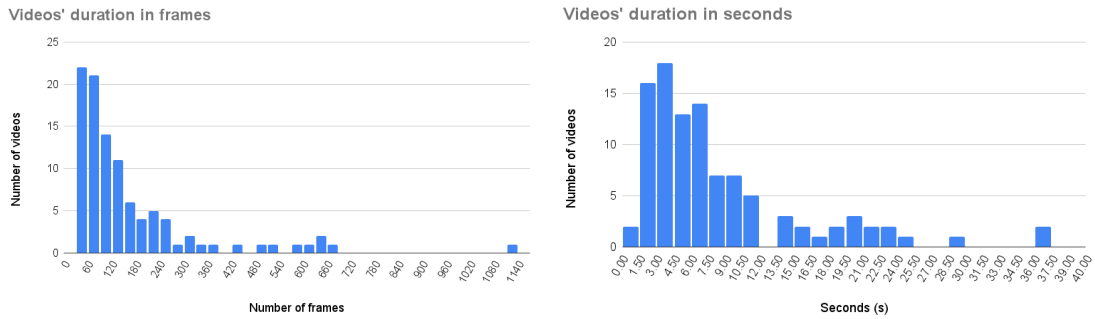


Figure 3.2: Percentages of color doppler and power doppler videos in the dataset.

The videos have an average duration of 8.4s and they were composed of approximately 165 frames on average (see the histograms 3.3a and 3.3b for the duration in frames and seconds of the videos constituting the dataset). Since these videos

belong from acquisitions performed in different hospitals where different ultrasound scanners may be employed, they have different dimensions as shown in the bar plot of figure 3.4.



(a) **(b)**

Figure 3.3: Figure (a) displays the number of videos vs the number of frames in which they were unpacked, whereas the histogram (b) shows the duration in seconds of the dataset's videos.

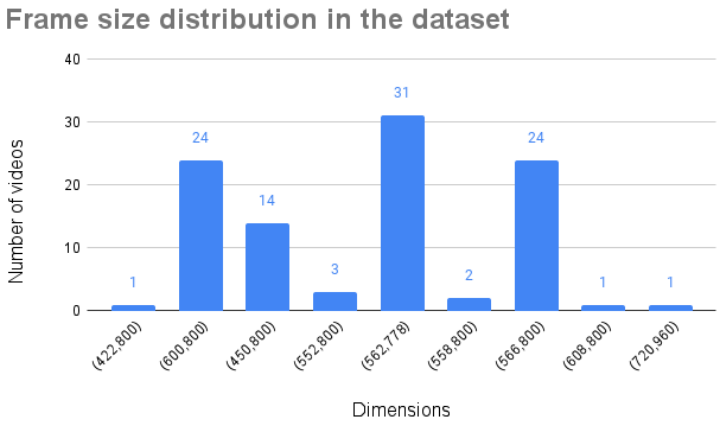


Figure 3.4: The figure shows the size of the frames constituting the dataset's videos. The frame size is defined as (width, height) of the images having color information encoded in R,G,B.

Moreover, both hospitals provided, when available, the hystological reports obtained for each case of ovarian lesions, as illustrated in figure 3.5. For a number of videos (27) there was no hystological report, while in 13 cases the adnexal mass was not surgically removed mainly because the ovarian neoplasm was benign ("no surgery").

Furthermore, the table 3.1 shows the number of tumors that were identified as benign, malignant and borderline (49, 17 and 8 respectively).

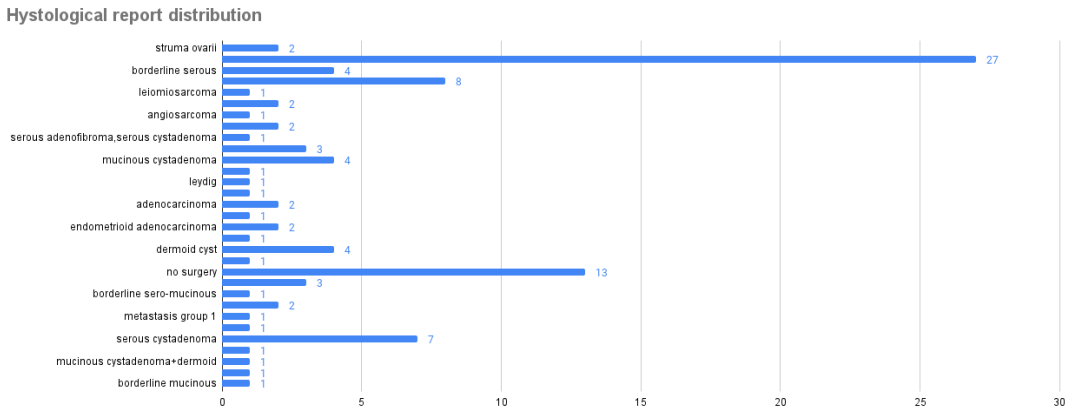


Figure 3.5: The histogram displays the results of the hystological reports for this dataset.

Tumor	Number of cases
Benign	49
Malignant	17
Borderline	8

Table 3.1: The table shows the numerosity of benign, malignant and borderline tumors diagnosed among the 101 videos of the dataset.

Finally, a color score from 1 to 4 was assigned to each video according to the IOTA guidelines by the Syndiag team and then approved by clinicians. The resulting distribution of color score in the dataset is shown in figure 3.6. In particular, there are 39 videos with color score 1, 29 with color score 2, 20 videos having color score equal to 3 and 13 videos to which a color score of 4 was assigned. The fact that the great majority of videos has a color score of 1 or 2 suggests that, in this dataset, malignant ovarian tumors, that normally show higher vascularization (CS = 3 or CS = 4, with CS meaning color score), are less frequent than benign ones that, in contrast, are generally less vascularized; this is coherent with the results shown in table 3.1.

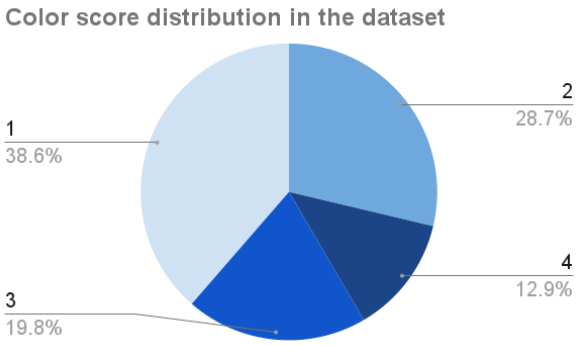


Figure 3.6: The graph shows how the color score is distributed in the dataset.

3.1.2 Dataset for assessing the performances of the artifact-removal algorithms

This dataset containing a total of 106 clinical cases was employed to assess the effect of the artifact-removal algorithms. Specifically, a decision tree was trained to predict the color score assigned by the clinicians to the cases, and the color score prediction was based on the doppler estimation obtained from the number of colored pixels present within the lesion.
The color score was assigned to 54 cases by a panel of six expert clinicians of A. O. Ordine Mauriziano and Policnico di Sant'Orsola, whereas the Syndiag team assigned the color score to the remaining 52 cases with the supervision of clinicians. As illustrated in figure 3.7, the dataset is composed of 90 (84.9%) videos acquired from clinicians at Mauriziano hospital and 16 videos of Sant'Orsola hospital.

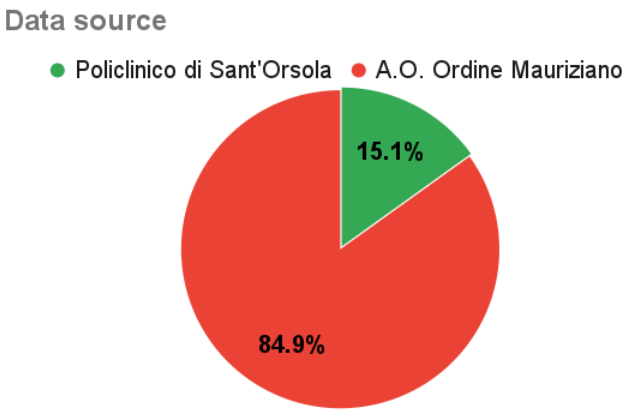
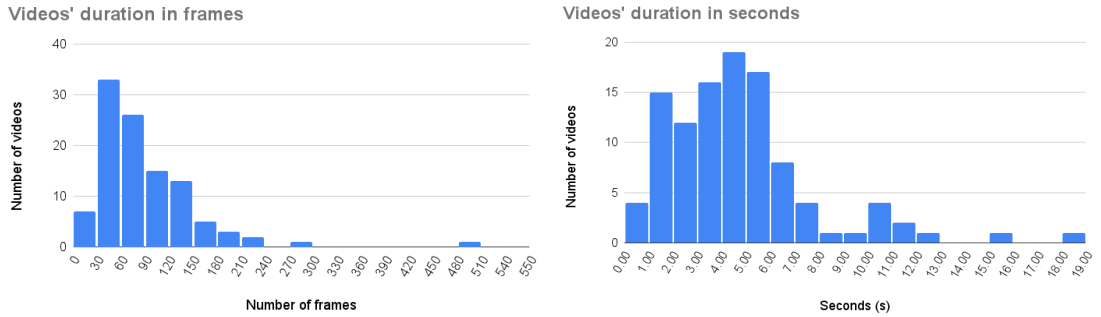


Figure 3.7: The figure shows the percentage of videos in the dataset that has been selected from the two hospitals.

These videos are composed, on average, of 88 three dimensional frames and last 4.7 s (see figures 3.8a and 3.8b for the duration in frames and in seconds of the single videos). Also concerning this dataset, the height and width of the frames can vary among different videos, as shown in figure 3.9.



(a) **(b)**

Figure 3.8: Figure (a) displays the number of videos vs the number of frames in which they were unpacked, whereas the histogram (b) shows the duration in seconds of this dataset's videos.

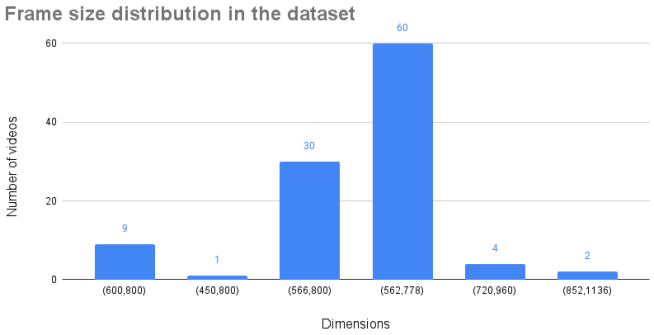


Figure 3.9: The figure shows the size of the frames constituting the 106 videos of the dataset. The frame size is defined as (width, eight) of the 3D images having color information encoded in R,G,B.

Also for this dataset the histological reports obtained for the 106 cases were analyzed and, as seen in figure 3.10, several types of adnexal cancer were identified. The table 3.2 displays the number of benign, malignant and borderline tumors that were diagnosed among the lesions included in this dataset, with benign neoplasms representing around 60% of the lesions.

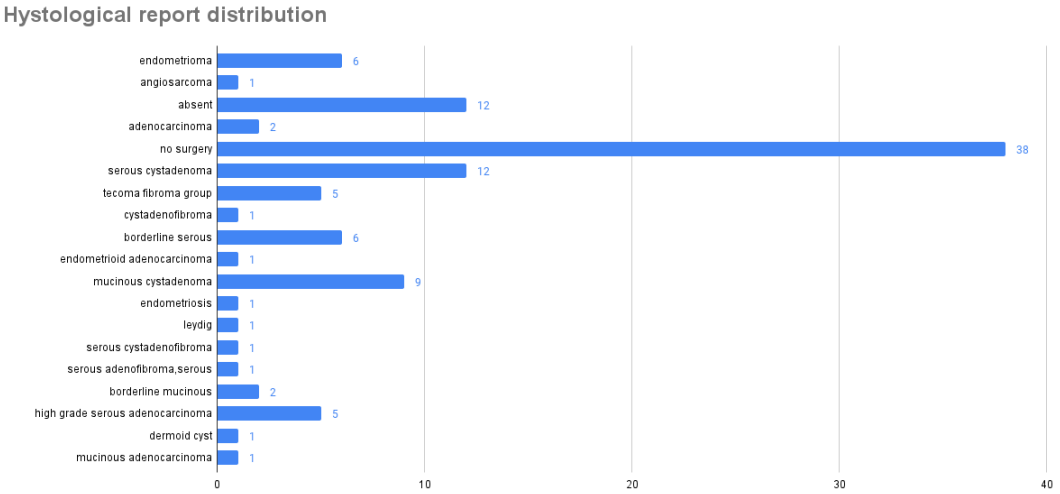


Figure 3.10: The histogram displays the results of the hystological reports for this dataset.

Tumor	Number of cases
Benign	61
Malignant	12
Borderline	11

Table 3.2: The table shows the numerosity of benign, malignant and borderline tumors diagnosed among the 106 videos of the dataset.

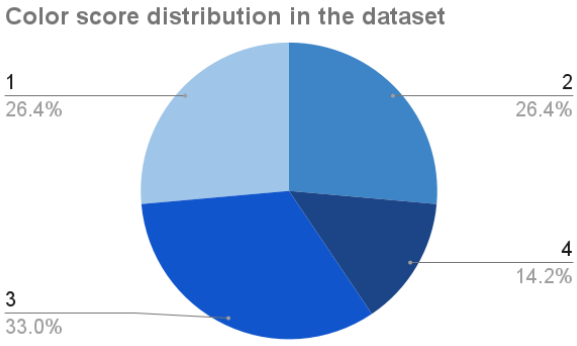


Figure 3.11: The graph shows how the color score is distributed in the dataset.

Moreover, as shown in figure 3.11, the color score is approximately equally distributed, with 28 videos having CS = 1 and CS = 2, and 35 videos with color score 3; while only 15 highly vascularized videos with CS = 4 could be selected.

3.2 Setup

All the codes have been developed using Python 3.7 as programming language and run on Mac OS in a dedicated Anaconda environment (version 4.12, [38]). Among the packages available in Python, the ones used in this study to implement the codes were: numpy [39], pandas [40], matplotlib [41], scikit-learn [42], imageio [43], scikit-image [44], cv2 [45], scipy [46], math [47] and graphviz.
A software platform for annotating medical datasets called Redbrick AI [48] was used to manually label both the adnexal masses and the doppler fan on frames of the videos included in the dataset described in section 3.1.2.

3.3 Pixel-based denoising algorithm

As discussed in the section 1.3.2, Color Doppler and Power Doppler imaging are characterized by a high number of artifacts. The presence of these artifacts may influence the assignment of color score that is a subjective semiquantitative parameter based, according to IOTA guidelines, on the amount of blood flow within the lesion [4].
In order to avoid wrong interpretations of the flow information due to artifacts that may be confused with the real signal, the Syndiag team has developed an algorithm, called pixel-based denoising algorithm, whose aim is removing artifacts from the doppler signal so that only the real blood flow remains (a more detailed description of this algorithm can be found in the master thesis of F. De Simone, [49]).
The denoising algorithm differentiates between real doppler signal and noise on the basis of the temporal persistence of color signal throughout the whole video, pixel by pixel. In particular, it is assumed that artifacts are not persistent, therefore the pixels that remain colored for a sufficient number of consecutive frames of the considered video are treated as real signal, the others are flagged as artifacts.
The flow chart of figure 3.12 shows the main steps of the algorithm pipeline.
The algorithm receives as input the frames, conveniently cropped, in which the doppler video was unpacked. Then, for each frame, a 3D binary mask where all colored pixels are set to 1 is generated. The temporal persistence is evaluated by calculating the lengths of all doppler activation sequences for obtaining the number of consecutive frames in which the investigated activation is persistent. A chosen threshold is equal to 90th percentile of the distribution of all activation

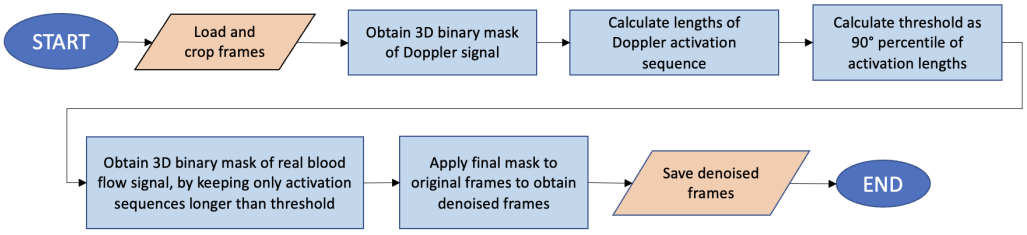


Figure 3.12: Flow chart of pixel-based denoising algorithm [49].

lengths of all the pixels. Hence if the length of doppler activations is larger than the threshold they are considered as real signal, otherwise they are seen as artifacts. The pixels constituting the artifacts are set to 0 in the resulting blood flow mask where, consequently, only the real blood flow signal remains. Eventually, the final mask is applied to the original noisy frames of the video to obtain the denoised frames that are finally saved.
The figure 3.13 illustrates an example of application of the denoising algorithm to a US-Doppler video. One frame and the corresponding denoised frame are represented and the successful suppression of the flash artifact can be noticed.

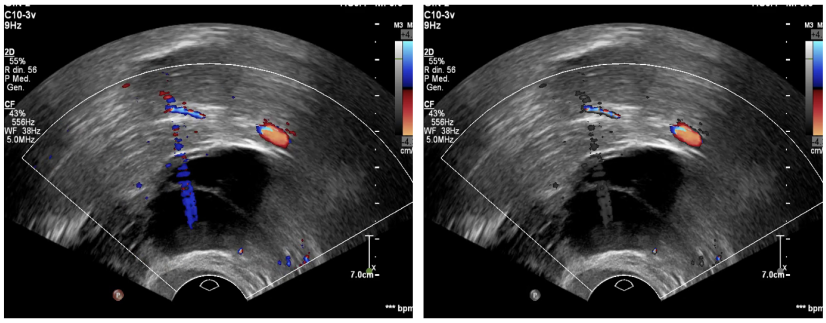


Figure 3.13: The image at the left represents the original frame, input of the denoising algorithm, meanwhile the right image shows the denoised frame resulting from the algorithm application where the flash artifact is removed and the real blood flow signal remains.

3.4 Component tracking algorithm

The main limitation of the denoising algorithm is that, to remove doppler artifacts, it only relies on the exact pixel correspondence in a frame sequence, taking into account only the temporal persistence of doppler activations during the whole video without considering the spatial persistence.

This aspect becomes a problem because, due to the movement of the ultrasound probe during the acquisitions, the investigated region is not static throughout the video, and individual pixels do not exactly correspond to fixed anatomical regions across the whole video, thus the spatial correspondence is lost. For the same reason, individual activation areas of doppler signal also shift through frames, following the investigated region movements. As a consequence, many real activations are flagged as artifacts and, thus, suppressed.
To avoid this problem, a component tracking algorithm has been implemented to be incorporated in the denoising algorithm.
This algorithm identifies every connected component as a doppler activation in the first frame of the considered video. It then associates a unique fixed identity to each activation, and is able to track it, i.e. to recognize the same doppler activation areas in the next frames even if they moved, changing their position. In this way, it is possible to track the different clusters of activation and follow them while they move by keeping track of the pixel coordinates occupied by each connected component – meaning each closed group of colored pixels that constitute a doppler activation – at each frame. By performing this tracking in combination with the denoising algorithm, both spatial and temporal persistence of signal activations during the whole video are considered.
The algorithm is characterized by some parameters:

- the distance between centroids. The centroid of each connected component is obtained for each frame. For each object (i.e., connected component) in a frame, the distance between its centroid and the centroids of all possible objects in the previous frame is calculated. The minimum computed distance is used to identify two objects in two consecutive frames as having the same identity.

- the overlap between connected components of consecutive frames. The overlap is calculated as the number of pixels of the intersection between an object and the candidate corresponding objects in the previous frame, divided by the number of pixels of the new object, as shown in equation 3.1

$$overlap = \frac{obj1 \cap obj2}{size2} \tag{3.1}$$

  where obj1 and obj2 represent the object of the previous frame and that of the current frame, respectively; whereas, size2 is the number of pixels of the object of the current frame.

- size similarity. In the case of multiple correspondences between objects in following frames, the pair of corresponding objects is chosen as the one that

maximizes size similarity. Size similarity is calculated as:

$$size\_similarity = \frac{|size1 - size2|}{size2} \tag{3.2}$$

where size1 corresponds to the number of pixels of the smaller object and size2 is the number of pixels of the larger object.

In the next section, the pipeline of the component tracking algorithm is described in details.

3.4.1 Algorithm Pipeline

1. The inputs of the algorithm are:
   - A list of binary masks, one for each frame of the video, where the colored doppler pixels are set to 1, while the rest are set to 0.
   - maxDist, a threshold corresponding to the maximum accepted distance, in pixels, between centroids of the same object in consecutive frames.
   - maxFrames, the maximum accepted number of frames where an object is still tracked, even though it is not visible.
   - minOverlap that corresponds to the minimum accepted value of overlap for which two objects are considered correspondent.
2. All the objects in the first mask are initialized and registered as new objects.
3. For each following mask:
   (a) Disappeared objects, namely objects that are marked as non-tracked, are updated.
   (b) It is checked if the mask is empty (so if all the pixels in the mask are set to 0 and no objects are present). If so, the frame is registered as empty and there is a check if there are disappeared objects.
   (c) The properties of all objects present in the mask are obtained. These properties are: coordinates of the object, rectangular bounding box extrema and centroid coordinates.
   (d) The overlap between all possible pairs of objects in current mask and tracked objects is calculated (this calculation is described in more details by the flow chart of figure 3.15).

(e) The distance between centroids of all possible pairs of objects in current mask and tracked objects is calculated (details of this calculation are included in figure 3.16). For each object pair:

- If overlap $\geq$ minOverlap OR distance $\leq$ maxDist, there is correspondence between the two objects which is registered.
- If overlap $<$ minOverlap OR distance $>$ maxDist, the absence of correspondence between the two objects is registered.

(f) Every new object can correspond to at most one tracked object. In the case of multiple correspondences between one object and objects in the previous frame, size similarity is considered: the correspondence having the highest value of size similarity is selected as the final correspondence.

(g) All the tracked objects for which no correspondence has been found are marked as unseen.

(h) A control to check the presence of disappeared objects is performed.

(i) For each tracked object for which at least one correspondence has been found:

- If the same tracked object corresponds to more than one new object, the binary mask of the merged new object is created, and the merged properties are obtained. The tracked object is updated with the properties of the merged object.
- If the tracked object has one correspondence, it is updated with the properties of the new object.

(j) Finally, all the new objects with no correspondences are registered as new.

4. The algorithm gives as output a list containing a number of collections equal to the number of objects that have been identified. Each collection corresponds to an object and contains the following fiels:

- The unique id assigned to each object in order of appearance in the video.
- The list of coordinates occupied by the object at each frame.
- A list whose elements, one corresponding to each frame of the video, are set to 1 if the object is identified in the corresponding frame, 0 otherwise.

The flow chart of figure 3.14 shows the pipeline of the component tracking algorithm.

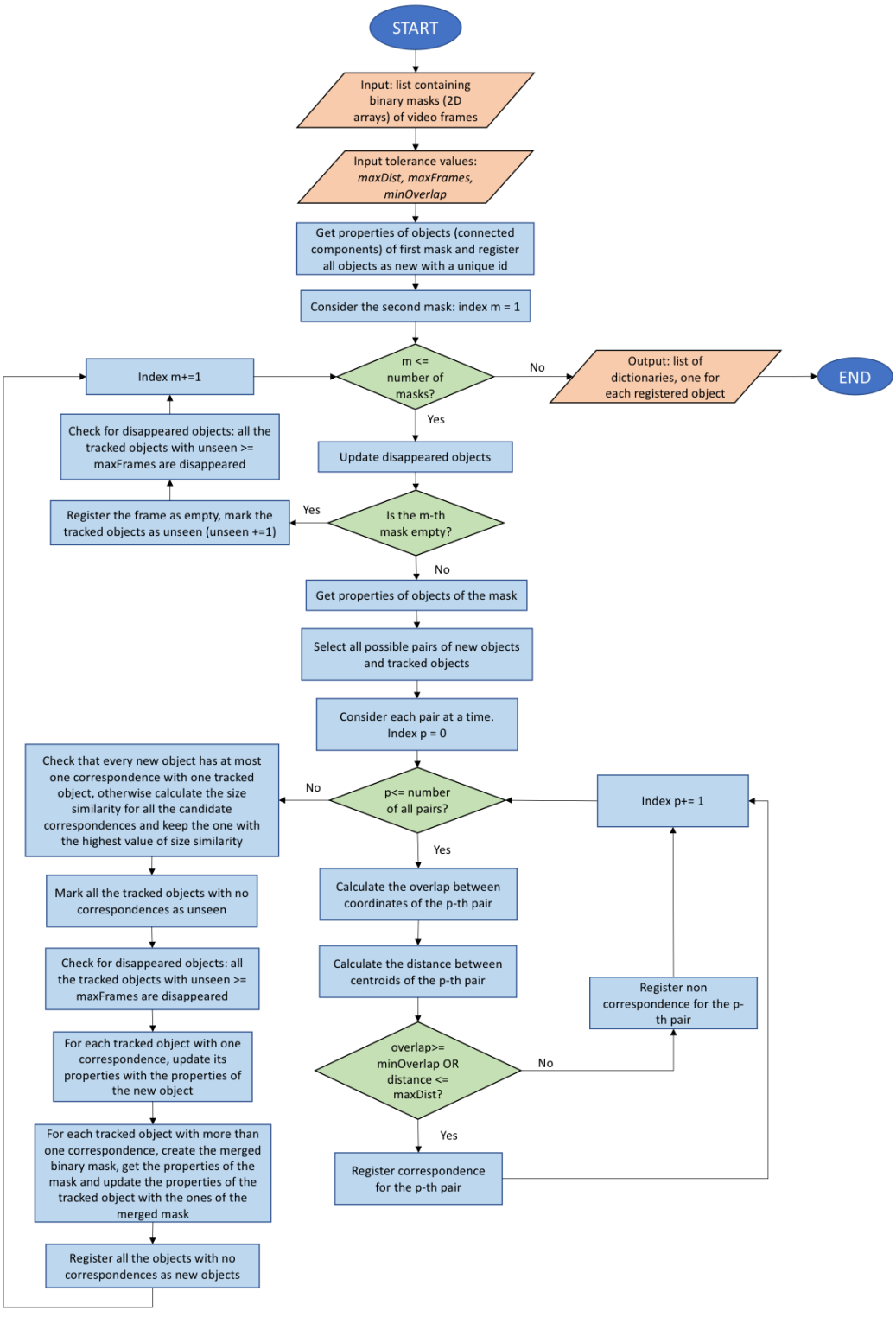


Figure 3.14: Flow chart of component tracking algorithm.

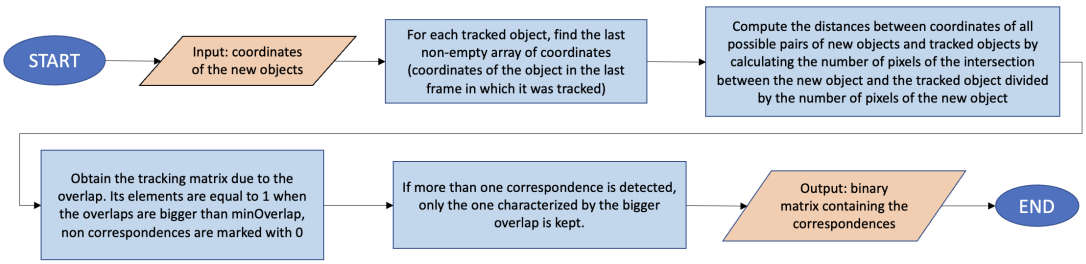


Figure 3.15: Flow chart of the function that calculates the overlap between coordinates of new objects and tracked objects.

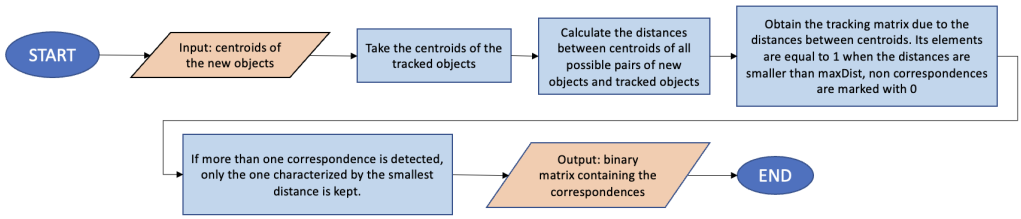


Figure 3.16: Flow chart of the function that calculates the distance between centroids of new objects and tracked objects.

The output of the algorithm is then displayed as images, one for each frame of the video of interest, that contain the identified objects, each represented with a different color that remains the same throughout the video. In this way, it is possible to keep track of the objects intuitively (see figures 3.17e, 3.17f, 3.17g and 3.17h for an example of plot of the tracking output considering four consecutive frames).
Finally, the tracking algorithm is then integrated with the denoising algorithm illustrated in section 3.3 resulting in the connected components-based denoising with component-tracking algorithm (in the following it will be also referred to as denoising with tracking algorithm or component-tracking denoising). In this case, the tracked components are the input of the denoising algorithm. The temporal persistence is evaluated by computing the lengths of doppler activation sequences for each object identified by the tracking algorithm, i.e. the activation length for each component is calculated as the number of frames in which that object is tracked. The activations for each object are then used to describe a distribution and the threshold to discriminate between real activations and artifacts is calculated as the 98th percentile of such distribution.
The figure 3.17 displays an example of the application of the component-tracking denoising algorithm, where, for the same four consecutive frames of the same video,

the frames obtained from the original video (i.e., video on which no algorithm to suppress artifacts is applied), the outputs of component-tracking algorithm and the outputs of the denoising with tracking algorithm are shown. From the figure, it can be noticed that the component-tracking denoising correctly keeps track of the real activations associated to the connected components identified by the tracking algorithm, while removing artifacts (as the one suppressed in the figure 3.17j).

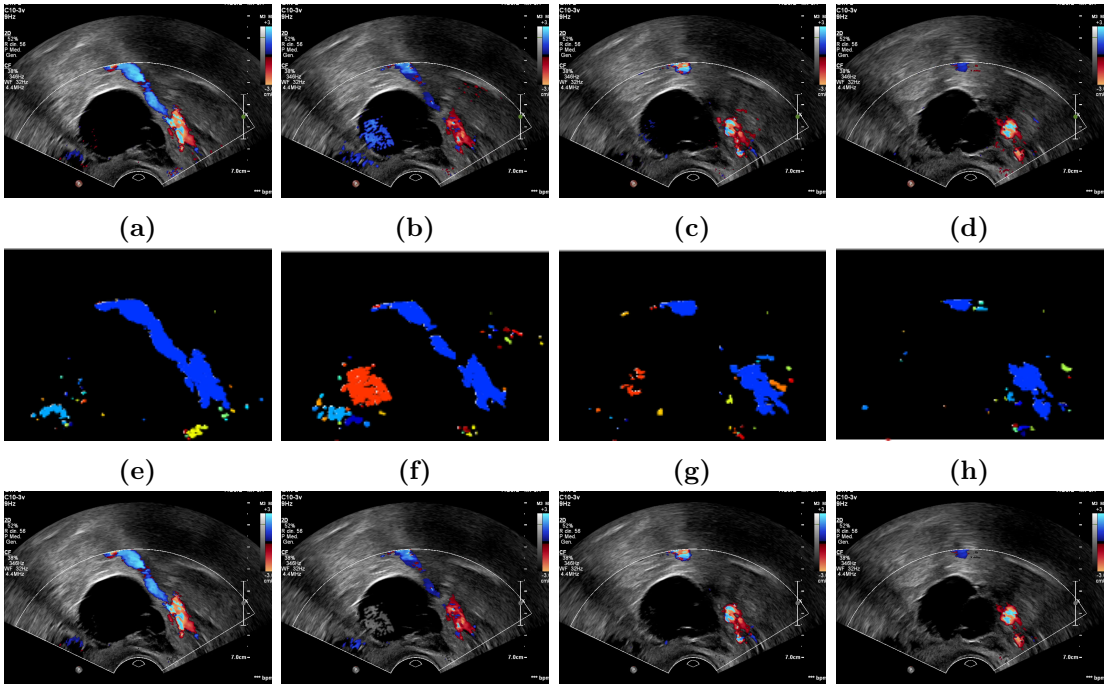


Figure 3.17: The figure shows an example of component-tracking denoising application. The images (a), (b), (c) and (d) in the first row correspond to four consecutive original frames of the same video. In the second row there are the outputs of the component-tracking showing the connected components identified for each frame, each represented with a different color that remains the same throughout the video. The third row of images consists of the corresponding denoised frames, output of the component-tracking denoising. In frame (b) an artifact appears, it is identified by the red connected component of figure (f), and it is correctly suppressed in the denoised frame (j). When the artifact disappears in frame (d), also the red object is not present anymore in the corresponding tracking output (h). The blue object identified in the images (e), (f), (g) and (h) is associated to a real activation that is correctly identified and tracked by the denoising algorithm in frames (i), (j), (k) and (l).

3.5 Algorithm's generalization and harmonization of the input data

3.5.1 Qualitative assessment of algorithms' outputs

In the first stage of the study, it was qualitatively evaluated the capacity of the component-tracking denoising to suppress artifacts. Therefore, a preliminary analysis was performed by applying the pixel-based denoising and the component-tracking denoising to the 101 videos included in the dataset illustrated in section 3.1.1. The outputs of these two algorithms were qualitatively evaluated. The assessment was performed considering two criteria: the presence of artifacts *survived* to the algorithms and the suppression of real blood flow inside and outside of the adnexal mass.
In particular, if the result of the algorithms' application led to a case where all the artifacts were removed while preserving the real doppler activation, it was positively evaluated as *good*, otherwise it was evaluated as *bad*.
As shown in figure 3.18, applying pixel-based denoising more than half dataset was positively evaluated, while the outputs of denoising with tracking algorithm resulted negative in almost 80% of cases.

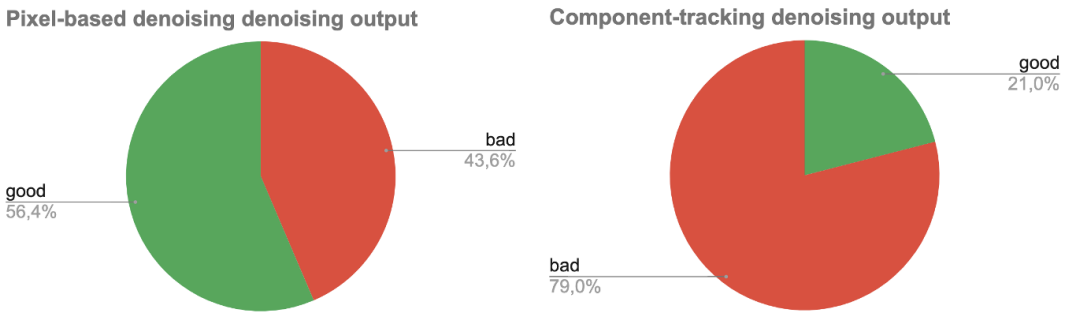


Figure 3.18: The pie chart at the left shows the evaluation of the pixel-based denoising outputs, while the right chart illustrates the evaluation of the component-tracking denoising outputs.

An analysis of the most frequent problems related to the denoising with tracking algorithm was conducted on the cases with negative evaluation.
From this analysis, it resulted that in more than 50% of the cases the component-tracking denoising did not perform well because there were artifacts overlapped to real doppler signal. When the connected components of the two overlapped, the algorithm wrongly recognized them as a single connected component (as seen in the example shown in figure 3.20).

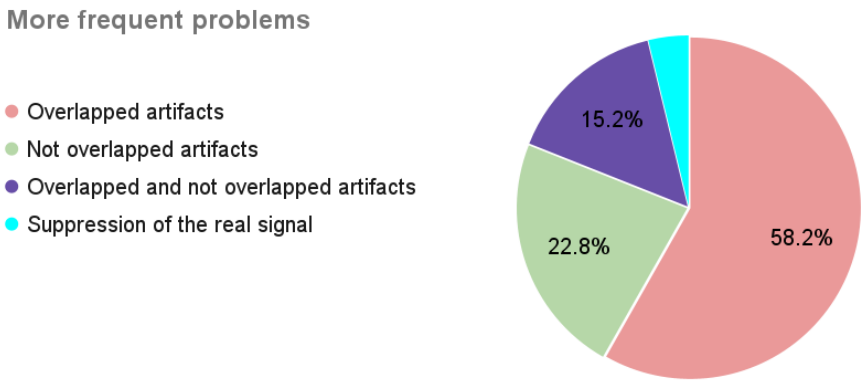
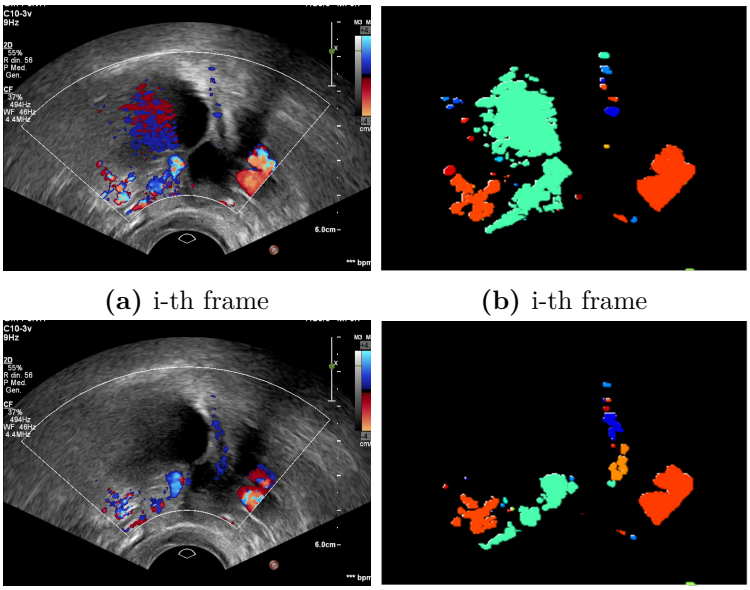


Figure 3.19: The figure displays the most frequent problems that led to negatively evaluated outputs of the component-tracking denoising algorithm.



(a) i-th frame **(b)** i-th frame

(c) (i+1)-th frame **(d)** (i+1)-th frame

Figure 3.20: The figure shows two consecutive original frames - (a) and (c) - and the corresponding outputs of the tracking algorithm depicted in (b) and (d). Note the presence of the artifact within the adnexal mass at the i-th frame, it overlaps with the real doppler signal that persists in the successive frame and they are identified by the same connected component in aqua.

Besides the artifacts that resulted overlapped with the real activations, there

were also artifacts *survived* to the component-tracking denoising that were not overlapped to the real doppler signal. These artifacts, in the following referred to as "not overlapped artifacts", were found in 18 videos (22.8%); whereas in 12 cases (15.2%) both types of artifacts were identified. The remaining 3 videos (3.8%) were characterized by the suppression of real doppler activation.
Moreover, it was also proved that the suppression of real doppler signal and the presence of the not overlapped artifacts occurred also when the pixel-based denoising algorithm was applied (as seen in figure 3.21), thus they are not characteristic of the component-tracking denoising.
In this context, the study proceeded focusing on the overlapping problem and on other edge cases (related to the component-tracking denoising or to the input data) that were identified during the development of the component-tracking denoising, in order to improve the performances of the denoising with tracking algorithm in terms of correctly tracking the real signal and removing artifacts. In particular, the main improvements that were implemented are:

- Exclusion of the pixels forming the fan outline that, being colored in some videos, was erroneously tracked by the component-tracking algorithm (see section 3.5.2).
- Implementation of a function to be integrated in the component-tracking denoising in order to manage the problem of large artifacts overlapping with real activations (see section 3.5.3).
- Managing the presence of empty binary masks where no object is identified by the tracking algorithm (see section 3.5.4).
- Analysis of several parameters of size and shape of the connected components to distinguish between real activations and artifacts (see section 3.5.5).

Afterwards, the component-tracking algorithm was tuned as illustrated in section 3.6.

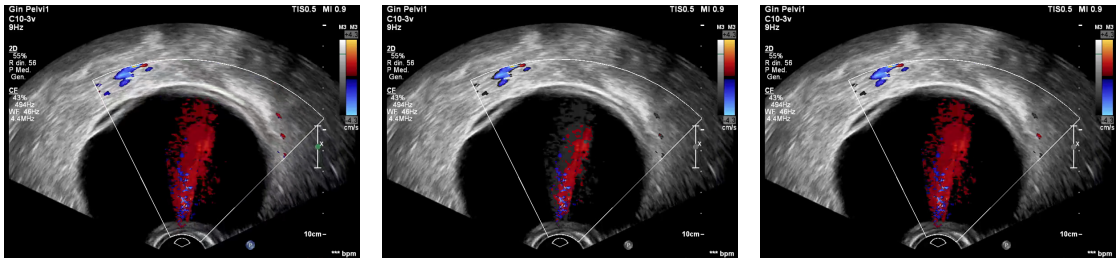


Figure 3.21: Example of flash artifact not overlapped to the real signal. From left to right: original frame, output of the pixel-based denoising, output of denoising with tracking algorithm. Note that the artifact is present in the three frames, despite the application of the denoising algorithms in the second and third image.

3.5.2 Debugging of videos with colored doppler fan

Some videos in some clinical cases were characterized by having a colored outline for the doppler fan (the area of the ultrasound image in which the doppler signal is collected). This caused the pixels in the outline to be included in the connected component analysis. In most cases this resulted in an identified connected component with the size of the whole doppler fan.
This issue is graphically represented in the figure 3.22 that shows a frame of the video taken as example, the corresponding binary mask obtained by setting to 1 the colored pixels, and the resulting output of the component tracking algorithm where the connected component associated to the colored fan overlaps with some doppler activations that, instead of being recognized as objects different from each others and from the fan, are associated to the connected component of the fan (indeed, they are represented with the same color of the fan).
This led to a wrong tracking of the doppler signal throughout the video and, therefore, also to errors in removing artifacts when the tracking algorithm was applied in combination with the denoising algorithm.

Figure 3.22: The figure shows, from left to right, the original frame, the corresponding binary mask where all the colored pixels are set to 1, and the corresponding output of the component tracking algorithm where there is the connected component associated to the doppler fan that overlaps with the doppler activations touching the fan.

Therefore, in order to fix this bug, it was implemented a method aimed at excluding the pixels forming the fan outline. The basic idea of this method is to apply the morphological operation of erosion to erode the colored pixels of the fan.
The pipeline of the implemented method is the following:

1. For each frame of the video of interest:
   (a) Starting from the original not denoised frame (see figure 3.23a), the binary mask where all colored pixels are set to 1 is obtained (see figure 3.23b).
   (b) The holes potentially present in the mask are filled by means of the function binary_fill_holes from the package scipy [46] (see figure 3.23c).

2. The masks resulting from point 1.b, obtained for all the frames of the video, are summed up to obtain a mask in which, afterwards, the pixels having intensity higher than 0 are set to 1, otherwise they maintain the 0 value. This way, the mask of the doppler fan is obtained (see figure 3.23d).

3. The morphological operation of erosion is applied with a 9x9 kernel to the mask of step 2 with the aim of deleting the colored pixels of the fan. In this way, in the resulting mask only the pixels within the fan are set to 1.

4. The mask is multiplied to each frame of the video to obtain the *eroded* frames (see figure 3.23e).

5. For each eroded frame resulting from step 4, the new binary mask of colored pixels is obtained. In this new mask the pixels of the doppler fan are set to zero (see figure 3.23f).

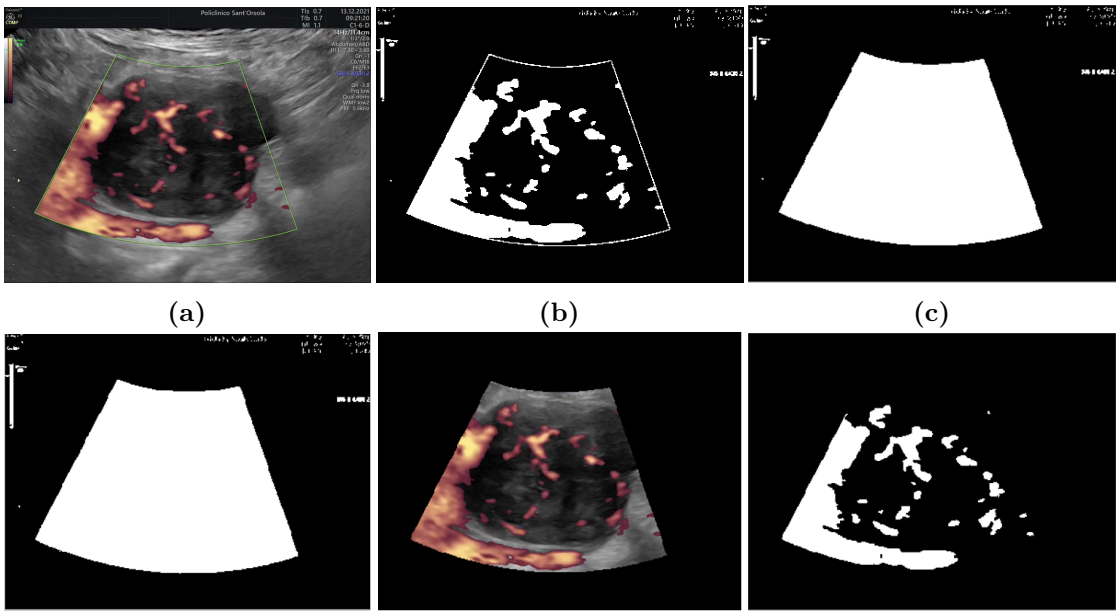


Figure 3.23: Step by step outputs of the pipeline of the method that eliminates the colored pixels of the doppler fan. (a) Original frame. (b) Doppler mask, depicting all pixels containing Doppler signal. (c) Mask where the holes are filled. (d) Binary mask resulting from the sum of the masks where the holes are filled for all the frames of the video. (e) Eroded frame, obtained by applying the sum of the filled masks to the original frame. (f) New doppler mask where the pixels of the fan are disappeared. Note that (c) and (d) are the same because in the considered frame the doppler fan was a closed polygon, but this is not necessarily the case.

The illustrated method was developed considering a small number of videos that were included in the dataset described in section 3.1.1, then its application was extended to the whole dataset.
Results were robust with videos having a grayscale doppler outline: the resulting binary masks remained unchanged with respect to the binary masks of colored pixels. On the other hand, regarding the 46 videos that presented the problem, in 39 the color was correctly eliminated from the doppler fan, while in the remaining 7 videos some pixels constituting the fan were still colored (see figure 3.24).

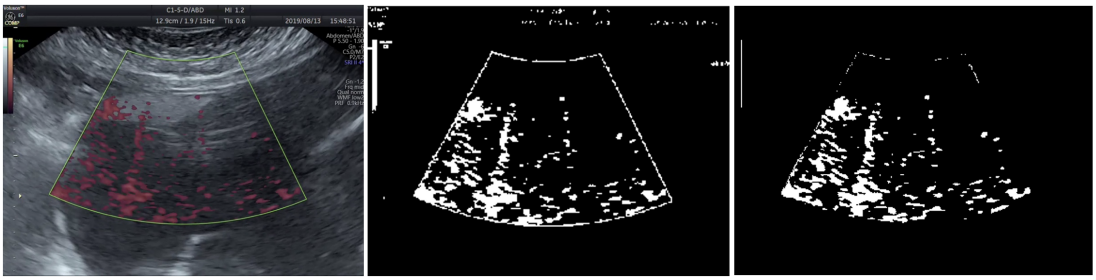


Figure 3.24: The figure shows, from left to right, the original frame, the corresponding binary mask containing all the colored pixels, and the corresponding doppler mask obtained after the application of the described method that, in this case, is not able to remove all the colored pixels constituting the doppler fan.

In order to fix the bug for all the videos, it was tried firstly to remove the small objects that were placed close to the doppler fan, considering different minimum dimensions of the objects (64, 128, 256 and 512 pixels). However, this approach did not solve the issue and the few colored pixels survived.
From the results obtained after the application of the methods that were implemented to try to fix this issue for all the videos, it emerges that these data have intrinsic complexities that may be related to the ultrasound instrument employed to perform the acquisitions and its settings. Moreover, further tests and trials are required to explore this feature more deeply.
The figure 3.25 shows an example of output of the component tracking algorithm run using the *new* mask in comparison with the tracking output obtained without applying the illustrated method, considering the same frame.
In addition, it has to be taken into account that, for the 7 videos where the doppler fan was still partially visible, the connected components associated to the remaining colored fan did not overlap with the surrounding doppler activations that were identified by different objects since the clusters of remaining colored pixels were spatially distant (see figure 3.26).

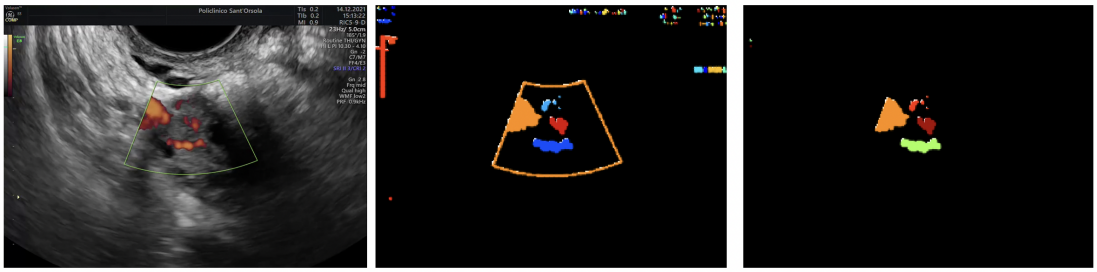


Figure 3.25: Original frame, output of the tracking algorithm where the colored pixels of the fan are considered as a connected component, tracking output without the presence of the object associated to the fan, considering the same frame. In this example, the described method worked well.

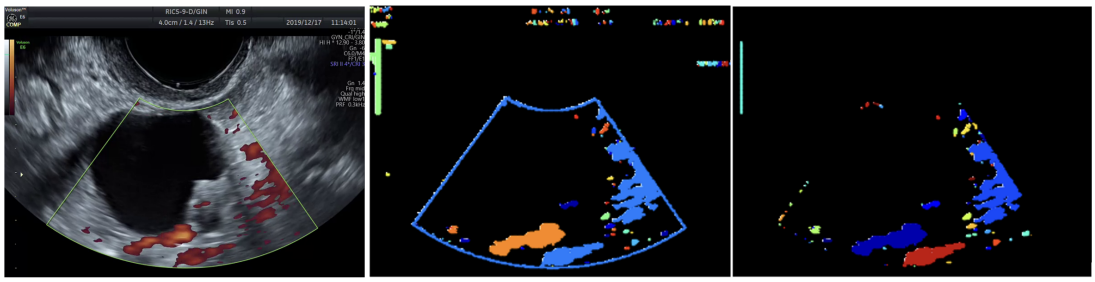
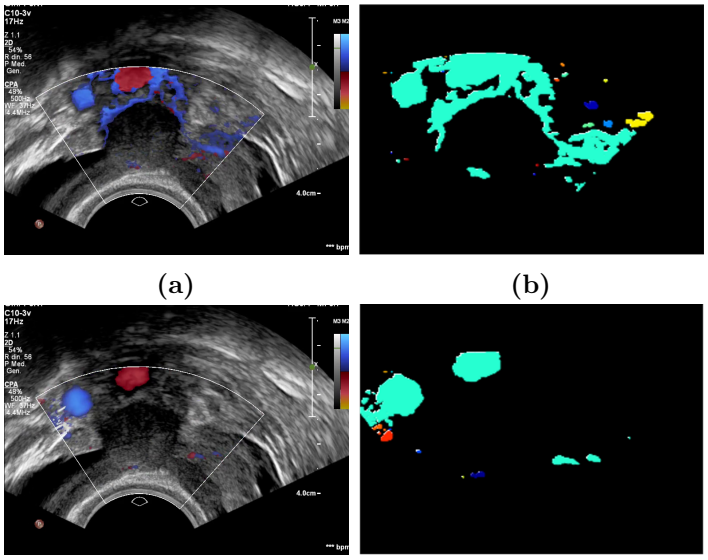


Figure 3.26: The figure shows, from left to right, the original frame, the corresponding tracking output that identifies an object in light blue associated to the fan, and the corresponding tracking output obtained giving as input to the algorithm the masks resulting from the application of the method. Even if in the third image there are few remaining pixels of the fan, the different doppler activations that touch the fan are identified as different objects, instead of what happens in the second image where they are identified as part of the same connected component in light blue.

3.5.3 Integration of the minimum distance function in the tracking algorithm

Afterwards, the component-tracking algorithm was further improved by managing an edge case that occurred in videos characterized by the presence of an artifact (for instance a flash artifact) of big dimensions that persists for several consecutive frames. Large artifacts tend to overlap to different doppler activations which can be quite distant between each other and, when the artifact disappears, they are associated to the same connected component (the one of the artifact) even if they should be identified by separate objects.

The figure 3.27 displays an example of this problem, where the large artifact present in the original frame 3.27a is identified by the aqua connected component and it is overlapped with the doppler activations that, when the artifact disappears in frame 3.27c, are still associated to the aqua object as shown in figure 3.27d.
This problem could lead to errors in keeping track of the doppler activations throughout the video.



(a) **(b)**

(c) **(d)**

Figure 3.27: The figure shows two original frames of the same doppler video - (a) and (c) - and the corresponding outputs of the tracking algorithm depicted in (b) and (d). Note the presence of the large artifact in the frame (a), it covers several doppler activations that, when the artifact disappears, are identified with the same connected component in aqua as seen in figure (d), despite they are distant from each other and not artifacts.

On the other hand, also the opposite situation could occur. In these cases, referred here as "split-merge problem", a doppler activation that appears at one frame splits into two (or more) connected components in the following frames. These new connected components should not be treated by the algorithm as different objects, but recognized as parts of the same object.
To account for the overlapping artifacts and the split-merge problem, I calculated the distances between all the possible pairs of objects in the same frame for which the algorithm has found a correspondence with the object of the previous frame. If the minimum distance between one object and the others is lower than a chosen threshold, the object is recognized as the same object of the previous frame; otherwise the object is tracked as a different new object.

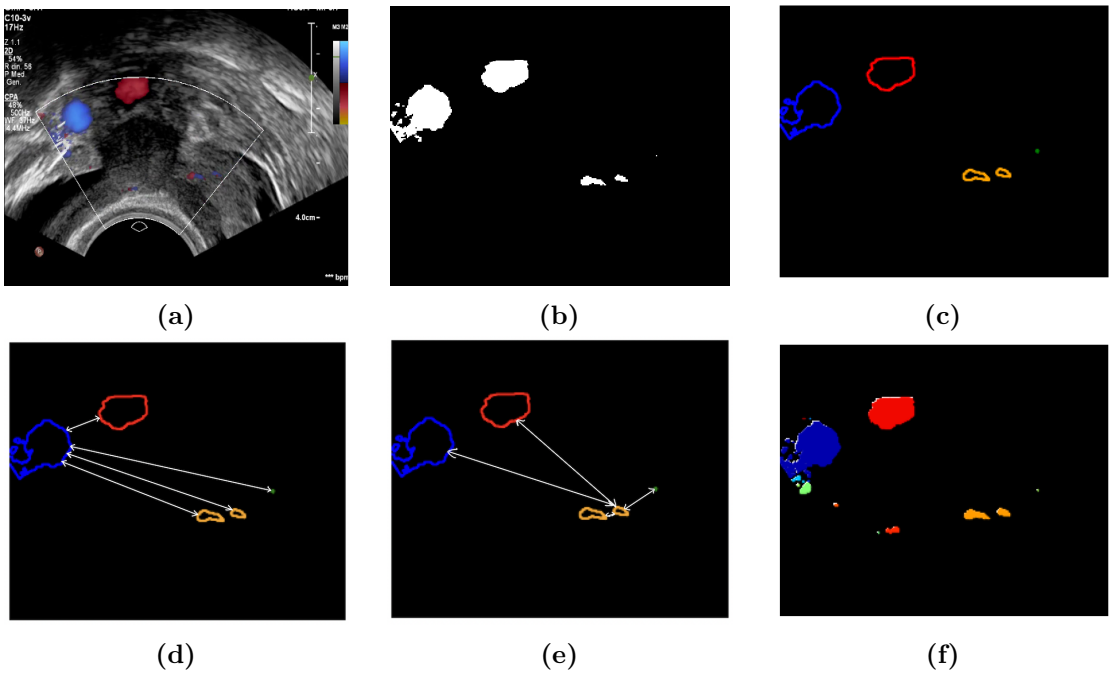


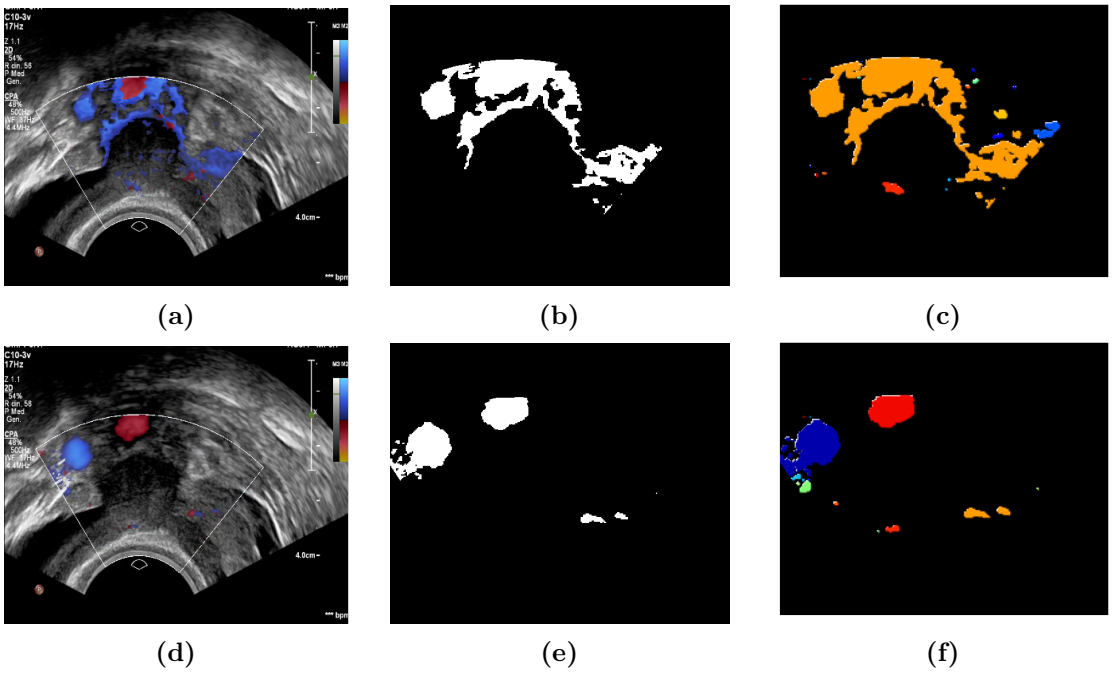
Figure 3.28: Step by step outputs of the function that calculates the distance between contours of objects in the same frame. (a) Original frame. (b) Binary mask containing the objects identified in frame (a), for which the correspondence with the artifact (present in the previous frame, shown in figure 3.27a) occurred. (c) Mask depicting the contours of these objects. (d) Calculation of the minimum distance between the contour of the blue object and the one of all the others. Since the minimum distance (the one from the red object) is larger than minDist, the blue object is tracked as a new object. (e) Calculation of the minimum distance between the contour of the smaller orange object and all the others. Since the minimum distance (the one between the two orange objects) is smaller than minDist, the two orange objects correspond to the connected component associated to the artifact. (f) Tracking output resulting from the function's application: the objects that are distant from each other are correctly identified by new connected components different from each other and from the artifact.

The pipeline of the implemented function is the following:

1. The input to the function consists in the list of the objects for which the correspondence with the object tracked at the previous frame occurred (i.e. a large overlapping artifact or a real activation that splits into these objects) and their coordinates.

2. The binary mask containing these objects is generated, in this mask the pixels identified by the coordinates of the objects are set to 1 (see figure 3.28b).

3. For each connected component, the contours are obtained with the function find_contours from the Python's package scikit-image [44] (see figure 3.28c for the mask containing the contours of the objects).
4. The distances between all possible pairs of the contour coordinates of all the pairs of objects are computed and the minimum distance between each object and the others is taken (see figures 3.28d and 3.28e).
5. Given a threshold minDist, chosen after tuning the algorithm:
   - If the minimum distance between the object and all the others $\leq$ minDist, this object corresponds to the tracked object, thus it can be merged with the others that correspond to the old object.
   - If the minimum distance between the object and all the others $>$ minDist, the object is registered as a new object.

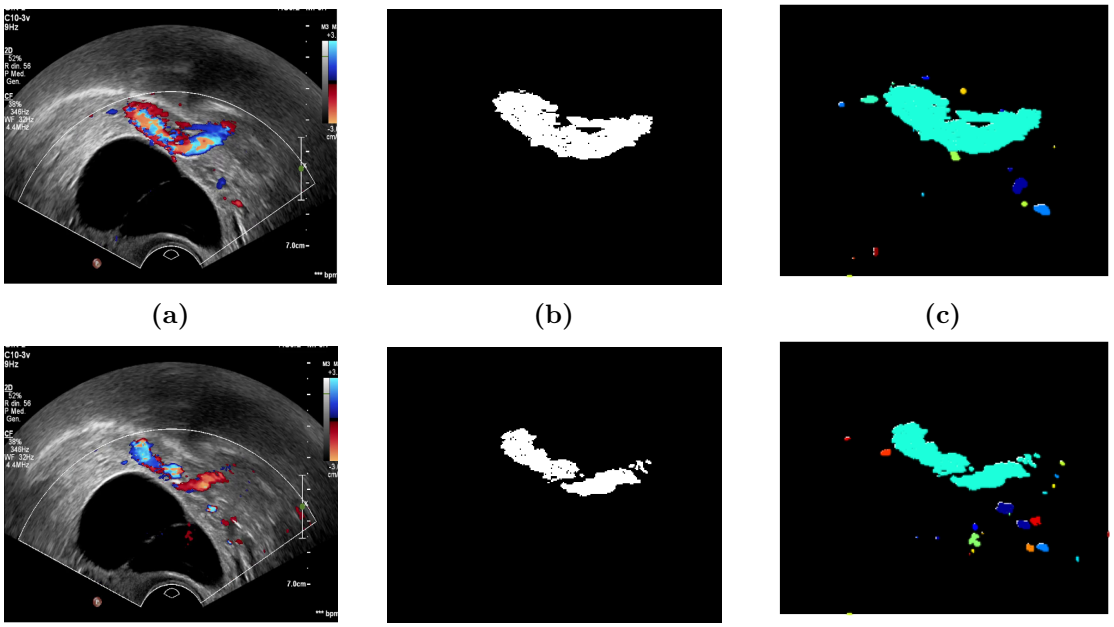


(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure 3.29: Example of application of the function that calculates the distance between contours of objects in the same frame. (a) Original frame where the large artifact is present. (b) Binary mask of the connected component associated to the artifact. (c) Tracking output where the artifact is represented by the orange object. (d) Original frame where the artifact is gone. (e) Binary mask containing the objects for which the correspondence with the artifact occurred. (f) Tracking output resulting from the function's application: the objects that are distant from each other are identified by new connected components different from the artifact.

Applying this function, when a large artifact that is persistent overlaps with several doppler activations that are distant between each other, they are correctly identified as different objects (see figure 3.29f). At the same time, the function ensures that if the resulting connected components are part of the same doppler signal they are still recognized as the same object (see figure 3.30).



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure 3.30: The figure shows an example where, considering two consecutive frames (a) and (d), the doppler activation identified by the aqua connected component in frame (c) splits into different objects that are correctly associated to the same connected component of the activation by the algorithm, thus they are both aqua, as depicted in figure (f). (a) Original frame. (b) Binary mask of the connected component of interest. (c) Tracking output at the frame (a). (d) Original frame. (e) Binary mask of the same connected component identified at the frame (d). (f) Tracking output at the frame (d) where the objects in which the activation split are recognized as part of the same connected component.

The flow chart of the described function is shown in figure 3.31.

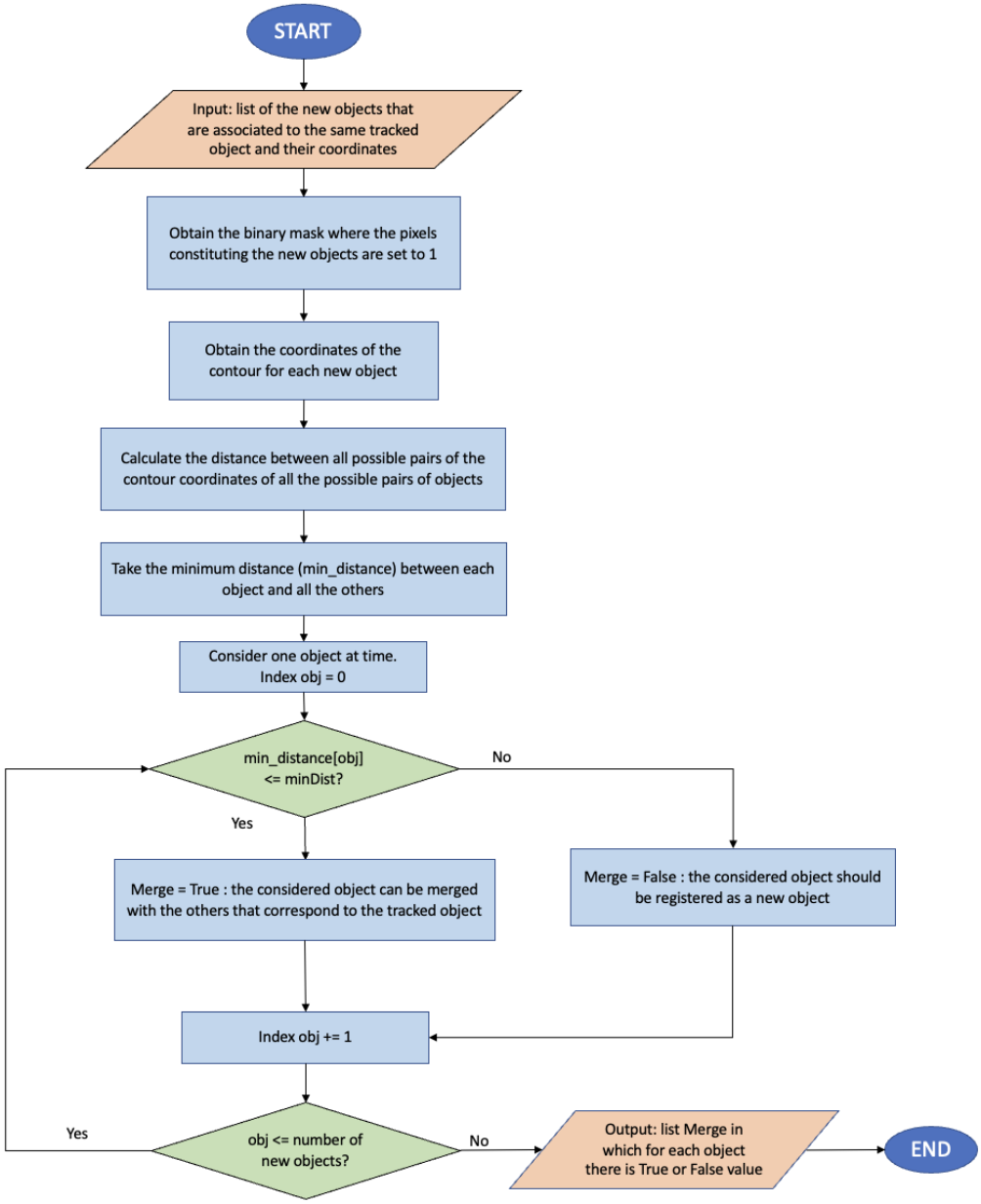


Figure 3.31: Flow chart of the function that calculates the minimum distance between the objects of the same frame, when these objects correspond to the same tracked object.

After the integration of this function in the tracking algorithm, the resulting pipeline is the one illustrated in figure 3.32 in the next page.

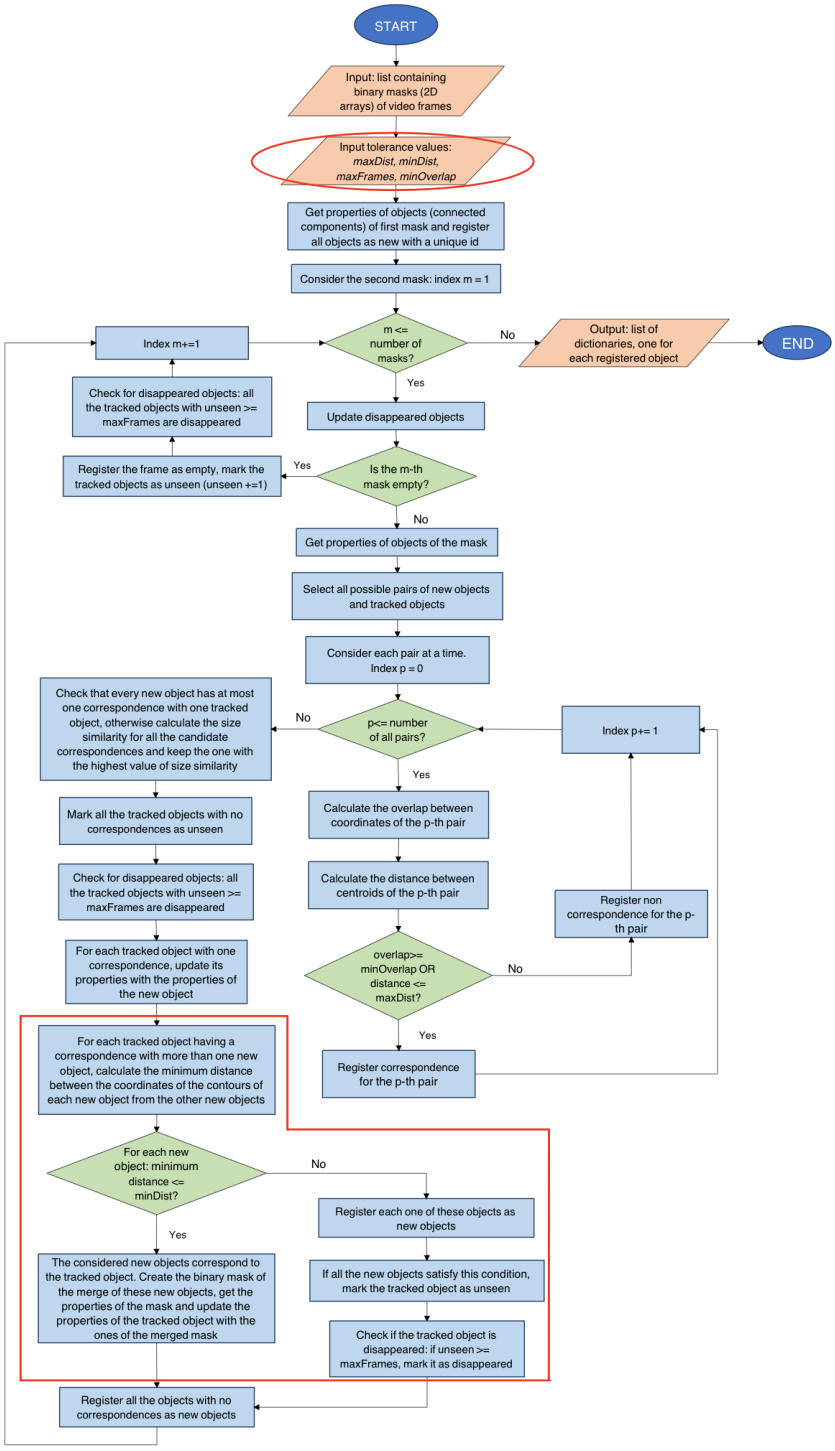


Figure 3.32: Pipeline of the tracking algorithm that incorporates the implemented function that calculates the minimum distance between objects of the same frame. The red boxes contain the sections that were modified and updated with respect to the pipeline of figure 3.14.

In particular, with respect to the pipeline described in section 3.4.1, there is one more input given to the algorithm that is the threshold minDist, defined as the minimum distance between contours of objects in the same frame that correspond to the same object of the previous frame, above which these objects are new objects. Moreover, in case of split-merge situations, for the objects that correspond to the same tracked object the merged binary mask is created, and the merged properties are obtained. The other objects which should not be merged are registered as new. The tracking algorithm incorporating this new function was applied in combination with the denoising algorithm to a subset of 20 videos included in the dataset described in section 3.1.1 and compared with the output of denosing with tracking algorithm resulting before the new implementation.
From this analysis, it resulted that the split-merge situations were correctly managed; whereas, considering the videos where the large artifact overlapped with real activations, the implementation solved the problem only when the artifact was associated to a single connected component of large dimensions. Meanwhile, if the large overlapping artifact was composed of several small connected components the problem remained.
This is explained by the fact that, despite the application of the implemented function, the distance between the residual small connected components of the artifact and the real activations (that result overlapped with the artifact) is smaller than the threshold minDist, thus the real activations are wrongly recognized by the algorithm as part of the artifact. Therefore, as a future development, it will be necessary to consider these cases as well.

3.5.4 Managing the empty binary masks in the tracking algorithm

Other two changes that were performed in the tracking algorithm concerned the management of the binary masks given as input to the algorithm in the cases where these masks were empty, i.e., masks in which all pixels are equal to 0 and, therefore, no objects are identified.
The algorithm described in section 3.4.1 started from considering the binary mask of colored pixels obtained from the first frame of the video, then the objects contained within this mask were registered as new objects and their properties were obtained. However, this version of the algorithm did not take into account the case in which the very first mask was empty. Therefore, to generalize the algorithm, it was implemented a function that, given the masks as input, returns the index of the first not empty mask of colored pixels whose connected components are tracked, so that the algorithm can start from this mask.
Another identified problem consisted in the fact that the previous version of the tracking algorithm did not manage the case in which no object was tracked due to the presence of an empty mask within the video. In particular, it happened that, if the (i-1)-th mask was empty, when considering the i-th mask which was not empty, the objects of this mask were correctly identified but the tracking metrics as the distance between centroids and the overlap between coordinates of these objects could not be computed since in the (i-1)-th mask no objects were detected and, thus, no centroids or coordinates could be obtained.
To avoid this issue and properly manage the empty binary masks of colored pixels, the following check was added to the algorithm before calculating the two metrics:

- if the number of already tracked objects, namely the objects identified in the previous frame, is larger than 0, the algorithm pipeline remains unchanged, thus the distance between centroids and the overlap between the coordinates are calculated;

- if there are no tracked objects it means that the previous mask is empty, thus every object of the current mask is registered as a new object.

The algorithm pipeline including both changes described in this section and the function illustrated in the previous one (see section 3.5.3) is presented in figure 3.33 of the next page.

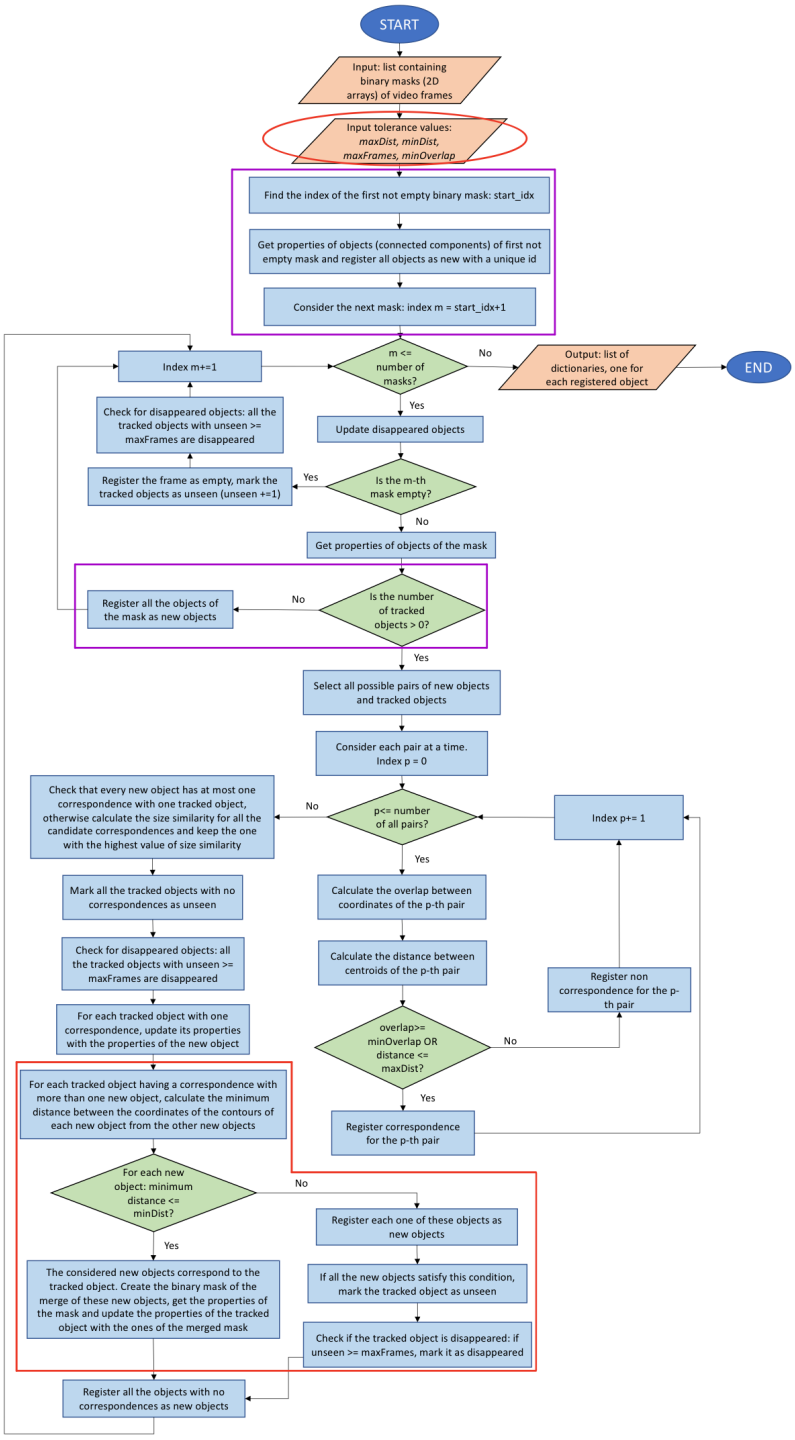


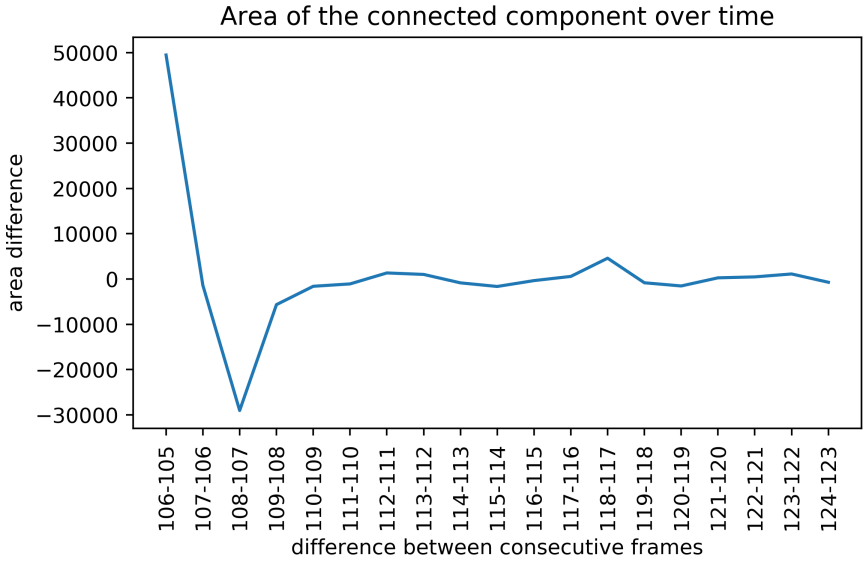
Figure 3.33: Flow chart of the tracking algorithm that includes the management of empty binary masks (the corresponding blocks are identified by the purple boxes) and the function described in section 3.5.3 incorporated inside the red boxes. The boxes contain the modified blocks with respect to the pipeline of figure 3.14.

3.5.5 Analysis of connected components over time

As resulted from the qualitative assessment of the outputs of the artifact-removal algorithms described in section 3.5.1, the presence of artifacts overlapping with real activations is the most common problem that led to poor performances of the component-tracking denoising algorithm according to the assessment criteria.
It has to be considered that when an artifact overlaps with a real doppler activation, it is assumed to be a connected component of high dimension characterized by a low temporal persistence (i.e., it disappears in a short number of consecutive frames) while the real signal is generally identified by a smaller connected component that persists for many consecutive frames of the video. Therefore, several parameters of size and shape of the connected components over time were analyzed. In particular, for each frame of the video, these parameters were calculated for the connected component corresponding to the real signal that, at a certain frame, overlaps with the artifact. By plotting the difference between these parameters of the connected component in consecutive frames, it is possible to appreciate when the artifact is superimposed to the signal because, in this case, the dimension or the shape of the resulting object changes, and a peak should appear in the graph for that frame. On the other hand, when the real doppler signal is not overlapped with the artifact, the value of the analyzed parameter should be approximately constant and, thus, the difference should approach zero.
This analysis was performed on 8 connected components coming from different videos of 8 clinical cases, among which 6 were characterized by the overlap between the doppler signal and an artifact, instead the remaining 2 objects, in which no overlapping was verified, were considered as a control case.
The parameters chosen to keep track of these connected components over time were:

- area: the number of pixels of the connected component
- convex area: area of the convex hull image, which is the smallest convex polygon that encloses the object
- major axis length: the length of the major axis of the ellipse that has the same normalized second central moments as the object
- minor axis length: the length of the minor axis of the ellipse that has the same normalized second central moments as the object
- aspect ratio: ratio of minor axis length to major axis length
- eccentricity of the ellipse that has the same second-moments as the object. The eccentricity is the ratio of the focal distance over the major axis length. The value is in the interval [0, 1). When it is 0, the ellipse becomes a circle

- perimeter of the object
- solidity: ratio of pixels in the region to pixels of the convex hull image



(a)

(b) Frame 105 **(c)** Frame 106 **(d)** Frame 107 **(e)** Frame 108

Figure 3.34: Example of the analysis of one connected component over time. (a) represents the plot of the difference between the areas of the analyzed connected component calculated for consecutive frames. In the x axis only the frames in which the connected component appears are inserted. Notice that there are two peaks: the highest appears when passing from frame 105 to frame 106, the second when moving from frame 107 to 108. Figures (b), (c), (d) and (e) depict the binary mask of the connected component at frames 105, 106, 107 and 108 respectively.

As seen from the example of plot shown in figure 3.34a, when the artifact was overlapped with the real signal, a significant change in dimensions of their connected component occurred and a peak (negative or positive) emerges in the plot for the corresponding frames, as expected.

Analyzing the plots of the difference of the parameters over time for different connected components, it is also possible to find a value of these parameters that can be used as a threshold above which the overlapping between the real signal and the artifact is verified. Below this threshold, the connected component consists of real signal only.
Therefore, the preliminary data produced during this analysis will be useful in the future, when a further study will be conducted in order to identify the optimal values of these thresholds for distinguishing between artifacts and real activations.

3.6 Tuning of the component tracking algorithm

In summary, the updated version of the tracking algorithm takes as input the binary masks where pixels corresponding to doppler activations are highlighted, and uses four tolerance values. These tolerance values are:

- maxDist, the maximum distance, in pixels, between centroids of the same object in two consecutive frames
- maxFrames, the maximum number of frames where an object is still tracked, even if it is not visible
- minOverlap, the minimum value of overlap for which two objects are considered correspondent
- minDist, the minimum distance, in pixels, between contours of objects in the same frame that correspond to the same tracked object above which these objects are new objects.

These parameters were tuned in order to enable the tracking algorithm to perform in the best possible way.
I empirically found out that the best tracking results were reached by setting the tolerance maxFrames equal to 0. That is, there is no tolerance for disappearing objects: if a tracked object is not seen in the considered frame, it is registered as disappeared and is not tracked anymore.
Regarding the overlap, it was computed as the number of pixels of the intersection between the new object and the already tracked object normalized with respect to the number of pixels of the new object. The selected value of minOverlap was 0.2.
The threshold maxDist was set to 10 pixels because this was proven to be the optimal value to keep track of the same object throughout the whole video.
Finally, the value of minDist was selected after having tested the performances of the tracking algorithm on a set of 20 videos included in the dataset described in section 3.1.1 considering different values of this parameter.

This dataset was composed of both videos where a doppler activation splits in two or more connected components passing from one frame to another and videos where there is a persistent artifact of large dimensions that overlaps with many distant connected components which, when the artifact disappears, should be recognized as different doppler activations and, therefore, different new objects.
The tested values of the threshold - applied to the distance between the contours of the objects present in the same frame - ranged from 20 to 70 at steps of ten.

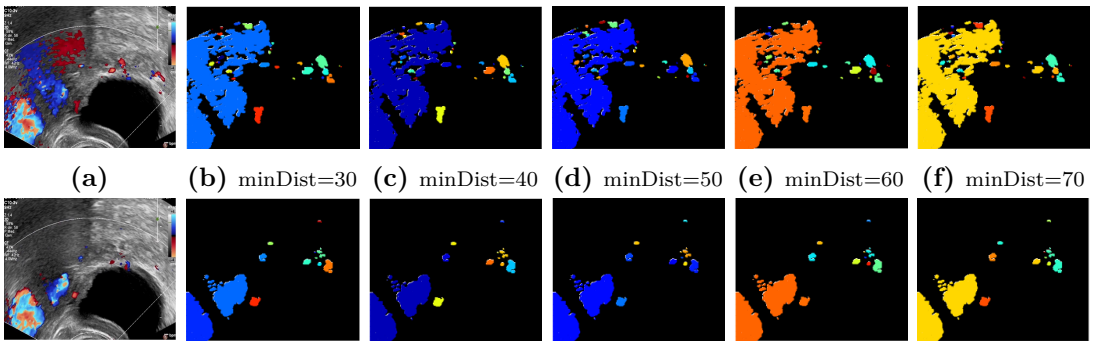


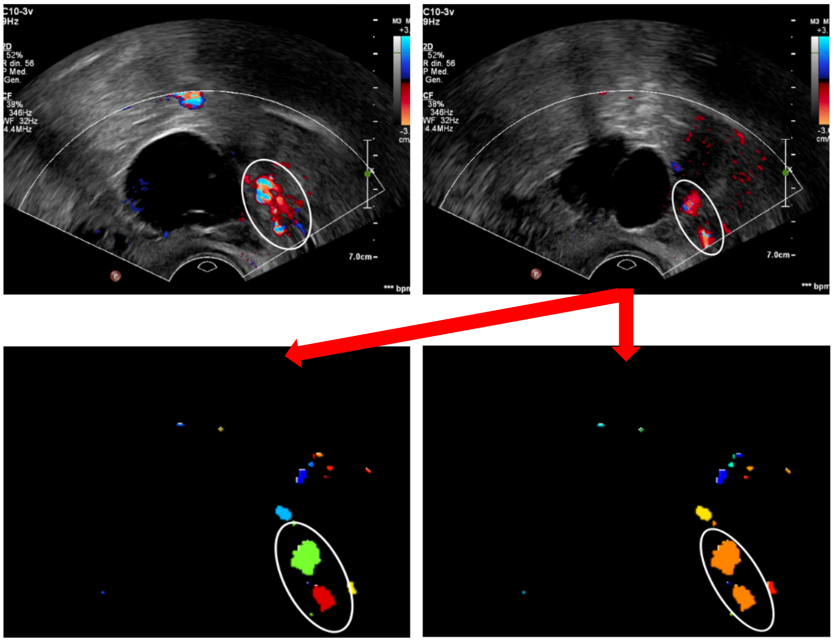
Figure 3.35: The figure shows an example of a video where at the i-th frame, shown in (a), there is an artifact that overlaps with the real signal, while in the (i+1) frame only the real activations remain, as shown in (g). The figures (b), (c), (d), (e) and (f) represent the outputs of the tracking algorithm at the i-th frame setting minDist equal to 30, 40, 50, 60 and 70 respectively. Notice that the artifact is depicted in light blue, blue, blue, orange and yellow in the five images. The figures (h), (i), (j), (k) and (l) show the output at the (i+1)-th frame obtained for the different values of minDist. Only when minDist is set to 30, the two connected components representing the two doppler activations are correctly identified with different colors, blue and light blue.

It was noticed that, when setting minDist = 40, 50, 60 and 70, objects clearly corresponding to different doppler activations resulted erroneously associated to the same connected component, (as seen in the figures 3.35i, 3.35j, 3.35k, 3.35l representing the outputs of the tracking algorithm obtained for the same frame using the four different thresholds), thus these tolerance values were too high.
On the other hand, setting the threshold to 20 the split/merge problem occurred, meaning that, in some videos, the two or more connected components in which a doppler activation split were wrongly identified as different objects (see figure 3.36 for the example).
Therefore, minDist = 30 was the optimal value that ensured, on one hand, that the split-merge problem did not appear again, and, at the same time, it allowed to

partially solve the problem caused by the presence of big and persistent artifacts during the video, as shown in figure 3.35h.
In summary, this version of the tracking algorithm uses five tracking measures – distance between centroids, number of frames where an object is tracked, overlap between coordinates, size similarity measure, distance between contours – and the following tolerance values:

- maxDist = 10
- maxFrames = 0
- minOverlap = 0.2
- minDist = 30

(a) i-th original frame **(b)** (i+1)-th original frame



(c) minDist = 20 **(d)** minDist = 30

Figure 3.36: The figure displays an example of a video where a doppler activation, indicated by the white circle in the frame (a), splits into two connected components shown in the frame (b). As indicated by the arrows, the figures (c) and (d) show the outputs of the component tracking algorithm corresponding to the original frame (b), obtained setting minDist equal to 20 and 30 respectively. While with minDist = 20 the two objects inside the white circle are wrongly associated to two different connected components (in green and red), using minDist = 30 they are correctly identified as the same object depicted in orange.

3.7 Tuning of the component-tracking denoising algorithm

Afterwards, the component tracking algorithm was integrated in the denoising in order to have an algorithm able to suppress artifacts while keeping track of the real signal throughout the video. The resulting denoising with tracking algorithm was applied to all 101 videos of the dataset.
The results were compared with the pixel-based denoising algorithm applied on the same dataset.
The denoising algorithms are characterized by the signal to artifact threshold, namely, the threshold that was applied to distinguish between artifacts and real doppler activations, based on the assumption that real signal activations are more persistent than artifacts. In particular, the 80th, 90th, 95th and 98th percentile of the distribution of the activation lengths of the pixels and the tracked objects were tested as possible threshold values for pixel-based denoising and component-tracking denoising algorithms respectively.
To tune the denoising algorithm and choose the appropriate signal to artifact threshold, a model was trained to predict the color score based on the estimation of the amount of doppler signal within the lesion evaluated on a dataset of denoised videos (the results of this assessment are illustrated in sections 4.2 and 4.3 for pixel-based denoising and component-tracking denoising respectively).
The signal to artifact threshold was chosen by means of this predictive model because the quantity of doppler signal within the adnexal mass changes based on the value of this threshold (this quantity decreases if, applying the threshold, some real doppler activations are suppressed because they are interpreted as artifacts, whereas it increases if a large number of artifacts survives to the denoising) thus leading to a color score prediction, performed by the model, that can be more or less accurate. Therefore, the optimal value of the signal to artifact threshold chosen for the denoising algorithms was the one that lead to higher classification performances.

3.7.1 Definition of the ground truth

The ground truth, selected to perform the quantitative assessment of the denoising, consisted in the color score values that were assigned by highly experienced clinicians to a dataset of ovarian cancer cases.
These color scores resulted from a study conducted by Syndiag in collaboration with A.O. Ordine Mauriziano and Policnico di Sant'Orsola hospitals, where six expert clinicians (three for each hospital) evaluated 100 cases of adnexal masses by assigning to them the color score. Afterwards, the Syndiag team evaluated the agreement among clinicians in assigning the color score.

From this analysis, 54 cases, on which a good agreement in the color score assignment was reached among clinicians, were selected. The dataset was then incremented with other 52 cases, whose color score was assigned by expert labelers with the supervision of the clinicians. The final dataset, described in section 3.1.2, resulted in a total amount of 106 videos.

Image segmentation

The color score is defined as the qualitative amount of doppler signal within an adnexal mass. To extract the amount of estimated doppler signal:

- the doppler colored pixels must be isolated
- it is necessary to identify the region of interest (ROI) where the doppler signal relevant for the analysis is present. From a medical point of view, this ROI consists of the intersection between the adnexal mass walls and the area of the doppler acquisition.

In order to find this ROI, the videos were labeled, then the amount of doppler signal within the ROI in the original videos and the one *survived* to the pixel-based denoising and component-tracking denoising algorithms was quantified (see section 3.7.1) and, finally, from this value, the color score was automatically predicted by the implemented models described in section 3.7.2.
A total number of 2986 frames were labeled. These frames were taken from the 106 cases of the dataset. It was decided not to label all the frames of the available cases but only the most representative ones so that the number of videos under analysis could be maximized. This is consistent with the medical doctors' usual assessment of doppler data.

The manual segmentation of the ROI was performed using RedBrick AI [48] that is a unified labeling platform that allows to label and manage data.
On each frame, two labels were positioned:

- the whole adnexal mass, by means of a pixel-sensitive segmentation tool
- the doppler fan area, by means of a polygon tool.

The platform returned as output the binary masks containing the labeled adnexal mass and the coordinates (normalized with respect to the size of the image) of the points that were placed by the labeler to form the closed polygon that contained the doppler fan.

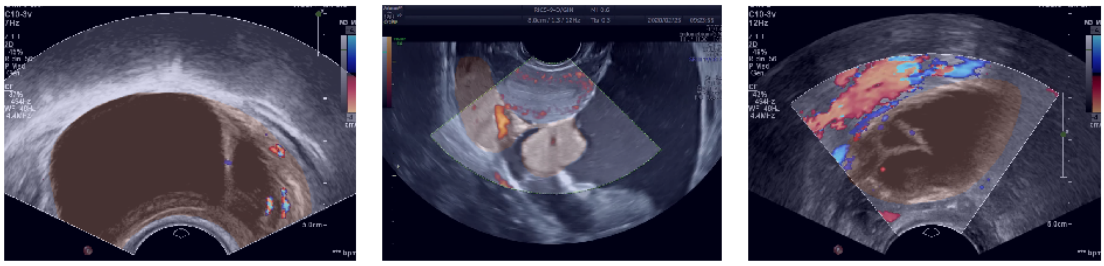


Figure 3.37: Examples of labeling performed on RedBrick AI platform. The adnexal masses are shown in orange, while the areas of the doppler fan are shown in white.

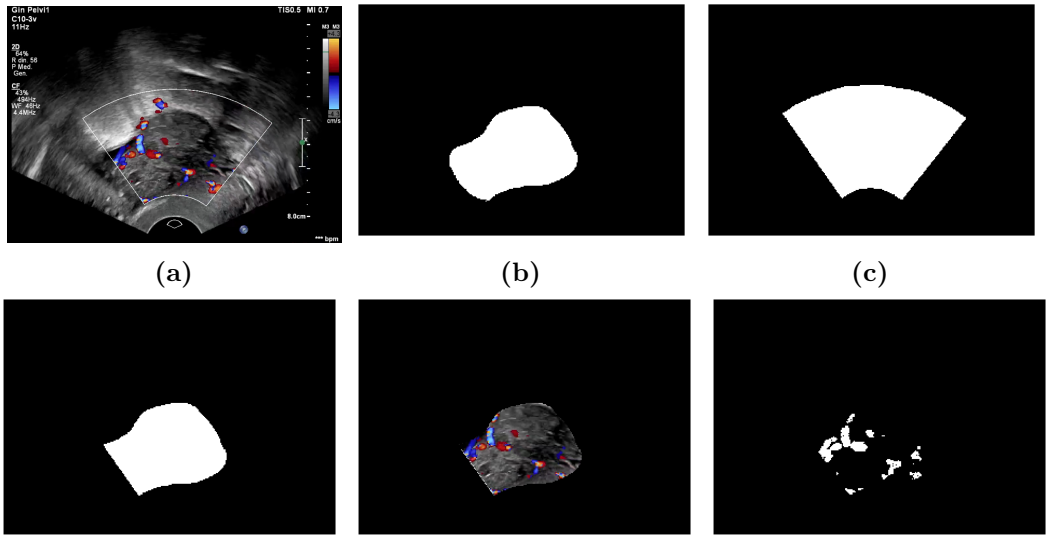
Calculation of doppler pixel count

After having labeled the doppler fan and the adnexal mass for the selected frames, the number of colored pixels corresponding to the doppler signal within the adnexal mass in each frame was calculated considering both original, denoised and denoised+tracked videos [1], this number was referred to as doppler pixel count. The doppler pixel counts are the inputs given to the model that predicts the color score based on the amount of doppler signal within the lesion (see section 3.7.2). According to the IOTA guidelines, the color score is defined as the amount of vascularization within the adnexal lesion. Therefore, this qualitative parameter was predicted based on the quantity of doppler signal within the ROI. This quantity was calculated in terms of doppler pixel count as the ratio between the number of colored pixels within the ROI and the area of the ROI itself:

$$doppler\ pixel\ count = \frac{number\ of\ colored\ pixels\ within\ the\ ROI}{area\ of\ the\ ROI} \tag{3.3}$$

Given that the color score is an ordinal variable and its definition is based on the amount of doppler signal, it is expected that the larger the number of colored pixels within ROI, the higher the assigned value of color score for the considered video. The normalization of the number of colored pixels with respect to the area of the ROI was performed because it is necessary to take into account that for different doppler videos the acquisition can be performed focusing on masses having different dimensions and setting different zoom levels.

[1] Original videos are intended as those videos on which no algorithm to remove artifacts is applied. Denoised videos are the ones resulting from the application of the pixel-based denoising algorithm. Finally, denoised+tracked videos are referred to as the videos on which the connected components-based denoising with component tracking is applied.



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure 3.38: Pipeline to evaluate the doppler pixel count. (a) Original frame taken as example. (b) Binary mask of the adnexal mass. (c) Binary mask of the doppler fan. (d) Binary mask of the intersection between the lesion and the fan, whose area represents the denominator of equation 3.3. (e) Mask resulting from the product between (a) and (d), it isolates the colored pixels within the lesion. (f) Binary mask where colored pixels are set to 1, the number of these pixels is the numerator of equation 3.3.

The steps performed to obtain the doppler pixel count for each labeled frame are:

1. From the original frame (or denoised frame or denoised+tracked frame), the binary mask where the pixels corresponding to the adnexal mass are set to 1 and the binary mask of the doppler fan are computed (see figures 3.38b and 3.38c).

2. The intersection between the mask of the adnexal mass and the mask of the fan is obtained and its area is calculated as the number of pixels constituting this ROI. (see figure 3.38d).

3. The considered frame is multiplied by the mask of the intersection (see figure 3.38e).

4. In the image resulting from step 3, the colored pixels associated to a doppler activation are identified as those pixels where there is a difference of at least 30 between R and G, R and B or G and B. The number of colored pixels is obtained (see figure 3.38f).

5. The amount of doppler signal within the mass for the considered frame is calculated as in the equation 3.3.

Applying this pipeline, for each case a list of doppler pixel counts (one value for each labeled frame) is obtained for both the original video, the video resulting from the application of the pixel-based denoising algorithm and the video on which the denoising with component tracking algorithm was applied.

Selection of threshold to apply on doppler pixel count

In the clinical practice, clinicians, in order to assign the color score to a clinical case, watch the whole video and then refer to those frames where the highest amount of what they see as real signal is present within the lesion. A similar approach was applied for the doppler pixel count, with two differences: first, the doppler pixel count is a quantitative and measurable parameter, while the color score is a subjective and qualitative variable; secondly, instead of considering the maximum as the most representative value of doppler pixel count to be given as input to the color score predictive model, a threshold was applied to the doppler pixel count to account for random fluctuations in the measurement.
The color score was predicted based on the doppler pixel count values obtained considering three different conditions:

- on original videos, without applying any artifact-removal algorithm,
- when the pixel-based denoising is applied,
- when the component-tracking denoising is applied.

The number of colored pixels conveniently normalized in original videos includes both real vascularization and artifacts, while, in the other two conditions, it consists in the fraction of pixels survived to the artifacts' suppression.
Different thresholds were proposed and an analysis was conducted in order to choose the optimal threshold that better characterized the videos.
Firstly, it was represented a bar plot where the average threshold values obtained for the videos having the same color score and the corresponding standard deviations were displayed as a function of the color scores for original, denoised and denoised+tracked videos (see figure 3.39 showing the bar plot obtained employing as a threshold the median of the doppler pixel counts larger than the 80th percentile of their distribution). This way, possible differences in the number of colored pixels within the masses in the three conditions could be seen.
The most important aspect emerging from the bar plots was the presence of large variability of the doppler pixel count values due to the high standard deviations.

The lowest variability was reached when the pixel-based denoising was applied. Moreover, the mean of doppler pixel counts for videos having color score 1 was always larger than zero for the three types of videos. This is in contrast with the definition of color score introduced by IOTA according to which a color score of 1 is assigned when no vascularization is shown within the mass. This means that the colored pixels present in those videos are artifacts and that the clinicians were able to correctly recognize those pixels as artifacts and not real signal.
Considering the three conditions, the doppler pixel count increases as the color score increases, as expected. For the denoised videos, the doppler pixel count values for color scores 1 and 2 are similar, therefore they may be difficult to distinguish. Instead, considering the denoised + tracked videos, the doppler pixel count for CS (i.e., color score) = 1 is higher than the one resulting from the denoised videos, hence too many artifacts are still considered as real signal by this algorithm.

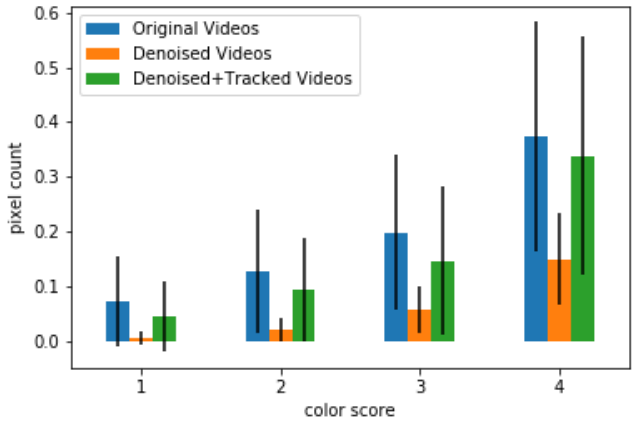
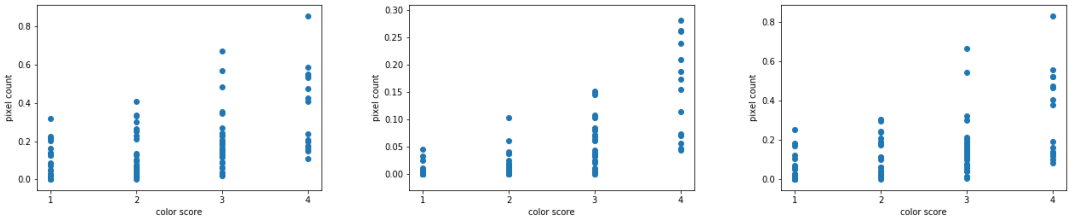


Figure 3.39: Bar plot showing the mean of the threshold values applied to doppler pixel counts of the videos having the same color score for original, denoised and denoised+tracked videos as a function of color score. The threshold used to obtain this plot corresponds to median of the doppler pixel count values larger than the 80th percentile of their distribution.

Afterwards, in order to show the correlation between the color scores and the threshold value for each video of the dataset, three separate scatter plots - where the threshold value for each video as a function of the color score was displayed - were made for original, denoised and denoised + tracked videos.
Considering the scatter plot obtained for the original videos (see figure 3.40a), color score 1 and 2 could not be distinguished and the same happened for color score 3 and 4. When the pixel-based denoising was applied, the doppler pixel count values were generally lower, and they differed more for the different color scores.



(a) original videos **(b)** denoised videos **(c)** denoised+tracked videos

Figure 3.40: Scatter plot of doppler pixel count values vs color scores for original (a), denoised (b), and denoised+tracked (c) videos. The threshold applied on the doppler pixel count values to obtain this plot corresponds to median of the doppler pixel count values larger than the 80th percentile of their distribution.

Finally, given the large variability that emerged from bar plots, the distributions of doppler pixel counts for the three conditions and for the videos having the same color score value were obtained in order to analyze their variability and to identify more easily the threshold to distinguish between different color scores.
Therefore, four distributions of doppler pixel counts groups each one corresponding to a color score were generated for the three types of videos – original, denoised, and denoised+tracked videos.
Considering denoised videos, there was no significant difference between color score 1 and 2 distributions as shown in figure 3.41. This algorithm removes the artifacts, but it makes difficult to differentiate between the four distributions meaning that it is also eliminating real activations, especially in videos with CS = 4.

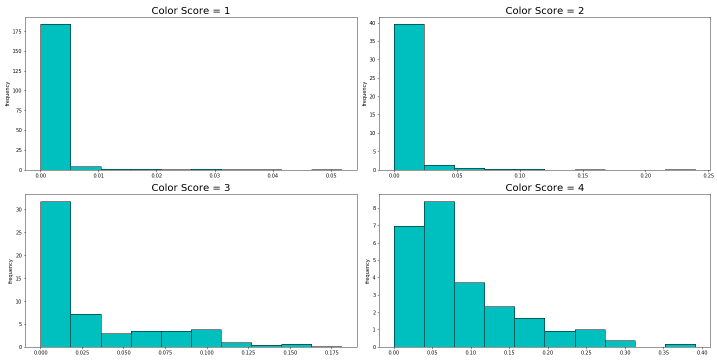


Figure 3.41: Distribution of doppler pixel count groups, each corresponding to a color score, for denoised videos. Other distributions were obtained also for the original and denoised+tracked videos.

From the conducted analysis, it resulted that the threshold that maximized the

difference between the distributions of doppler pixel count values obtained for the different color scores and that, thus, better represented the single videos was the median of the doppler pixel count values larger than the 80th percentile of the doppler pixel count distribution. Therefore the model, described in the next section, was trained receiving this threshold as input.

3.7.2 Color score predictive model

Once obtained the most representative value of doppler pixel count for each video of the dataset, it was used as a feature to train a model, which is a classifier, able to automatically predict the color score.
For the three conditions, the same model was built – only the input changed, i.e. the doppler pixel counts obtained for the original, denoised or denoised+tracked videos – performing a supervised learning, meaning that the model gets trained on a labelled dataset consisting of both the input parameters represented by the doppler pixel counts and the output parameters that are the corresponding color scores assigned by clinicians.
The performances of the classifiers were evaluated in terms of accuracy, sensitivity and specificity and, then, compared to show which method led to better results.
The implemented classifier is a Decision Tree which gives as output three thresholds that divide the doppler pixel counts in 4 groups each belonging to one of the 4 classes represented by the 4 color scores. After the identification of these thresholds, they can be applied to assign the color score to completely new samples.

A decision tree is a non-parametric and supervised machine learning algorithm often employed for classification problems. It has a hierarchical, tree structure characterized by a root node, branches, internal nodes, and leaf nodes.
In particular, a decision tree starts with a root node from which branches develop and arrive to the internal (or decision) nodes where a test is performed on an attribute. The outcomes of these tests are represented by the branches of the tree that lead to the leaf nodes (or terminal nodes), each one holding a class label.
Decision tree learning uses a divide and conquer strategy by performing a greedy search where the starting dataset splits into subsets based on the attribute value test. This process of splitting is repeated in a recursive manner until all, or the majority of the samples in each subset have been classified under the same class label.

The splitting criterion employed for the built decision trees was the Gini Impurity. This score gives an idea of how good a split is by how mixed the response classes are in the groups created by the split.
Considering a dataset D that contains samples from k classes and defined p_i the probability of samples belonging to class i at a given node, the Gini Impurity of D

is denoted by the formula:

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

A perfect class purity occurs when a group contains all inputs from the same class, in which case G = 0, whereas a node having a 50–50 split of classes in a group has the worst purity (G = 0.5).

However, when the decision tree grows in size and, consequently, its complexity increases, it becomes increasingly difficult to arrive to pure leaf nodes where all the samples belong to the same class. When this increase in complexity and dimensions occurs, it can often lead to overfitting.
Therefore, to avoid overfitting and start generalizing well to new data, pruning was employed. In particular, the implemented type of pruning was the minimal cost-complexity pruning, a process that controls the size of the tree by removing the branches that have lower importance after that the complete tree was constructed. In this way, the complexity of the tree is reduced and its predictive power increases. The algorithm used to apply the pruning is based on the complexity parameter α used to weigh whether nodes can be removed. The higher the value of α, the higher the number of nodes pruned and, hence, the lower the tree's complexity.
This method recursively finds the node with the "weakest link" that is characterized by an effective α; the nodes with the smallest effective α are pruned first.
In particular, the first step of this pruning algorithm consists in finding the pruning path which gives the effective αs and the corresponding total leaf impurities at each step of the pruning process. This means that, for each step, the effective α is evaluated, and the corresponding node is pruned: this process continues until one node remains, thus the minimum complexity of the tree and the maximum value of α are reached. These values of α are then used to train a Decision Tree and, for each training, the accuracy on both training set and test set is computed. Finally, the optimal α is selected as the one that leads to the best compromise between train and test accuracy.

In order to evaluate the model's fit and make the performances of the classifiers comparable, the cross-validation was performed.
The most used cross-validation technique is K-Fold that divides the dataset into k subsets called folds. k-1 folds constitute the training set, while the remaining fold is held for testing, and this split is repeated until each of the folds is employed as test set. Then, the K folds are fit and evaluated, and the mean accuracy for all these folds is computed. However, this process works well for balanced classification tasks, but it fails for imbalance classes. This is because the k folds are obtained by splitting the data randomly without taking care of the class imbalance.

In this context, the available dataset is small and unbalanced because it contains a large number of samples having color score 1, 2 or 3 and only few samples whose color score is 4. Therefore, the stratified K-fold cross-validation with k = 5 was performed because it considers the class imbalance by maintaining the same class ratio throughout the K folds as the ratio in the original dataset.

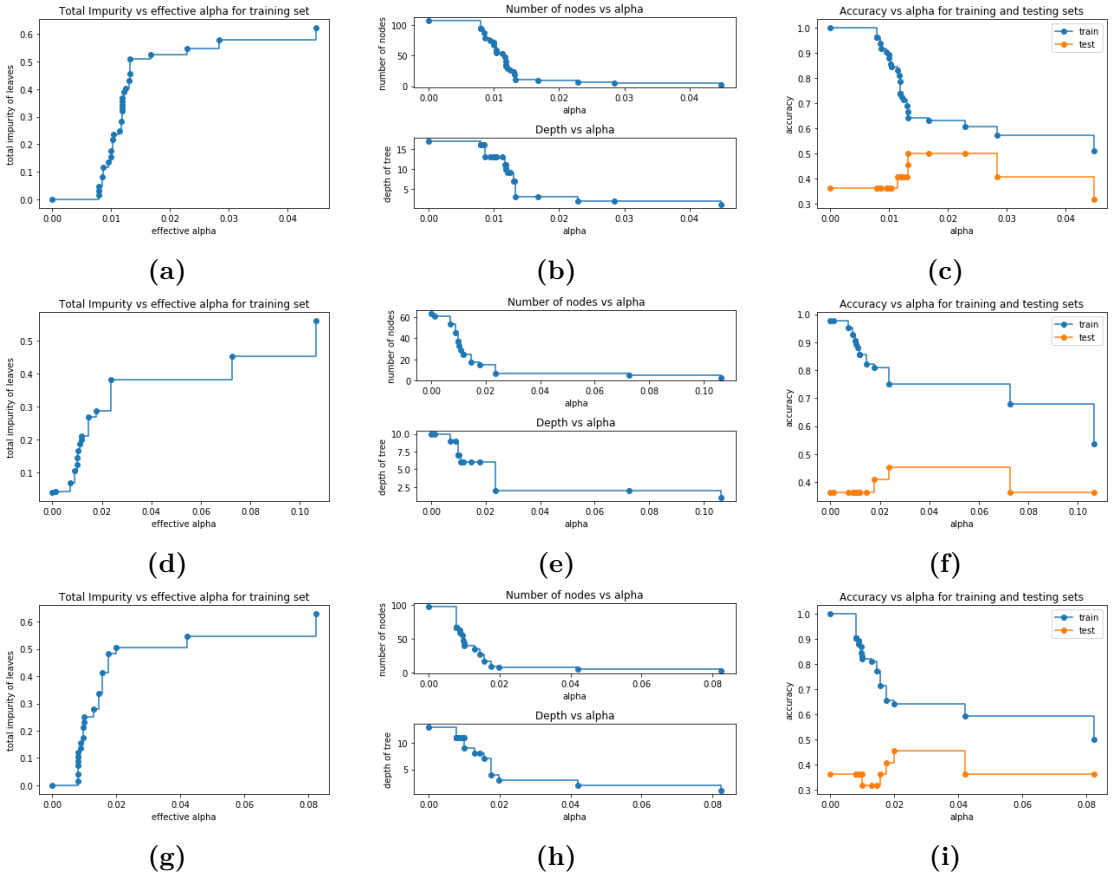


Figure 3.42: The figure shows the steps performed to choose the optimal value of α during the pruning procedure. The figures (a), (b) and (c) were obtained for the original videos; (d), (e) and (f) for the denoised videos; (g), (h) and (i) for the denoised+tracked videos. (a), (d) and (g): pruning path with effective αs and the corresponding number of total leaf impurities. As α increases, more of the tree is pruned, which increases the total impurity of its leaves. (b), (e) and (h): number of nodes and tree depth as a function of α. Both variables decrease as α increases. (c), (f) and (i): accuracy vs α for training and set sets: as α increases, more of the tree is pruned, thus creating a decision tree that generalizes better. The optimal values of α are 0.02, 0.05 and 0.04 for original, denoised and denoised + tracked videos respectively, as they provide the best compromise.

In summary, each of the three Decision trees was built following these steps:

1. The doppler pixel counts obtained for original videos (denoised, or denoised + tracked videos) of the dataset, representing the only feature, and the corresponding color scores, used as labels, are taken to be given as inputs to the classifier.

2. The doppler pixel counts are divided into training set and test set (80% and 20% of pixels counts respectively).

3. The pruning path is calculated and the effective αs with the corresponding impurities are obtained (their relationship is shown in figures 3.42a, 3.42d and 3.42g).

4. The effective αs are used to train a decision tree and the corresponding accuracies on training and test set are computed (see figures 3.42c, 3.42f and 3.42i).

5. The optimal value of α is chosen as the one that ensures the best compromise between train and test accuracy. In this case the chosen values were 0.02, 0.05, 0.04 for original, denoised and denoised + tracked videos respectively.

6. The stratified K-fold cross-validation with k = 5 is performed to validate the model that includes the pruning with the optimal α.

Chapter 4

Artifacts suppression from Doppler signal results in an improved color score prediction

As described in section 1.2.1, the color score is a scoring system that indicates the amount of blood flow within the adnexal mass. According to IOTA [4], vascularization can be described as having color score 1 if no blood flow is present within a lesion, 2 if the blood flow is minimal, 3 if the flow is moderate and 4 when the adnexal mass is highly vascular. Medical doctors assign color score considering the maximum amount of doppler signal (i.e. colored pixels) that is visible within the lesion. Color score is a powerful yet qualitative measurements, that is user dependent and, thus, can be affected by the presence of doppler artifacts within the ultrasound video, leading to low agreement.
In order to ease clinicians' evaluation of the lesions and improve their concordance in the assignment of color score, two algorithms that remove doppler artifacts – the pixel-based denoising algorithm and the connected components-based denoising with component-tracking – were developed (they are illustrated in sections 3.3 and 3.4 - 3.5) on 101 ovarian cancer cases.
To assess the effect of the artifact-removal algorithms, a predictive model, namely a decision tree, was trained to predict the color score based on the doppler estimation on both noisy videos and denoised doppler videos of 106 ovarian cancer cases. Each video of these cases has its correspondent color score label, assigned by clinicians. Meanwhile, to estimate the doppler amount, I calculated the doppler pixel count, defined as the number of colored pixels present within the lesion. The rationale

was that a well-tuned artifact removal algorithm should lead to more accurate color score prediction starting from *survived* doppler activations, than a prediction simply based on the complete original, not denoised, doppler activations.
In particular, three experiments were conducted: in the first, a predictive model was trained and tested on original videos; in the second applying the pixel-based denoising with different signal to noise thresholds; in the third applying the component-tracking denoising and considering, also in this case, different threshold values.
The classification performances in terms of accuracy on test and training sets obtained from the three experiments were then compared and the impact that the artifact-removal algorithms may have on clinical practice was discussed.

4.1 Experiment 1: Color score prediction from original videos

The first experiment that was performed consisted in training and testing the decision tree (whose features are described in section 3.7.2) on the original noisy videos. Therefore, the doppler pixel count value obtained for each video (calculated as the median of the doppler pixel counts larger than the 80th percentile of the distribution of the doppler pixel count values obtained for each frame of the video) of the dataset included both the colored pixels associated to real doppler activations and the ones representing doppler artifacts that were not removed in this case.
Before training the model, for the three experiments, the doppler pixel count values were divided into four groups based on their color score and the characteristics of these groups were analyzed.
As previously mentioned, the employed dataset is not balanced towards color scores 1, 2 and 3, since only few samples having color score 4 were available, reflecting the average distribution of this pathology in the clinical practice.
In particular, the numerosity of each class of color scores is the following:

- 28 samples with color score 1
- 28 samples with color score 2
- 35 samples have color score 3
- 15 samples have color score 4

The figure 4.1 shows a box plot of the doppler pixel counts obtained considering the whole dataset, where on the y-axis there are the doppler pixel count values, whereas on the x-axis the color scores are displayed.
From the box plot, it is possible to see a trend where the higher median values

of the doppler pixel counts are associated to higher color scores. However, the four boxes, each one representing a color score, are partially – or completely, as it happens for boxes associated to CS = 2 and CS = 3 - overlapped, meaning that when considering original noisy videos it is difficult to distinguish between different color scores (especially 2 and 3) on the basis of the doppler pixel count values.
Moreover, the doppler pixel counts for the samples having color score 1 (i.e. absence of doppler activation within the lesion) are larger than zero, suggesting that these values include doppler artifacts.

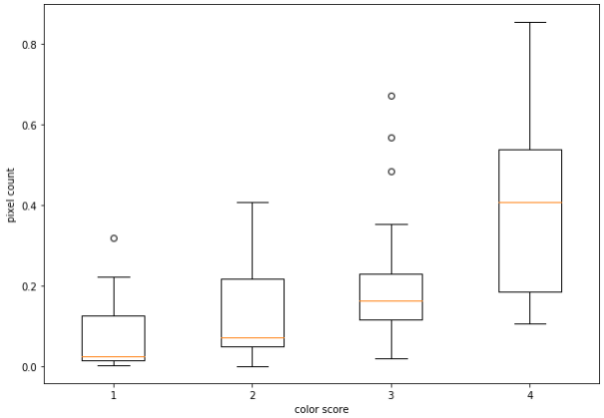


Figure 4.1: Box plot of doppler pixel count values for original videos vs color scores.

The Decision Tree including minimal cost complexity pruning and validated using the stratified K-fold was then trained on a 80% - 20% train - test partition of the 106 videos of the dataset.
The accuracy on the training set was calculated for each cross-validation "cycle" (see table 4.1a), the obtained values were averaged, and their standard deviation was calculated and the resulting accuracy is equal to 0.62 and +/- 0.02.
Moreover, the test accuracy was evaluated for each of the 5 folds that constituted, in turn, the test set. The overall accuracy was computed as the mean of these accuracy values +/- their standard deviation and it resulted equal to 0.53 (standard deviation = +/- 0.06).
The cumulative confusion matrix of table 4.1b was obtained by summing up the confusion matrices of each fold constituting the test set, hence considering all the samples of the dataset. It can be seen that there is a high number of samples, 13, that have color score 2 but they are misclassified as color score 3. Moreover, 10 samples with color score 1 are misclassified as class 3.

k	train accuracy	test accuracy
1	0.64	0.50
2	0.58	0.62
3	0.62	0.48
4	0.61	0.48
5	0.62	0.57

(a)

True labels \ Predicted labels	1	2	3	4
1	14	4	10	0
2	3	11	13	1
3	2	6	24	3
4	0	1	7	7

(b)

Table 4.1: Classification performances for the first experiment conducted on original videos considering 4 labels, each one corresponding to one color score value. (a) Train and test accuracy values for each fold used in cross-validation of the model. (b) Cumulative confusion matrix on whole dataset.

Trial	train accuracy	test accuracy
start	0.62 +/- 0.02	0.53 +/- 0.06
1	0.68 +/- 0.02	0.48 +/- 0.12
2	0.60 +/- 0.02	0.49 +/- 0.12
3	0.55 +/- 0.06	0.52 +/- 0.03
4	0.62 +/- 0.02	0.53 +/- 0.06
5	0.62 +/- 0.02	0.53 +/- 0.06
6	0.62 +/- 0.02	0.53 +/- 0.06

Table 4.2: The table shows the accuracy values on train and test sets obtained testing different configurations of the decision tree on the original videos. The trial named start refers to the model whose characteristics are described in section 3.7.2. In trial 1 the criterion to measure the quality of a split was changed from gini to entropy. In trial 2 the weights associated with the classes were adjusted based on the proportion of each class frequencies. In trial 3 the best random split was chosen at each node. In trial 4 and 5, the minimum number of samples required to split a node and the maximum depth of the tree were respectively changed, they were set to all the integers ranging from 3 to 10. In trial 6, the maximum number of leaf nodes was set to all the integers ranging from 4 to 10.

The obtained value of test accuracy suggests that in almost the half of the cases the classifier is not able to assign the correct color score when it receives as input the doppler pixel count values of the original noisy videos.

In this context, in order to improve the classification performances, I tuned the

model considering different values of the following parameters given as input to the classifier:

- the criterion to measure the quality of a split. It was changed from the Gini Index to the Entropy (trial 1 of table 4.2)
- the weights associated with the classes. In trial 2 of table 4.2, the weights were balanced, meaning that they were adjusted according to the proportion of each class frequencies
- the strategy used to choose the split at each node. It was changed from *best* where the best split is chosen, to *random* in which the best random split is chosen (trial 3)
- the minimum number of samples required to split a node. It was set to all the integers ranging from 3 to 10, the default value was 2 (trial 4)
- the maximum depth of the tree. It was set to all the integers ranging from 3 to 10 (trial 5)
- the maximum number of leaf nodes. It was set to all the integers ranging from 4 to 10 (trial 6).

However, tuning of these parameters did not improve the classification performance (see table 4.2).
Considering the large overlap between the distribution of pixel count for color score 2 and 3 (as seen in the boxplot of figure 4.1) and the consequent misclassification between the two (as shown in the confusion matrix of table 4.1b), I considered a new prediction model trained merging cases with color score 2 and 3 in a unique, intermediate, condition. This decision was supported by the fact that, differently from CS = 2 and CS = 3, 1 and 4 represent the color score values that are more significant from the clinical and diagnostic point of view because the absence of vascularization (to which a CS = 1 is associated) is a sign of benignity, while an highly vascularized lesion is identified as a malignant tumor, as suggested by IOTA's Simple Rules illustrated in section 1.2.2.
In this context, the doppler pixel count values were divided into three groups with the following numerosities: 28 videos having color score 1, 63 samples with color score 2 or 3, 15 videos with color score equal to 4.
In the resulting box plot shown in figure 4.2, the groups appear more separated than the previous one, but still partially overlapped.

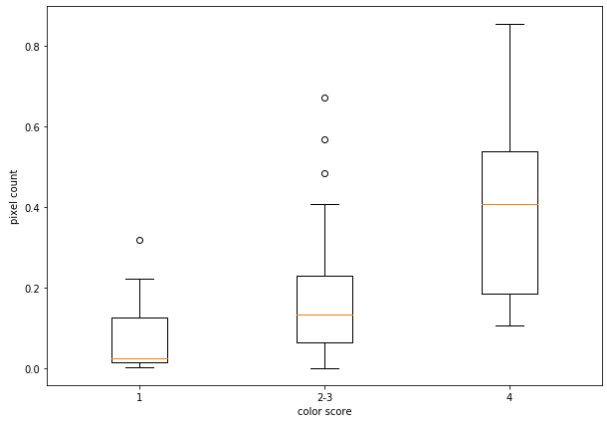


Figure 4.2: Box plot of doppler pixel count values for original videos vs the three labels (1, 2-3, 4).

k	train accuracy	test accuracy
1	0.74	0.73
2	0.74	0.71
3	0.73	0.76
4	0.72	0.81
5	0.77	0.62

(a)

		Predicted labels		
		1	2-3	4
True labels	1	16	12	0
	2-3	5	54	4
	4	0	8	7

(b)

color score	sensitivity	specificity
1	0.57	0.94
2-3	0.86	0.53
4	0.47	0.96

(c)

Table 4.3: Classification performances for the first experiment conducted on original videos. (a) Train and test accuracy values for each fold used in cross-validation of the model. (b) Cumulative confusion matrix on whole dataset. (c) Sensitivity and specificity values for each of the three color score classes.

The classification accuracy on training set resulted equal to 0.74 +/- 0.02, whereas the test accuracy is 0.73 +/- 0.06 (see table 4.3a for accuracies obtained for each fold), both values are higher than the ones obtained considering the four color scores as separate classes meaning that a high portion of wrong color score predictions performed by the model involved classes 2 and 3, now merged into one.
The corresponding confusion matrix of table 4.3b shows that the large majority of samples having color scores 2 or 3 are now correctly classified; while class 1 has the highest number of misclassified samples, all assigned to class 2-3. Considering color score 4, more than the 50% of samples belonging to this class are misclassified and identified as belonging to class 2-3.
In addition, the sensitivity and specificity values for each color score class were calculated and they are shown in table 4.3c. Color score 4 has the lowest sensitivity (0.47) meaning that a high number of samples having CS = 4 were erroneously assigned to another class; while class 2-3 has the highest sensitivity equal to 0.86.

Meanwhile, both color scores 1 and 4 have high specificity (0.94 and 0.96) meaning that, when the model classifies the clinical cases into class 1 or 4, the majority of these cases actually have color scores 1 or 4 respectively. In the contrary, class 2-3 has a specificity of 0.53, so the majority of the misclassified samples belonging to class 1 and 4 were predicted as having color score equal to 2 or 3.

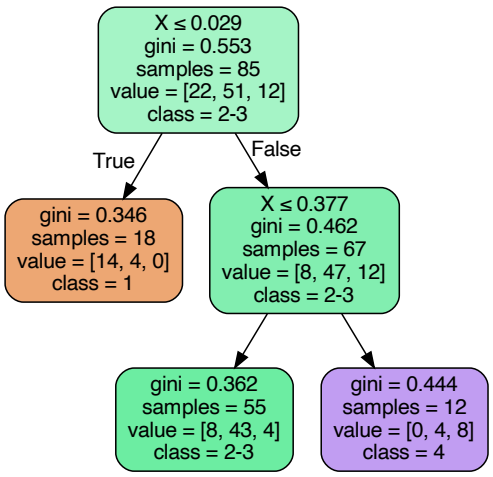


Figure 4.3: The figure shows the structure of the decision tree trained on original videos: class 1 node is depicted in orange, class 2-3 in green and class 4 in purple. Each node indicates the test that is performed on the samples at that node, the purity of the node defined by the Gini index, the number of samples that reach that node, the number of samples for each class, and the predominant class label.

The tree structure obtained for this experiment is shown in figure 4.3. Only 3 nodes were not pruned, hence the tree has a depth of two levels. Each leaf node shows the count of all samples that reached that node during the training, for a total of 85 samples, as the tree structure was obtained training the 80% of the dataset.
Notice that 8 samples having CS = 1 and 4 with CS = 4 reached leaf nodes not associated to class 1 and 4 respectively, coherently with the low values of sensitivity obtained for these two classes with respect to class 2-3. Moreover, the final node depicted in green, associated to class 2-3, contains the highest number of videos that actually belong to the other classes, as shown by the low value of specificity obtained for this class.
From the three leaf nodes left in the tree structure, it was possible to obtain the two thresholds that divide the dataset into the three classes:

- doppler pixel count $\leq 0.029 \rightarrow$ color score = 1

- $0.029 <$ doppler pixel count $\leq 0.377 \rightarrow$ color score $= 2$ or 3
- doppler pixel count $> 0.377 \rightarrow$ color score $= 4$

4.2 Experiment 2: Assessment of pixel-based denoising in color score prediction at different threshold values

In the second experiment, the pixel-based denoising algorithm was applied to the 106 ovarian lesion cases constituting the dataset, considering different values of the threshold that discriminates between artifacts and real doppler signal (referred to as signal to artifact threshold) in order to establish the one that led to better classification performances. The tested values were: 80th, 90th and 98th percentile of the distribution of the activation lengths of the pixels.
Similarly to what was done in the first experiment, the doppler pixel count values were firstly separated into four groups based on the color score and the box plot showing the doppler pixel counts vs color scores was obtained (see figure 4.4 for the box plot where the 90th percentile is used as threshold).

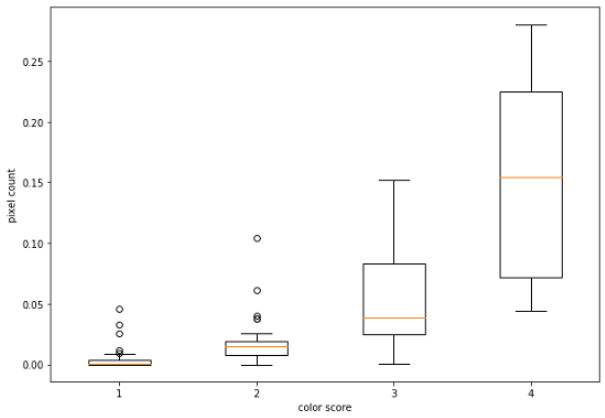


Figure 4.4: Box plot of doppler pixel count values for denoised videos vs color scores.

The box plot shows that the values of doppler pixel count clearly increase as the color score becomes higher, thus the four pixel count groups are distributed coherently to the color score values. Moreover, the box representing denoised videos having color score 1 is characterized by doppler pixel count values that are close to zero, suggesting that the pixel-based denoising algorithm is able to suppress a

significant portion of doppler artifacts identified within the lesion. However, the four boxes are still partially overlapped as in the previous experiment.

	test accuracy	
threshold	**mean**	**standard deviation**
80th percentile	0.56	0.06
90th percentile	0.62	0.08
98th percentile	0.51	0.09

Table 4.4: The table manifests the values of test mean accuracy and standard deviation obtained for the experiment conducted applying the pixel-based denoising at different threshold values. Here the 80th, 90th and 98th percentile of activation lengths' distribution were considered as thresholds.

The decision tree was trained and validated on the denoised videos for the different threshold values. The overall test accuracies and their corresponding standard deviations obtained for the 3 tested signal to artifact thresholds are shown in table 4.4. It is possible to notice that, employing the 90th percentile of the activation lengths' distribution as threshold, the highest accuracy, equal to 0.62 +/- 0.08, was reached (see table 4.5a for the accuracy values for each fold), thus the 90th percentile was chosen as final threshold. Meanwhile, the mean accuracy on the training set resulted equal to 0.69 +/- 0.01.

k	train accuracy	test accuracy
1	0.70	0.64
2	0.68	0.71
3	0.71	0.67
4	0.68	0.48
5	0.69	0.62

(a)

		Predicted labels			
		1	**2**	**3**	**4**
True labels	**1**	22	3	3	0
	2	9	14	5	0
	3	6	5	23	1
	4	0	0	8	7

(b)

Table 4.5: Classification performances for the second experiment conducted on denoised videos considering 4 labels, each one corresponding to one color score value. (a) Train and test accuracy values for each fold used in cross-validation of the model. (b) Cumulative confusion matrix on whole dataset. The tables are obtained using the 90th percentile of the activation lengths' distribution as signal to artifact threshold.

From the cumulative confusion matrix of table 4.5b, it emerges that, while almost all the samples having CS = 1 are correctly classified, the other three classes – (CS = 2, 3 and 4) have a high number of misclassified samples (14, 12 and 8

respectively) with respect to their numerosity.
In this context, in order to improve the classification performances, the predictive model was tuned by changing the values of the following parameters that characterize the decision tree:

- the criterion to measure the quality of a split,
- the weights associated with the classes,
- the strategy used to choose the split at each node,
- the minimum number of samples required to split a node,
- the maximum depth of the tree,
- the maximum number of leaf nodes.

However, no significant improvement of the classification performances occurred (see table 4.6).

Trial	train accuracy	test accuracy
start	0.69 +/- 0.01	0.62 +/- 0.08
1	0.66 +/- 0.06	0.59 +/- 0.06
2	0.62 +/- 0.06	0.52 +/- 0.06
3	0.64 +/- 0.03	0.66 +/- 0.06
4	0.69 +/- 0.01	0.62 +/- 0.08
5	0.69 +/- 0.01	0.62 +/- 0.08
6	0.69 +/- 0.01	0.62 +/- 0.08

Table 4.6: The table shows the accuracy values on train and test sets obtained testing different configurations of the decision tree on the denoised videos. The trial named start refers to the model whose characteristics are described in section 3.7.2. In trial 1 the criterion to measure the quality of a split was changed from gini to entropy. In trial 2 the weights associated with the classes were adjusted based on the proportion of each class frequencies. In trial 3 the best random split was chosen at each node. In trial 4 and 5, the minimum number of samples required to split a node and the maximum depth of the tree were respectively changed, they were set to all the integers ranging from 3 to 10. In trial 6, the maximum number of leaf nodes was set to all the integers ranging from 4 to 10. The tuning was performed employing the 90th percentile as signal to artifact threshold.

For a more accurate comparison with experiment 1, I merged CS = 2 and CS = 3 into a single, intermediate class, thus obtaining three classes: 1, 2-3, 4.

The resulting box plot where the doppler pixel counts are divided into the three groups basing on their color score is illustrated in figure 4.5. The boxes associated to color score 1 and class 2-3 appear well separated and not overlapped, and the pixel count groups are distributed coherently to the color scores. Moreover, the doppler pixel count values belonging to class 1 are nearly zero, thus the majority of the artifacts was suppressed by the algorithm.

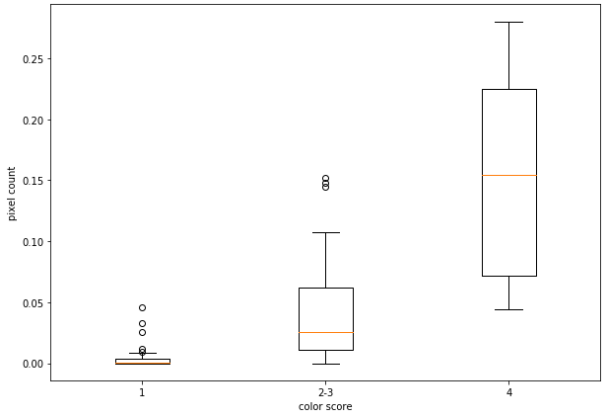


Figure 4.5: Box plot of doppler pixel count values for denoised videos vs the three labels (1, 2-3, 4).

At this point, the decision tree was trained and tested on the videos on which the pixel-based denoising with the 90th percentile as threshold was applied, being this the threshold with an associated higher accuracy (see table 4.4).

k	train accuracy	test accuracy
1	0.80	0.82
2	0.80	0.76
3	0.79	0.76
4	0.80	0.76
5	0.82	0.62

(a)

		Predicted labels		
		1	2-3	4
True labels	1	17	11	0
	2-3	7	55	1
	4	0	8	7

(b)

color score	sensitivity	specificity
1	0.61	0.91
2-3	0.87	0.56
4	0.47	0.99

(c)

Table 4.7: Classification performances for the second experiment conducted on denoised videos. (a) Train and test accuracy values for each fold used in cross-validation of the model. (b) Cumulative confusion matrix on whole dataset. (c) Sensitivity and specificity values for each of the three color score classes.

The resulting accuracy on training set and that on test set are 0.80 +/- 0.01 and 0.75 +/- 0.07 respectively (see table 4.7a for accuracy values obtained for the individual folds). Thus, treating the two intermediate color scores as a single class led to improved classification performances and this aspect emerges also from the

confusion matrix of table 4.7b and from the sensitivity and specificity values for each class shown in table 4.7c.
The high sensitivity of class 2-3 reflects the fact that, among the 63 samples belonging to this class, 55 are correctly classified, only 1 is classified as class 4 and the remaining 7 are assigned to class 1 by the model. The other two classes have lower sensitivity values because 11 samples among the 28 samples having CS = 1 and 8 among the 15 samples with CS = 4 are assigned by the model to another class (i.e. class 2-3). Concerning specificity, class 4 has the highest value (0.99), thus almost all the samples to which the model assigned CS = 4 actually belonged to this class. Instead, class 2-3 has a specificity of 0.56 because the majority of samples having color score 1 or 4 are wrongly associated to this class.
The resulting tree - whose complexity and dimensions were reduced thanks to the pruning process - is depicted in figure 4.6. From the tree, it can be seen that almost all the samples having CS = 2 or CS = 3 are assigned to class 2-3 by the model, this result justifies also the high value of sensitivity. Instead, class 1 and class 4 have 10 and 4 misclassified samples respectively, all of them assigned to class 2-3.

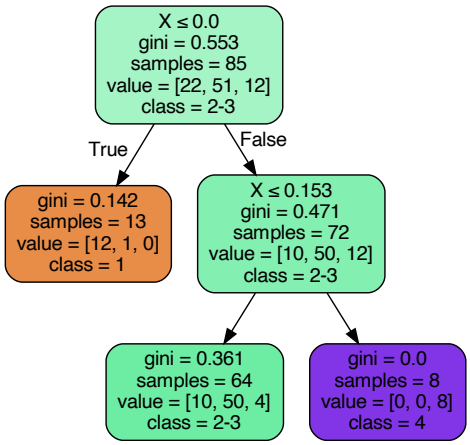


Figure 4.6: The figure shows the structure of the decision tree trained on denoised videos: class 1 node is depicted in orange, class 2-3 in green and class 4 in purple. Each node indicates the test that is performed on the samples at that node, the purity of the node defined by the Gini index, the number of samples that reach that node, the number of samples for each class, and the predominant class label.

Finally, also for this experiment, from the remaining leaf nodes were identified the two thresholds that divided the videos of the dataset into three groups:

- doppler pixel count = 0.0 → color score = 1

- 0.0 < doppler pixel count $\leq$ 0.153 $\rightarrow$ color score = 2 or 3
- doppler pixel count > 0.153 $\rightarrow$ color score = 4.

4.3 Experiment 3: Assessment of component-tracking denoising in color score prediction at different threshold values

In the third experiment, the color score predictive model was trained and tested on the videos output of the component-tracking denoising algorithm at different values of the signal to artifact threshold, the parameter that determinates how much the artifact-removal is conservative, which needed to be tuned. The optimal threshold value was chosen - among the 80th, 90th, 95th and 98th percentile of the distribution of the activation lengths of the objects tracked by the algorithm - as the one that led to a more accurate prediction of the color score.
The steps to perform this assessment are the same as those applied to the previous experiments.
In particular, initially the box plot showing the four doppler pixel count groups as a function of their color scores was obtained considering the different thresholds. In figure 4.7 there is the one obtained using the 95th percentile. Again, the four boxes result overlapped, meaning that there is not a range of doppler pixel count values that is representative of a single color score value, this is particularly true for color scores 2 and 3.

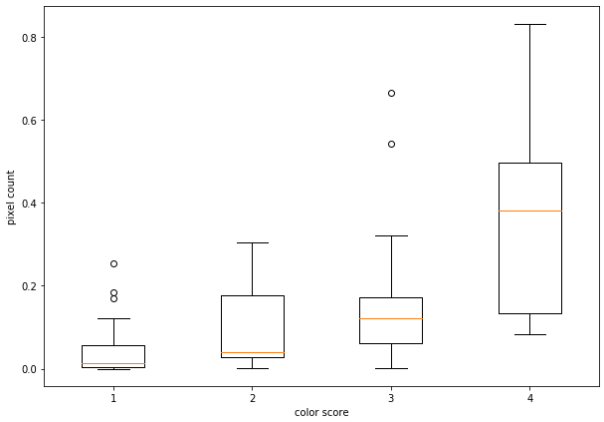


Figure 4.7: Box plot of doppler pixel count values for denoised + tracked videos vs color scores.

The accuracy values obtained after training and testing the decision tree on denoised+tracked videos at the different thresholds were not satisfactory, they

reached a maximum of 54% for 80th and 95th percentile and similar standard deviations (see table 4.8). This means that almost the 50% of ovarian cancer cases were wrongly classified by the model.

	test accuracy	
threshold	**mean**	**standard deviation**
80th percentile	0.54	0.07
90th percentile	0.49	0.06
95th percentile	0.54	0.08
98th percentile	0.52	0.09

Table 4.8: The table displays the values of mean accuracy and standard deviation obtained for the experiment conducted applying the component-tracking denoising at different threshold values. Here the 80th, 90th, 95th and 98th percentile of activation lengths' distribution were considered as thresholds.

In table 4.9a, there are the values of accuracy on train and test sets obtained for each of the 5 folds in which the dataset was divided during the cross-validation, considering 95th percentile as signal to artifact threshold. The corresponding cumulative confusion matrix, instead, is shown in table 4.9b and it is evident that class 2 is the most misclassified, with a total of 22 samples - among 28 - that are assigned by the model to classes different than class 2. Also the number of samples belonging to class 1 and wrongly associated to color score 3 is high (8).

k	train accuracy	test accuracy
1	0.62	0.55
2	0.59	0.67
3	0.62	0.48
4	0.52	0.43
5	0.61	0.57

(a)

		Predicted labels			
		1	**2**	**3**	**4**
True labels	**1**	20	0	8	0
	2	8	6	14	0
	3	9	1	23	2
	4	0	0	7	8

(b)

Table 4.9: Classification performances for the third experiment conducted on denoised+tracked videos considering 4 labels, each one corresponding to one color score value. (a) Train and test accuracy values for each fold used in cross-validation of the model. (b) Cumulative confusion matrix on whole dataset.

Also for this experiment, to improve the performances, I changed the values of the parameters that characterize the decision tree. The tuned parameters were: the criterion to measure the quality of a split, the weights associated with the classes,

the strategy used to choose the split at each node, the minimum number of samples required to split a node, the maximum depth of the tree and the maximum number of leaf nodes. However, as seen in the table 4.10, the classification performances did not substantially improve.

Trial	train accuracy	test accuracy
start	0.59 +/- 0.04	0.54 +/- 0.08
1	0.61 +/- 0.05	0.50 +/- 0.13
2	0.62 +/- 0.06	0.44 +/- 0.09
3	0.56 +/- 0.04	0.46 +/- 0.08
4	0.59 +/- 0.04	0.54 +/- 0.08
5	0.59 +/- 0.04	0.54 +/- 0.08
6	0.59 +/- 0.04	0.54 +/- 0.08

Table 4.10: The table shows the accuracy values on train and test sets obtained testing different configurations of the decision tree on the denoised+tracked videos. The trial named start refers to the model whose characteristics are described in section 3.7.2. In trial 1 the criterion to measure the quality of a split was changed from gini to entropy. In trial 2 the weights associated with the classes were adjusted based on the proportion of each class frequencies. In trial 3 the best random split was chosen at each node. In trial 4 and 5, the minimum number of samples required to split a node and the maximum depth of the tree were respectively changed, they were set to all the integers ranging from 3 to 10. In trial 6, the maximum number of leaf nodes was set to all the integers ranging from 4 to 10. The tuning was performed employing the 95th percentile as signal to artifact threshold.

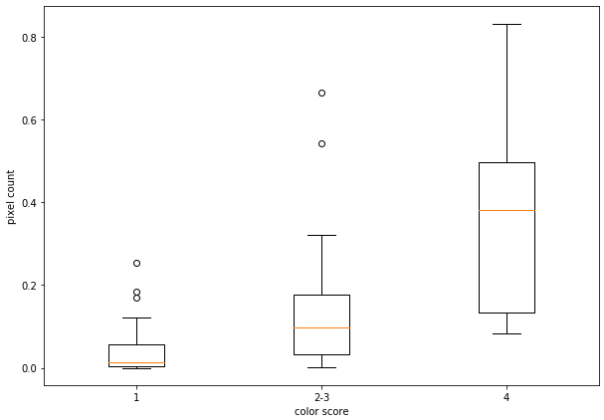


Figure 4.8: Box plot of doppler pixel count values for denoised+tracked videos vs the three labels (1, 2-3, 4).

For comparison with the previous experiments, the classification task was performed using only three labels instead of four, by putting color scores 2 and 3 into the same class, referred to as class 2-3.
By dividing the doppler pixel count values into three groups on the basis of the color score class (1, 2-3, 4), the doppler pixel count groups appear distributed coherently to the color score values with the higher doppler pixel counts correctly associated to higher color score values, as shown in figure 4.8.
Only 80th and 95th percentiles were tested as thresholds in this case because they were the ones that previously conducted to the highest accuracy.
The resulting accuracy on training set was equal to 0.74 +/- 0.02 and 0.75 +/- 0.02, while the cross-validation accuracy resulted 0.69 +/- 0.05 and 0.72 +/- 0.07 for 80th percentile and 95th percentile respectively. Therefore, the 95th percentile of the distribution of the components' activation lengths was chosen as optimal signal to artifact threshold due to the higher classification performance (see table 4.11a for accuracy values obtained in the 5 folds separately). The tables 4.11b and 4.11c illustrate the confusion matrix and both the sensitivity and specificity values respectively. It can be noticed that, firstly, class 4 has the lowest sensitivity due to the fact that 7 samples - among the 15 videos having color score equal to 4 – are assigned to class 2-3. On the contrary, this class has high specificity (0.98) because nearly all the videos that the model associated to this class actually have color score 4. Instead, the specificity of class 2-3 is only 0.58 because a high number of samples belonging to the other classes are assigned to the class 2-3 by the model. Regarding color score 1, 11 samples of this class are misclassified as class 2-3, thus resulting in a sensitivity of around 60%.
From these tables, it emerges that, grouping together class 2 and class 3 into a single label, the number of misclassified samples significantly decreased with respect to that obtained considering 4 separate classes of color score (30 vs 49 samples), and this brought to higher accuracy in color score prediction.

k	train accuracy	test accuracy
1	0.73	0.73
2	0.75	0.67
3	0.73	0.76
4	0.78	0.81
5	0.77	0.62

(a)

True labels \ Predicted labels	1	2-3	4
1	17	11	0
2-3	10	51	2
4	0	7	8

(b)

color score	sensitivity	specificity
1	0.61	0.87
2-3	0.81	0.58
4	0.53	0.98

(c)

Table 4.11: Classification performances for the third experiment conducted on denoised+tracked videos. (a) Train and test accuracy values for each fold used in cross-validation of the model. (b) Cumulative confusion matrix on whole dataset. (c) Sensitivity and specificity values for each of the three color score classes.

As seen from the figure 4.9, the final node associated to class 2-3 contains the highest number of samples belonging to the other two classes, coherently with the low specificity of this class. As a matter of fact, 6 samples with CS = 1 and 4 with CS = 4 are associated to the wrong class (class 2-3) by the tree, this confirms the low sensitivity values obtained for these labels.
Finally, also in this case, the two thresholds that separate the cases of adnexal masses into three classes were obtained:

- doppler pixel count $\leq 0.025 \rightarrow$ color score = 1
- $0.025 <$ doppler pixel count $\leq 0.351 \rightarrow$ color score = 2 or 3
- doppler pixel count $> 0.351 \rightarrow$ color score = 4

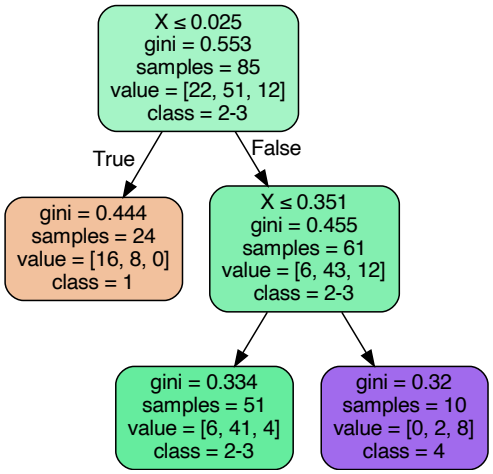


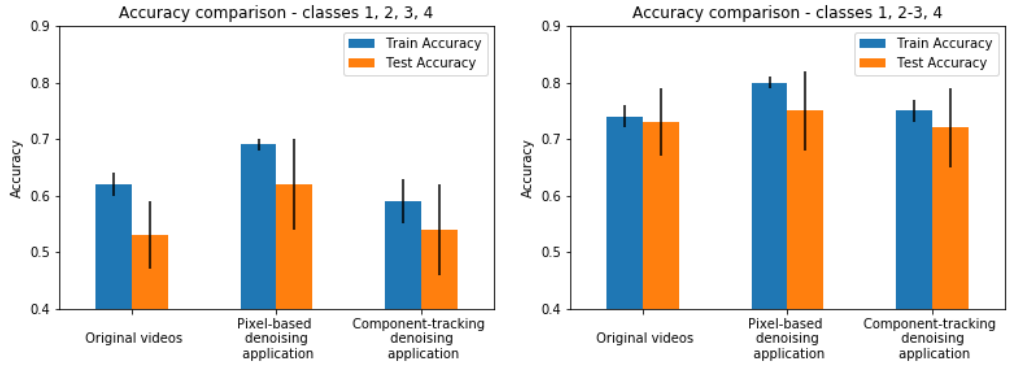
Figure 4.9: The figure shows the structure of the decision tree trained on denoised+tracked videos: class 1 node is depicted in orange, class 2-3 in green and class 4 in purple. Each node indicates the test that is performed on the samples at that node, the purity of the node defined by the Gini index, the number of samples that reach that node, the number of samples for each class, and the predominant class label.

4.4 Comparison of the experiments

I compared the results obtained in the three experiments conducted by training and testing the model on the original videos (Experiment 1), applying the pixel-based denoising (Experiment 2) and applying the component-tracking denoising

(Experiment 3).
From the boxplots obtained by dividing the doppler pixel count values into three groups based on the color score class, some important aspects emerged (see figures 4.2, 4.5 and 4.8).
Firstly, for the three box plots, the higher median values of the doppler pixel counts are associated to higher color scores, thus the three doppler pixel count groups are distributed coherently to the color score values. However, the groups result partially overlapped, in particular when considering the first and the third experiment, hence for some cases different color scores are associated to the same values of doppler pixel count making difficult for the model to predict the true color score based on this parameter.
Moreover, despite color score 1 being assigned to a lesion if no blood flow is present according to the IOTA guidelines, the doppler pixel count group associated to color score 1 includes values larger than zero when considering original videos and denoised+tracked videos. This was expected for the first experiment since no artifact-removal algorithm was applied to the original videos, thus doppler pixel count values also included the pixels that are colored due to the artifacts. However, the non-zero values appear also in the doppler pixel count group of CS = 1 obtained in the third experiment, meaning that the component-tracking denoising was not able to remove a significant portion of artifacts. Instead, applying the pixel-based denoising, almost all the artifacts were removed, and the doppler pixel count values for CS = 1 stand around zero, as expected.
In addition, the overlap between color score 1 and color score 2, that can be seen in the box plots of the first and the third experiment, is explained by the fact that CS = 2 is assigned if there is minimal flow within the lesion, thus samples with this color score may have very small doppler pixel count values. However, if there are artifacts in videos having CS = 1, similar values of doppler pixel count can be reached, so these doppler pixel counts for CS = 1 and CS = 2 overlap.
Furthermore, color scores 2 and 3 overlapped in the three cases. Distinguishing when the vascularization is minimal (CS = 2) and when it is moderate (CS = 3) is complex, qualitative and based on the clinician's perception. For this reason, I merged color score 2 and 3 in a single, intermediate, class for further analysis.
Eventually, in the three experiments the overlap between color score 3 and 4 occurred. According to some clinicians in contact with Syndiag, a possible explanation of this overlap is that clinicians tend to assign color score 4 to adnexal masses where the vascularization is equally distributed, and CS = 3 to the cases in which the amount of doppler signal is high but limited to a small portion of the lesion. However, the parameter used in this study to predict the color score (i.e., the doppler pixel count) does not consider the dispersion of colored pixels within the ROI, thus the same doppler pixel count can be associated to these two cases even if their color score is different, resulting in the overlap of CS = 3 and CS = 4.

The performances of the classifiers were also compared. The bar plot of figure 4.10b shows the values of train and test accuracy and their standard deviations resulting from the three experiments. From this figure, it can be noticed that all the classification performances are larger than the 70%, and the application of pixel-based denoising algorithm is associated to the higher accuracy both on training and test set. It is important to underline that the performances obtained for the three experiments increased when the intermediate color score values were considered as a unique class, with respect to the performances - shown in figure 4.10a - that were reached using 4 classes, one for each color score value.
Nevertheless, the experiments suggest that the pixel-based denoising algorithm generally improved classification performances with respect to original noisy videos and provided a more accurate estimate of doppler signal for color score assessment. Indeed in the second experiment, class 1 and class 2-3 have the highest number of correctly classified samples; while in the third experiment class 2-3 contains the highest number of misclassified samples (12) that actually have CS = 1 or CS = 4. Moreover 8 samples of class 4 are correctly classified (they are 7 in the first two experiments) and the number of samples of class 1 that are correctly classified is equal to that obtained applying the pixel-based denoising algorithm (see figures 4.7b and 4.11b).
In addition, the pixel-based denoising algorithm produced better results than the component-tracking one whose performances were comparable to the ones obtained on the original noisy videos, suggesting that this artifact-removal algorithm needs to be improved and further tuning of its parameters is required.



(a) **(b)**

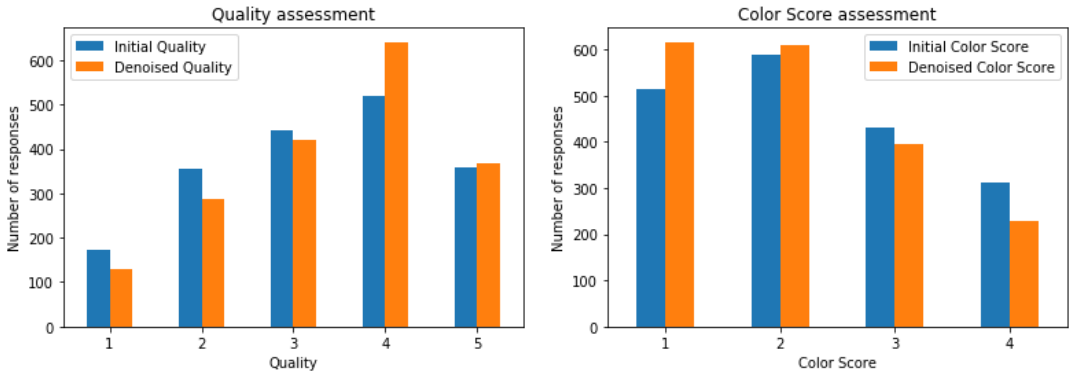
Figure 4.10: The bar plots show the accuracy values on training and test sets and their corresponding standard deviations obtained for the three conducted experiments. (a) shows the classification performances reached when the four color scores were used as labels. In (b) there are the results obtained considering three classes (1, 2-3 and 4) with color scores 2 and 3 constituting a single label.

4.5 Assessment of clinical impact

Moreover, we evaluated the impact that the denoising algorithms could have on the clinical practice as tools that support clinicians in performing the assessment of the ovarian lesions vascularization.
In the clinical practice, particular attention is given to CS = 1 and CS = 4. This is because the absence of vascularization within the mass, to which a CS = 1 is associated, is a sign of benignity, while an highly vascularized lesion, and thus a color score equal to 4, is characteristic of a malignant tumor, as stated in the IOTA's Simple Rules illustrated in section 1.2.2.
Therefore, when assigning the color score to a video, discriminating between when there is no blood flow within the lesion (color score equal to 1) and when the vascularization is minimal (CS = 2) is, at the same time, of outmost importance and complex for clinicians. In the same way, distinguishing between color score 4 (highly vascular mass) and 3 (moderate vascularization) becomes essential.
In this context, removing the artifacts through the application of denoising algorithms can simplify the work of clinicians who, this way, when assigning the color score have to distinguish only between the survived artifacts and the real signal.
As a matter of fact, as seen in tables 4.3c, 4.7c and 4.11c, class 1 has a sensitivity of 0.61 in the second and third experiment, and the sensitivity becomes 0.57 when the non processed videos are considered. This means that, when artifacts are partially removed, more videos having color score 1 are correctly associated to class 1 by the model, with respect to the original videos. Meanwhile, the sensitivity value of class 4 remains the same for the first two experiments (0.47) and increases to 0.53 in the third. This value is justified by the fact that, applying the component-tracking denoising, there are 8 videos with color score 4 that are correctly assigned to class 4 by the model, instead of 7. However, this sensitivity remains non satisfactory because there is still a high number of samples that are wrongly assigned to class 3. At the same time, class 4 maintains a high specificity value for the three experiments (0.96, 0.99, 0.98), because almost all the samples to which the model assigned CS = 4 actually belonged to this class. This occurs also for class 1, but with slightly lower values (0.94, 0.91, 0.87) since few samples of class 2-3 were assigned to class 1, underestimating the degree of vascularization within the mass and treating the real doppler activations as artifacts.

In this context, given the classification performances obtained for the three experiments, the corresponding specificity and sensitivity values evaluated for each class and the role of subjectivity in the color score assignment, applying the pixel-based denoising on the noisy ovarian cancer cases can be useful to suppress a good portion of artifacts that occur within the lesion, allowing clinicians to more clearly assign the color score by reducing the risk of confusion between artifacts and real signal.

This consideration is confirmed by the results of a study conducted by the Syndiag team in collaboration with Mauriziano and Sant'Orsola hospitals. In this study, 20 clinicians with different levels of experience (6 expert clinicians, 5 clinicians with less than 10 years of experience and 9 juniors) were asked to evaluate 100 videos coming from 100 unique clinical cases of ovarian lesions before the pixel-based denoising was applied and, then, evaluate the same videos after the application of the denoising algorithm. The videos were randomly presented to the doctors. The clinicians' evaluation of the videos was performed through the assignment of color score and of an index ranging from 1 to 5 indicating the quality of the analyzed video, with 1 meaning poor quality and 5 optimal quality.



(a) **(b)**

Figure 4.11: Results of the study performed to assess the clinical impact of the pixel-based denoising algorithm. (a) The bar plot shows the number of clinicians' responses vs the videos' quality before (blue bars) and after (orange bars) the denoising application. The quality index is 1 when the quality is poor, 2 when it is not sufficient, 3 if it sufficient, 4 if it is decent, 5 if the quality is optimal. (b) The bar plot shows the number times that clinicians assigned a certain color score before (blue bars) and after (orange bars) the denoising application.

As seen from figure 4.11a, the quality of the videos perceived by clinicians increases when denoising is applied since the number of responses equal to 1 and 2 is decreased while a higher number of responses where quality = 4 and 5 was registered.
Regarding the color score assignment, instead, a relevant trend emerges: when the pixel-based denoising was applied, a lower number of CS = 4 responses was given by clinicians, while they tended to assign more often a color score equal to 1 (see figure 4.11b). This proves the presence of a clinical impact due to the application of this denoising algorithm. Specifically, these results suggest that, when artifacts were not removed, clinicians tended to assign higher color score values. Therefore, this study demonstrates that the presence of doppler artifacts

influences the assignment of color score and the clinicians' evaluation.
On the other hand, the component-tracking denoising algorithm, as it is today, has lower performance and possibly a lower impact on the clinical practice. Indeed, when applied to the 106 videos of the dataset, a number of artifacts still survives, thus further improvements are needed.

Chapter 5

Conclusions, limitations and future developments

Ovarian cancer is the eighth most common cancer in women worldwide and it is a very aggressive type of tumor characterized by a mortality of 4.3%. Indeed, in Europe, it is the main cause of death among gynecologic malignancies [1, 2].
The low survival rate associated to this neoplasm is explained by the difficulty of its diagnosis at early stages since it is usually asymptomatic, or symptoms are not specific. Ovarian cancer is diagnosed at advanced stages in about 70% of cases, and its survival rate after 5 years is lower than 30%. In contrast, if the ovarian cancer is detected early the survival becomes longer than 5 years for more than 90% of patients [2, 9].
Therefore, detecting the adnexal masses as soon as possible in combination with an adequate discrimination between benign and malignant lesions is of outmost importance for a correct management of the patient.
Nowadays, ultrasound represents the method of choice to evaluate the adnexal masses thanks to its non-invasiveness, low-cost and widely diffusion. Moreover, the advent of Color Doppler and Power Doppler imaging techniques allowed clinicians to evaluate the lesions' vascularization that is an important indicator of malignancy.
In this context, a group of researchers founded the International Ovarian Tumor Analysis (IOTA) group with the aim of producing a standardized terminology that characterizes adnexal masses and developing diagnostic tools in ultrasound for the prediction of malignancy.
Regarding vascularization, among the definitions introduced by IOTA, the color score is particularly interesting since it is a scoring system, ranging from 1 to 4, employed to assess the amount of blood flow within the septa, cyst walls or solid tumor area. The color score is assigned by clinicians to the lesion and can assume the following values: 1 if no blood flow is shown in the mass, 2 if the flow is minimal,

3 when moderate flow is present, 4 is assigned to highly vascular lesions [4].
The color score was proven to be a good predictor of malignancy and it has been included in several models that assess the probability that a lesion is malignant or benign (the IOTA's Simple Rules for instance). However, the main issue is that the estimation of the color content within an adnexal mass is based on the subjective evaluation of clinicians.
Moreover, Doppler techniques are characterized by the presence of several types of artifacts that make color score assignment even more complex for clinicians because the flow information becomes more difficult to interpret.

Consequently, two algorithms were developed in order to remove artifacts from doppler signal with the aim of easing clinicians' assessment of adnexal lesions. The first is the pixel-based denoising algorithm that suppresses doppler artifacts on the basis of the temporal persistence of doppler activations during the whole video, pixel by pixel; assuming that artifacts are less persistent than real activations. However, this algorithm has the limit of relying only on the exact pixel correspondence in a frame sequence, thus some real activations are identified as artifacts because they shift through frames due to the movement of the probe during the acquisitions.
To overcome this limitation, a second artifact-removal algorithm was developed: the connected components-based denoising with component-tracking that considers both temporal and spatial persistence and relies on connected components rather than single pixels. In this case, the component tracking algorithm is integrated with the denoising giving rise to an algorithm that suppresses artifacts while keeping track of the activations' clusters during the video.
These two algorithms were developed on 101 ovarian cancer cases resulting from the acquisitions performed by expert clinicians of two hospitals: A.O. Ordine Mauriziano in Turin and Policnico di Sant'Orsola in Bologna.
During the development of the component-tracking denoising several edge cases were identified and treated during this study. The solutions that were implemented to address the identified edge cases are:

- A method that excludes the pixels of the doppler fan that, for some data, resulted colored and, therefore, it was recognized by the component-tracking algorithm as an object to be tracked.

- A function to be integrated in the component-tracking algorithm that solves the problem of overlapping between a large artifact and real activations that, due to the overlap, were not correctly tracked.

- A function to be integrated in the component-tracking algorithm that manages the presence of empty binary masks, i.e. masks where all the pixels are set to 0 and, thus, no object is identified by the algorithm.

Moreover, the tuning of the tolerance values employed in the component-tracking algorithm was performed.
Afterwards, the updated version of the component tracking algorithm was integrated in the denoising and the effect of the two artifact-removal algorithms (i.e., pixel-based denoising and component-tracking denoising) was evaluated.

In order to quantitatively assess the performances of the artifact-removal algorithms, a Decision Tree was trained to predict the color score based on the doppler estimation obtained on both original and denoised doppler videos included in a second dataset. This dataset contained 106 ovarian cancer cases and the corresponding color scores, used as labels, were assigned by 6 expert clinicians. The resulting dataset reflected the distribution of malignancy of this tumor in the clinical practice, with only 15 cases having color score 4, while 28 videos with color score 1 and 2 and 35 videos having color score equal to 3 were selected.
To estimate the amount of doppler signal within the lesion, i.e. the input to the decision tree, the videos of the dataset were manually segmented to identify the region of interest (i.e. the intersection between the mass and the doppler acquisition fan). The doppler pixel count was calculated as the ratio between the number of colored pixels within the ROI and the area of this region.
The model was trained on the original noisy videos in a first experiment, applying the pixel-based denoising at different signal to artifact threshold values in a second, and applying the component-tracking denoising at different thresholds in a third. The results of the three experiments were compared.

The resulting cross-validation accuracies obtained when the Decision Tree model was tested were equal to 53% (standard deviation = +/- 6%), 62% (standard deviation = +/- 8%) and 54% (standard deviation = +/- 8%) for the first, second and third experiment respectively. These classification performances were reached using the optimal values of signal to artifact thresholds coinciding with the 90th percentile and the 95th percentile of the distribution of the activations' lengths in the pixel-based denoising and the component-tracking denoising algorithms respectively.
However, in almost the half of the ovarian cancer cases the classifier was not able to correctly predict the color score.
Tuning of the decision tree did not improve the performance, but a closer look at the data revealed that the doppler pixel count groups associated to scores 2 and 3 resulted strongly overlapped. I consequently merged the intermediate values of color score into a single class and trained the Decision Tree model using three labels (1, 2-3 and 4). Another reason that brought to this choice concerns the definition of color score and its relevance from the diagnostic point of view. Scores 1 and 4 are the ones with a clearer diagnostic value because, when assigned to a lesion,

they are characteristic of benignity and malignancy respectively, as indicated by IOTA's Simple Rules 1.2.2.
Using three labels, the dataset consisted of 28 videos having color score 1, 63 samples with color score 2 or 3 and 15 videos with color score equal to 4.
In this way, the model showed good performances in the three experiments where overall accuracy values of 73%, 75% and 72% were reached on original, denoised and denoised+tracked videos respectively. In the three cases, the class 2-3 had the highest sensitivity (above 80%), while the lowest sensitivity of approximately 50% occurred for color score class 4 with a high number of samples (7 or 8, depending on the considered experiment, among the 15 total samples) being assigned to color score 3.
Finally, what emerges from results obtained using three and four labels is that, in this study, applying the pixel-based denoising algorithm improved the classification performances with respect to original noisy videos. On the contrary, the accuracy of the model trained after denoising with tracking algorithm is comparable to the one obtained when the model was trained on original noisy videos.

These results suggest that further studies and tuning are needed to improve and optimize the tracking approach.
In particular, the analysis, described in section 3.5.5, where several parameters of size and shape of single connected components were calculated for each frame of the video, could be further deepened in order to find values of these parameters useful to distinguish between when the analyzed connected component includes only the real signal and when it includes both the real activation and the artifact.
Moreover, the function that was introduced in the tracking algorithm to manage the presence of large artifacts overlapping with real activations (see section 3.5.3) solved this problem - and thus allowed the algorithm to keep track of the real activations that were correctly distinguished from each other and from the artifact - only if the artifact was composed by a single connected component of large dimensions. In the future, this overlapping problem needs to be solved also when the artifact is composed of several small connected components, so that all the types of artifacts can be suppressed.
Other future developments involve the model's construction.
In particular, it must be considered that clinicians might assign the color score not only on the basis of the amount of doppler signal within the lesion (that is a quantitative and objectively measurable parameter whose estimation is given by the pixel count) but also taking into account other factors that are not considered by the implemented model, such as the dispersion of the real signal within the region of interest, the number of ramifications of the vessels and their localization. Therefore, in the future, these factors can be included in the model as new features, together with the doppler pixel count, to be used to estimate the doppler signal in

order to improve the color score prediction. This improvement can be particularly useful in contributing to reduce the dependence of the model from the employed dataset.
In addition, it would be interesting to conduct the three experiments training a random forest (i.e. a classification algorithm that combines the output of multiple decision trees to reach a single result), instead of using a single decision tree, in order to see if this model brings to more accurate color score predictions.
Moreover, due to the low numerosity of the dataset and its imbalance – which represent strong limitations – the performances of the model decrease a lot if few samples are removed from the dataset or few label values are changed, despite the application of the pruning procedure. Beside this, the starting labels result particularly noisy due to the fact that the it is difficult for clinicians to assign the color score to the adnexal lesions. Therefore, including more videos with color score 4 and, in general, increasing the number of ovarian cancer cases to include in the dataset would bring to more robust performances and cushion the noise.
Furthermore, it must be also considered that clinicians, when performing the acquisitions of doppler videos, often focus not on the whole adnexal mass but only on the portion of the mass in which they are interested in. Therefore, the dataset results biased from this point of view, with the bias being the assumption of what part of the adnexal mass should be examined according to the clinician. This means that, when the six expert clinicians had to assign the color score to the videos, they were influenced by this choice.
Finally, another relevant step to be performed is increasing the number of expert clinicians who are going to evaluate the original videos and assign them a color score, and, afterwards, asking them to evaluate also the videos on which the pixel-based denoising is applied and the ones where the component-tracking denoising is applied. This way, it is possible to see whether clinicians are effectively helped in their evaluation by the artifact-removal algorithms and measure their agreement in color score assignment when the two algorithms are applied on the noisy videos in order to understand if it increases or not with respect to when non processed videos are evaluated.

List of Figures

List of Tables

Bibliography

[1] P. V. Foti, G. Attinà, S. Spadola, et al. «MR imaging of ovarian masses: classification and differential diagnosis». In: *Insights Imaging* 7 (2016), pp. 21–41. DOI: 10.1007/s13244-015-0455-4 (cit. on pp. 1–8, 34, 105).

[2] N. Sehgal. «Efficacy of Color Doppler Ultrasonography in Differentiation of Ovarian Masses». In: *J Mid-life Health* 10 (2019), pp. 22–28. DOI: 10.4103/jmh.JMH_23_18 (cit. on pp. 1, 2, 8, 34, 105).

[3] B. Van Calster, K. Van Hoorde, L. Valentin, et al. «Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study». In: *BMJ* (2014). DOI: 10.1136/bmj.g5920 (cit. on pp. 1–3).

[4] D. Timmerman, L. Valentin, T. H. Bourne, 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 (2000), pp. 500–505. DOI: 10.1046/j.1469-0705.2000.00287.x (cit. on pp. 1, 5, 8–14, 34, 43, 83, 106).

[5] D. M. Twickler and E. Moschos. «Ultrasound and assessment of ovarian cancer risk». In: *AJR Am J Roentgenol.* 194 (2009), pp. 322–329. DOI: 10.2214/AJR.09.3562 (cit. on pp. 1, 13, 14).

[6] M. A. Pozniak, J. A. Zagzebski, and K. A. Scanlan. «Spectral and color Doppler artifacts.» In: *RadioGraphics* 12 (1992), pp. 35–44. DOI: 10.1148/radiographics.12.1.1734480 (cit. on pp. 1, 19, 21, 23–27).

[7] R. Massobrio. «Diagnosi ultrasonografica delle masse annessiali: studio della concordanza nell'interpretazione delle immagini, con applicazione della terminologia IOTA». MA thesis. Università degli studi di Torino, scuola di Medicina, 2018 (cit. on pp. 2, 31).

[8] World Cancer Research Fund International. *Worldwide cancer data.* URL: https://www.wcrf.org/cancer-trends/worldwide-cancer-data/ (visited on 11/30/2022) (cit. on p. 2).

[9] S. Wei, H. Li, and B. Zhang. «The diagnostic value of serum HE4 and CA-125 and ROMA index in ovarian cancer». In: *Biomedical reports* 5 (2016), pp. 41–44. DOI: 10.3892/br.2016.682 (cit. on pp. 2, 3, 15, 34, 105).

[10] National Cancer Institute. *Definition of adnexal mass*. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/adnexal-mass (visited on 09/13/2022) (cit. on p. 2).

[11] Merriam-Webster. *Medical Definition of etiopathogenesis*. URL: https://www.merriam-webster.com/medical/etiopathogenesis (visited on 08/25/2022) (cit. on p. 3).

[12] V. W. Chen, B. Ruiz, J. L. Killen, et al. «Pathology and classification of ovarian tumors». In: *ACS Journals* 97 (2003), pp. 2631–2642. DOI: 10.1002/cncr.11345 (cit. on pp. 4, 7).

[13] Ovarian Cancer Research Alliance (OCRA). *Types of Ovarian Cancer*. URL: https://ocrahope.org/patients/about-ovarian-cancer/types-ovarian-cancer/#:~:text=How%5C%20many%5C%20types%5C%20of%5C%20ovarian,that%5C%20make%5C%20up%5C%20the%5C%20ovary. (visited on 09/13/2022) (cit. on p. 4).

[14] M. Gil-Martin, B. Pardo, and M. P. Barretina-Ginesta. «Rare ovarian tumours. Other treatments for ovarian cancer». In: *European Journal of Cancer Supplements* 15 (2020), pp. 96–103. DOI: 10.1016/j.ejcsup.2019.11.002 (cit. on p. 5).

[15] International Ovarian Tumor Analysis. *Materials*. URL: https://iota.education/educational-materials/ (visited on 08/29/2022) (cit. on pp. 5, 9–17).

[16] International Ovarian Tumor Analysis. *About IOTA*. URL: https://www.iotagroup.org/welcome/about-iota (visited on 08/29/2022) (cit. on p. 9).

[17] D. Timmerman, F. Planchamp, T. Bourne, et al. «ESGO/ISUOG/IOTA/ESGE Consensus Statement on preoperative diagnosis of ovarian tumors». In: *Ultrasound Obstet Gynecol* 58 (2021), pp. 148–168. DOI: 10.1002/uog.23635 (cit. on pp. 9, 14–16, 18, 34).

[18] A. M. Abbas, K. M. Zahran, A. Nasr, et al. «A new scoring model for characterization of adnexal masses based on two-dimensional gray-scale and colour Doppler sonographic features». In: *Facts Views Vis Obgyn.* 6 (2014), pp. 68–74 (cit. on pp. 16, 33).

[19] L. Ameye, D. Timmerman, L. Valentin, et al. «Clinically oriented three-step strategy for assessment of adnexal pathology». In: *Ultrasound in Obstetrics & Gynecology* 40 (2012), pp. 582–591. DOI: 10.1002/uog.11177 (cit. on p. 16).

[20] D. Timmerman, A. C. Testa, T. Bourne, et al. «Logistic Regression Model to Distinguish Between the Benign and Malignant Adnexal Mass Before Surgery: A Multicenter Study by the International Ovarian Tumor Analysis Group». In: *Journal of Clinical Oncology* 23 (2005), pp. 8794–8801. DOI: 10.1200/JCO.2005.01.7632 (cit. on p. 16).

[21] International Ovarian Tumor Analysis. *IOTA Simple Rules and SRrisk calculator to diagnose ovarian cancer*. URL: https://www.iotagroup.org/research/iota-models-software/iota-simple-rules-and-srrisk-calculator-diagnose-ovarian-cancer (visited on 08/29/2022) (cit. on pp. 16, 18).

[22] B. Van Calster, K. Van Hoorde, W. Froyman, et al. «Practical guidance for applying the ADNEX model from the IOTA group to discriminate between different subtypes of adnexal tumors». In: *Facts Views Vis Obgyn.* 7 (2015), pp. 32–41 (cit. on pp. 18, 19, 30).

[23] D. H. Evans, J. A. Jensen, and M. B. Nielsen. «Ultrasonic colour Doppler imaging». In: *Interface Focus* 1 (2011), pp. 490–502. DOI: 10.1098/rsfs.2011.0017 (cit. on pp. 19–22).

[24] Radiopaedia. *Doppler shift*. URL: https://radiopaedia.org/articles/doppler-shift (visited on 09/21/2022) (cit. on pp. 20, 21).

[25] W. N. McDicken and T. Anderson. «The difference between Colour Doppler Velocity Imaging and Power Doppler Imaging». In: *Eur. J. Echocardiogr.* 3 (2002), pp. 240–244. DOI: 10.1053/euje.2002.0150 (cit. on pp. 21, 22, 24, 27, 28).

[26] J. M. Rubin. «Power Doppler». In: *European Radiology* 9 (1999), pp. 318–322. DOI: 10.1007/pl00014064 (cit. on pp. 22, 27–29).

[27] D. J. Rubens, S. Bhatt, S. Nedelka, et al. «Doppler Artifacts and Pitfalls». In: *Radiologic Clinics* 44 (2006), pp. 805–835 (cit. on pp. 23–27).

[28] Radiopaedia. *Blooming artifact (ultrasound)*. URL: https://radiopaedia.org/articles/blooming-artifact-ultrasound?lang=us#image_list_item_44335068 (visited on 09/21/2022) (cit. on p. 23).

[29] «Artefatti Doppler». In: 8 (2008), pp. 313–325 (cit. on pp. 25, 26).

[30] Radiopaedia. *Twinkling artifact*. URL: https://radiopaedia.org/articles/twinkling-artifact?lang=us#image_list_item_10771401 (visited on 09/21/2022) (cit. on p. 25).

[31] L. Wynants, D. Timmerman, T. Bourne, et al. «Screening for data clustering in multicenter studies: the residual intraclass correlation». In: *BMC Med Res Methodol.* 13 (2013). DOI: 10.1186/1471-2288-13-128 (cit. on p. 30).

[32] L. Zannoni, L. Savelli, L. Jokubkiene, et al. «Intra- and interobserver reproducibility of assessment of Doppler ultrasound findings in adnexal masses». In: *Ultrasound in Obstetrics & Gynecology* 42 (2013), pp. 93–101. DOI: 10.1002/uog.12324 (cit. on p. 30).

[33] A. C. Fleischer, W. H. Rodgers, D. M. Kepple, et al. «Color Doppler sonography of ovarian masses: a multiparameter analysis». In: *J Ultrasound Med.* 12 (1993), pp. 41–48. DOI: 10.7863/jum.1993.12.1.41 (cit. on p. 32).

[34] S. M. Stein, S. Laifer-Narin, M. B. Johnson, et al. «Differentiation of Benign and Malignant Adnexal Masses: Relative Value of Gray-Scale, Color Doppler, and Spectral Doppler Sonography». In: *AJR Am J Roentgenol.* 164 (1995), pp. 381–386. DOI: 10.2214/ajr.164.2.7839975 (cit. on p. 32).

[35] D. Timmerman, T. H. Bourne, A. Tailor, et al. «A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: The development of a new logistic regression model». In: *Am J Obstet Gynecol.* 181 (1999), pp. 57–65. DOI: 10.1016/s0002-9378(99)70436-9 (cit. on p. 32).

[36] L. Ameye, L. Valentin, A. C. Testa, et al. «A scoring system to differentiate malignant from benign masses in specific ultrasound-based subgroups of adnexal tumors». In: *Ultrasound Obstet Gynecol.* 33 (2009), pp. 92–101. DOI: 10.1002/uog.6273 (cit. on p. 33).

[37] C. Van Holsbeke, B. Van Calster, L. Valentin, et al. «External Validation of Mathematical Models to Distinguish Between Benign and Malignant Adnexal Tumors: A Multicenter Study by the International Ovarian Tumor Analysis Group». In: *Clin Cancer Res.* 13 (2007), pp. 4440–4447. DOI: 10.1158/1078-0432.CCR-06-2958 (cit. on p. 33).

[38] Anaconda. *ANACONDA DISTRIBUTION*. URL: https://www.anaconda.com/products/distribution (cit. on p. 43).

[39] Charles R. Harris et al. «Array programming with NumPy». In: *Nature* 585 (2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2 (cit. on p. 43).

[40] Wes McKinney et al. «Data structures for statistical computing in python». In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. 2010, pp. 51–56 (cit. on p. 43).

[41] J. D. Hunter. «Matplotlib: A 2D graphics environment». In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55 (cit. on p. 43).

[42] F. Pedregosa et al. «Scikit-learn: Machine Learning in Python». In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 43).

[43] Almar Klein et al. *imageio/imageio: v2.22.4*. Version v2.22.4. Nov. 2022. DOI: 10.5281/zenodo.7297715. URL: https://doi.org/10.5281/zenodo.7297715 (cit. on p. 43).

[44] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. «scikit-image: image processing in Python». In: *PeerJ* 2 (2014), e453 (cit. on pp. 43, 60).

[45] G. Bradski. «The OpenCV Library». In: *Dr. Dobb's Journal of Software Tools* (2000) (cit. on p. 43).

[46] Pauli Virtanen et al. «SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python». In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2 (cit. on pp. 43, 54).

[47] Guido Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020 (cit. on p. 43).

[48] RedBrick AI. *Collaborative Rapid Medical Data Annotation*. URL: https://redbrickai.com/ (cit. on pp. 43, 73).

[49] F. De Simone. «Doppler Flow Imaging in support of ovarian tumor diagnosis: automated denoising and evaluation of color score in accordance with IOTA guidelines». MA thesis. Politecnico di Torino,Master of Science in Biomedical Engineering, 2020 (cit. on pp. 43, 44).