Cancer Detection in Full Field Optical Coherence Tomography Images

Diana Mandache

Thèse CIFRE préparée à LLTech et Institut Pasteur de Paris

delivrée par Sorbonne Université

École Doctorale Informatique, Télécommunications, Électronique de Paris n°130

Unité d'Analyse d'Images Biologiques

Cancer Detection in Full Field Optical Coherence Tomography Images

Détection du cancer dans les images de tomographie par cohérence optique plein champ

Par Diana MANDACHE

Thèse de doctorat d'Informatique

Dirigée par Jean-Christophe OLIVO-MARIN

et co-encadrée par Vannary MEAS-YEDID

Présentée et soutenue publiquement le 7 décembre 2022

Devant un jury composé de

Christine DECAESTECKER	Directrice de recherche, Université Libre de Bruxelles	Examinatrice
Petr DOKLADAL	Chargé de recherche, MINES Paris	Rapporteur
Auguste GENOVESIO	Directeur de recherche, ENS	Examinateur
Nedra MELLOULI-NAUWYNCK	Maîtresse de conférences, Université Paris 8	Rapportrice
Vannary MEAS-YEDID	Ingénieur de recherche, Institut Pasteur	Co-encadrante
Jean-Christophe OLIVO-MARIN	Directeur de recherche, Institut Pasteur	Directeur de thèse

What is real is not the appearance, but the idea, the essence of things.

- Constantin Brâncuși

Abstract

Cancer is a leading cause of death worldwide making it a major public health concern. Despite its growing incidence, this pathology still has many unknowns subject to an active research. Different biomedical imaging techniques accompany both research (cell biology) and clinical (screening, diagnosis, treatment) efforts towards improving patient outcome.

In this work we explore the use of a new family of imaging techniques, static (FFOCT) and dynamic (DCI) full field optical coherence tomography. Based on the principle of light coherence, they provide an image of the micro-architecture and cellular activity of the tissue in depth without any preparation. Therefore, they allow for a much faster analysis of the tissue compared to the gold standard histopathology. However, their novel and unique contrast - different from common techniques, makes them difficult to adopt in clinical settings.

In order to facilitate the interpretation of FFOCT and DCI images, we develop several aid-to-diagnosis algorithms by exploiting the data collected from clinical studies. We employ suitable data curation techniques to build training sets fitted for data-driven methods. We propose an analytical method for a better characterization and part-based separation of the raw dynamic interferometric signal, as well as multiple diagnostic support methods for FFOCT / DCI images based on deep learning. Accordingly, convolutional neural networks were designed and trained under various paradigms: (i) fully supervised learning, whose prediction capability surpasses the pathologist performance; (ii) multiple instance learning, which accommodates the lack of expert annotations; (iii) contrastive learning, which exploits the multi-modality of the data. Moreover, we highly focus on method validation and we decode the trained "black box" models to obtain an interpretation of the reasoning behind the predicted diagnostic. This ensures the good generalization of the models and ultimately leads to finding specific biomarkers in our images. This thesis presents an ensemble of original methods with significant results dedicated to FFOCT / DCI data representation, a pioneering endeavor which sets the grounds for further research directions.

Keywords: full field optical coherence tomography, convolutional neural networks, representation learning, digital pathology, cancer detection, aid to diagnosis

Résumé

Le cancer est une des principales cause de décès dans le monde, ce qui fait de lui un problème majeur de santé publique. Malgré son incidence croissante, cette pathologie comporte encore de nombreuses inconnues faisant l'objet d'une recherche active. Différentes techniques d'imagerie bio-médicale accompagnent les efforts de la recherche (biologie cellulaire) et de la pratique clinique (dépistage, diagnostic, traitement) pour améliorer le pronostic des patients.

Dans ce travail, nous étudions l'utilisation d'une nouvelle famille de techniques d'imagerie, la tomographie par cohérence optique plein champ statique (FFOCT) et dynamique (DCI). Reposant sur le principe d'interférométrie à faible cohérence optique, elles ont l'avantage de fournir une image de la microarchitecture et du contenu cellulaire du tissu en profondeur sans aucune préparation du tissu. Cela permet une analyse beaucoup plus rapide du tissu par rapport à la technique de référence en histopathologie. Cependant, leur contraste novateur et singulier, différent des techniques classiques, rendent difficiles leur adoption en milieu clinique.

Afin de faciliter l'interprétation des images FFOCT et DCI, nous développons plusieurs algorithmes d'aide au diagnostic en exploitant les données recueillies lors d'études cliniques. Nous utilisons des techniques appropriées de gestion des données pour créer des ensembles d'entraînement adaptés aux méthodes pilotées par les données. Nous proposons une méthode analytique pour une meilleure caractérisation et séparation des signaux interférométriques dynamiques bruts, ainsi que de multiples méthodes d'aide au diagnostic pour les images FFOCT / DCI basées sur l'apprentissage profond. Pour cela, des réseaux de neurones convolutifs ont été conçus et entraînés en utilisant différents paradigmes: (i) l'apprentissage entièrement supervisé, dont la capacité de prédiction dépasse la performance du pathologiste; (ii) l'apprentissage à instances multiples, qui permet de surmonter le manque d'annotations d'experts; (iii) l'apprentissage contrastif, qui exploite la multi-modalité des données. En outre, nous portons une attention particulière à la validation des méthodes et nous déchiffrons les modèles "boîte noire" entraînés afin d'obtenir une interprétation du raisonnement derrière le diagnostic prédit. Ceci garantit une bonne généralisation des modèles et conduit finalement à la découverte de biomarqueurs spécifiques dans nos images. Cette thèse constitue un ensemble de méthodes originales avec des résultats significatifs dédiés à la représentation des données FFOCT / DCI, un effort pionnier qui pose aussi les bases de nouveaux développements de recherche.

Mots clés: tomographie par cohérence optique plein champ, réseaux de neurones convolutifs, apprentissage de représentations, pathologie numérique, détection du cancer, aide au diagnostic

List of Publications

[Mandache2023] **D. Mandache**, E. Benoit á la Guillaume, J-C. Olivo-Marin and V. Meas-Yedid, *Cross-Modal Contrastive Learning for Robust Representation of the Extracellular Matrix in Static and Dynamic Full-Field OCT Images*, IEEE International Symposium on Biomedical Imaging (ISBI), Cartagena de Indias, Colombia, 2023. (submitted)

[Mandache2022] **D. Mandache**, E. Benoit á la Guillaume, Y. Badachi, J-C. Olivo-Marin and V. Meas-Yedid, *The Lifecycle of a Neural Network in the Wild: a Multiple Instance Learning Study on Cancer Detection from Breast Biopsies Imaged with Novel Technique*, IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022. DOI: 10.1109/ICIP46576.2022.9897596.

[Mandache2021b] **D. Mandache**, E. Benoit á la Guillaume, M-C. Mathieu, J-C. Olivo-Marin and V. Meas-Yedid, *Leveraging Global Diagnosis for Tumor Localization in Dynamic Cell Imaging of Breast Cancer Tissue Towards Fast Biopsying*, IEEE International Symposium on Biomedical Imaging (ISBI), Nice, France, 2021. DOI: 10.1109/ISBI48211.2021.9434110.

[Mandache2021a] **D. Mandache**, E. Benoit á la Guillaume, J-C. Olivo-Marin, V. Meas-Yedid, *Blind Source Separation in Dynamic Cell Imaging Using NonNegative Matrix Factorization Applied to Breast Cancer Biopsies*, IEEE International Symposium on Biomedical Imaging (ISBI), Nice, France, 2021. DOI: 10.1109/ISBI48211.2021.9434128.

[Gonzalez2019] D. Gonzalez, **D. Mandache**, J-C. Olivo-Marin, V. Meas-Yedid, *Icytomine: A User-Friendly Tool for Integrating Workflows on Whole Slide Images*, European Congress on Digital Pathology (ECDP), Warwick, UK, 2019. DOI: 10.1007/978-3-030-23937-4_21.

[Mandache2018] **D. Mandache**, E. Dalimier, J. Durkin, A. C. Boccara, J-C. Olivo-Marin and V. Meas-Yedid, *Basal Cell Carcinoma Detection in Full Field OCT images using Convolutional Neural Networks*, IEEE International Symposium on Biomedical Imaging (ISBI), Washington, DC, 2018. DOI: 10.1109/ISBI.2018.8363689.

[Thouvenin2021] O. Thouvenin, J Scholler, **D. Mandache**, M-C. Mathieu, A. Ben Lakhdar, M. Darche, T. Monfort, C. Boccara, J-C. Olivo-Marin, K. Grieve, V. Meas-Yedid, E. Benoit, *Automatic Diagnosis and Biopsy Classification with Dynamic Full-Field OCT and Machine Learning*, 2021. DOI: 10.21203/rs.3.rs-371207/v1. (preprint)

Contents

Chapter I

Introduction

Contents

I.1 Motivation

This manuscript opens with a quote from the modernist sculptor Constantin Brâncuși (1876, Hobiţa, Romania - 1957, Paris, France) saying *"What is real is not the appearance, but the idea, the essence of things"*. The choice is not inadvertent, aside from the author's fondness towards the artist, it is mainly due to the symbolism conjured up in his works which is beautifully summed up in this quote and serves as inspiration for the present study.

Bird in Space (1923, marble) by Constantin Brâncuși, source: Metropolitan Museum of Art, New York City.

As opposed to the realist current in art which aimed to accurately capture what the naked eye could see, modern art movements (e.g. symbolism, expressionism, etc.) capture concepts, thoughts, ideas by resourcing more intricate creative processes. Instead of snapshots of the surface, they offer unique depictions of the core, filtered by the artist's

creative "lens". To exemplify, Brâncuși's work *Bird in Space* (see image), through its lean minimalist lines, communicates the notion of flight itself rather than describing the appearance of a particular bird. We find that this duality of appearance *vs.* essence can be analogous to imaging modalities by comparison to standard visible light photography.

Imaging techniques do not share the intricacies of an artist's spirit, as they are the result of entrenched laws of physics, rather than metaphysics of the creative process. However, what they share is that they all provide different renderings of reality. In medical imaging, the same organ could be perceived via different modalities and different aspects would be revealed, be they anatomical or molecular.

In this work we are handling a novel family of imaging techniques destined for fast tissue analysis, *Full-Field Optical Coherence Tomography* (FFOCT) and *Dynamic Cell Imaging* (DCI), which even if they have a well-established mathematical formulation, the underlying revealed phenomena from biological tissue are still partly a mystery. We shall step in the shoes of an "art critic", given that is a person who is charged with *analyzing, interpreting, and evaluating* art, and we shall try to decode the FFOCT / DCI portrayals of cancerous tissue. In this regard, we shall be guided not only by mindful intuition but we are also turning to data driven approaches from the field of *Deep Learning* (DL) which aim to encode the essence of the data and deliver a generalized characterization of it.

The main motivation of this study is increasing the adoptability of this disruptive imaging technique which together with automated diagnosis and interpretation tools has vocation to improve patients outcome.

I.2 Context

I.2.1 Bio-Medical Imaging for Cancer Diagnosis

Medical Imaging History The term "exploratory surgery" might raise a few eyebrows now in modern times, however, it used to be a reality [1] as it was the only way to look inside the human body until near the end of the 19^{th} century, when medical imaging techniques emerged. Historically, there has been an everlasting race of going beyond what the naked eye can see, ever since antiquity there is evidence on the use of lenses [2], the magnifying glass is firstly documented in 1100 and the microscope in 1751. However, the major breakthrough that actually gave birth to the field of *medical imaging* happened in 1895 when Wilhelm Conrad Röntgen discovered radiography, which allowed to look inside the human body non-invasively for the first time. Technological advances in the field of computers and mathematical theory (i.e. Radeon transform) pushed the field of radiography further with the invention of the computed tomography (CT scan) in the 1970's. CT scanners use a rotating X-ray generator and a row of detectors placed all around the gantry to measure X-ray attenuations by different tissues

inside the body, these measurements taken from different angles are then processed using reconstruction algorithms to produce tomographic (cross-sectional) images of the body.

There is now a plethora of well-established medical imaging techniques, whose field of application is determined by the trade-off between spatial resolution and penetration depth: going from *magnetic resonance imaging* (MRI), *computed tomography* (CT) or *ultrasound* used in studying organ anatomy (big penetration depth, low resolution of the order of millimeters), passing by *optical coherence tomography* (OCT) and *confocal microscopy* ($1\,\mu\mathrm{m}$-$10\,\mu\mathrm{m}$ resolution and $100\,\mu\mathrm{m}$-$1\,\mathrm{mm}$ penetration), *classical optical microscope* (surface imaging down to diffraction limited resolution of $0.2\,\mu\mathrm{m}$), *super-resolution microscopy* [3] (which overcomes the diffraction limit but comes with photo-toxicity, photo-bleaching sample degradation disadvantages due to high excitation intensity or extended exposure times) to *electron microscopy* (EM) that can capture details of $0.1\,\mathrm{nm}$ at great inconveniences of high cost and sensitivity to vibration and external magnetic fields, which are used for single-molecule or organelle level analysis.

Cancer History Medical imaging is especially of service in the treatment of *cancer* [4], which is a silent killer especially in the early stages, when it is also most effective to detect it in order to improve the chances of survival. The world's oldest suggestion of cancer was found on Egyptian papyrus dating from 1500 BC that was documenting a breast tumor treated by destroying the tissue with a hot instrument, what is now known as cauterization. However, the origin of the term "cancer" is attributed to the father of medicine, the Greek physician Hippocrates (460-360 BC). Later, Bernard Peyrilhe (1735–1804), a French surgeon, set the foundations of experimental cancer research and tried to explain the, still not fully understood to this day, causing factors of the malady [5]. In modern times, cancer knew an important surge, leading to the U.S. declaring a "War on Cancer" by increasing the funding and support for cancer research, in the 1970s. Since then, cancer research and ultimately cancer therapy are also profiting from the scientific advances in the domains of imaging [4, 6], genomics [7], mathematics [8, 9] and more recently, artificial intelligence [10]. Focusing on current practices in cancer therapy, there are multiple levels of clinical examination adjacent to treatment (which is mostly surgery): screening, diagnosis and interventional methods; they are not limited to, but primarily represented by, imaging related procedures.

Cancer Screening Beforehand, there are the screening examinations, which are either mass screenings meant to detect very early conditions that are highly prevalent in a population or selective screening, dedicated to certain patients with a higher risk towards some pathologies (based on family history, for example). As they are routine, preventive procedures they should be minimally invasive and cost-effective. They can be as harmless as a naked eye visual examination by a dermatologist to look for skin cancer.

In the field of women's health, screening can start from palpating the breast to check for lumps. X-ray imaging, i.e. mammograms, are the gold standard of breast cancer screening, ultrasound can also be used for detecting dense breast masses, but also ovarian cancer or other more common conditions (ovarian cyst, endometriosis).

Another type of minimally invasive screening exams are *cytopathology* exams which consist in the analysis of cell types often found in fluid specimens (e.g. urine) or tissue smears (e.g. cervix). A routine such test is the Papanicolaou test that can find abnormal cells in the cervix which may turn into cancer. Pap smears were associated with a reduction of hundreds of thousands of cases of cervical cancer over the past three decades in the U.S. [11]. It is usually concomitantly done with a test for HPV (Human Papilloma Virus), which is a virus that can cause the cell changes leading to cervical cancer. The HPV test is a molecular biology test (e.g. PCR[1]) that looks for a known DNA sequence of the virus.

There are even genetic tests to screen for breast cancer, but they are selective tests meant for patients with a close family history of this malignancy. BRCA1 and BRCA2 are two different genes that have been found to impact a person's chances of developing breast cancer. The domain of cancer screening is under continuous development [12] and the most promising new method is conducted via a mere blood test to look for cell-free tumor DNA [13].

Cancer Diagnosis In the unfortunate event of a suspicious screening result, the only method for accurate diagnosis is *biopsy*[2] which implies sampling the abnormal mass localized through screening, this is usually done by a surgeon or an interventional radiologist. Biopsies can differ both in composition: liquid or tissue, extraction procedure: by needle, incision, forceps snap, etc. and adjacent navigation technique: X-ray, ultrasound, endoscopy (e.g. bronchoscopy), choice dependent on the organ sampled. Regardless of the biopsying method, all samples undergo the same procedure for analysis, covered by the field of *histopathology*[3]. It involves the microscopic analysis by a specialized doctor - the pathologist, of a so-called histology slide obtained through heavy processing of the biopsied tissue.

Histolopathology Immediately after excision, the biopsy is usually placed and stored in a formalin solution meant to keep intact the tissue architecture by stopping the degrading biological processes - i.e. fixation. Then, at the histopathology lab, the sample is dehydrated with alcohol and then embedded in a wax block which is then cut into very thin slices with a microtome. One or more sections are selected and placed on glass slides, stained to enhance contrast, covered with another protective glass slide and

[1] Polymerase Chain Reaction (PCR) is a technique in molecular biology for creating multiple copies of DNA from a sample to be able to study it in detail.

[2] The term "biopsy" comes from the Greek *bios*, "life," and *opsis*, "a sight".

[3] The term "histopathology" means "study of diseases of organic tissues" from the Greek words *histós* "web, tissue" + *pathos* "experience, suffering" + *logia* "sutdy of".

only then studied under a microscope. The most widely used stain and gold standard is Hematoxylin & Eosin (H&E), a combination of two biochemical dyes: the hematoxylin, which stains cell nuclei purple by binding to nucleic acids, and eosin, which stains the extracellular matrix and cytoplasm pink, by binding to proteins, other structures taking on different shades as a combinations of these colors. With this staining, there can be examined cell appearance criteria like their number, size, nucleus to cytoplasm ratio, cell organization, etc., based on which malignancy can be determined, together with the origin of the cancer, its subtype and grade. The grade is sometimes expressed as a number on a scale of 1 to 4 and is determined by how cancer cells look under the microscope, according to well established "scoring systems" (e.g. the Elston-Ellis system for breast cancer grading, the Gleason system for prostate cancer grading). Low-grade (grade 1) cancers are generally the least aggressive and high-grade (grade 4) cancers are generally the most aggressive, information which may help guide treatment options.

Other special tests on the cancer cells can also help to guide treatment choices, like *immunohistochemistry* (IHC) based staining, which uses antibodies, accompanied by a dye, to check for certain antigens (markers) in the tissue. Therefore, if the antigen-antibody binding takes place in the cells, the dye shall also "stick" to the sample and the colored binding site will signal the presence of the researched biomarker. IHC is routinely used, along with H&E, for breast cancer in order to reveal the tumor's responsiveness to certain targeted therapies. IHC stains are used to look for cancer biomarkers like estrogen receptors (ER), progesterone receptors (PR) and human epidermal growth factor 2 receptors (HER2), the worst prognostic being for the tumors non responsive to any of these (i.e. triple negative cancer). The analysis protocol, all the observations and final diagnosis are recorded by the pathologist via a written pathology report.

Extemporaneous Analysis The entire histology process is lengthy taking 2–3 days before a microscopic slide is ready for diagnosis. However, in some situations, the sample (either biopsy or surgical resection) may require immediate examination. Such scenarios might include verifying a biopsy's adequacy for diagnosis by ensuring the appropriate tissue was sampled sufficiently [14], tumor margin analysis to ensure complete removal of malignancy [15] or intra-operative assessment of sentinel lymph nodes to check for metastasis [16]. Extemporaneous analysis is normally done by *imprint cytology* or *frozen section histology*. The latter implies replacing the fixation, dehydration and embedding steps by freezing the tissue. Both techniques are part of the standard practice with a sensitivity of 83% and a specificity of 95% for frozen section histology and a sensitivity of 72% and a specificity of 97% for imprint cytology, figures quite low due to the important imaging artifacts. Moreover, both techniques are labor-intensive and still require specialized personnel and equipment, plus a waiting time of at least half an hour. Other well-established (although in other fields) imaging techniques include X-ray or ultrasound to examine the specimen, but the insufficient resolution does not ensure a sensitive diagnosis,

i.e. for radiography a sensitivity of 41% and a specificity 78%, while for ultrasound a sensitivity of 44% and a specificity of 94%.[4]

Novel Prototypes for Extemporaneous Analysis Since the extemporaneous tissue analysis is a field of major impact and the existing techniques are either too laborious or offer only a coarse resolution, novel unconventional techniques exploring this application are rapidly emerging. In [18] they review 16 groups of techniques for intraoperative margin assessment under a common framework: fluorescence, advanced microscopy, ultrasound, specimen radiography, optical coherence tomography, magnetic resonance imaging, elastic scattering spectroscopy, bio-impedance, X-ray computed tomography, mass spectrometry, Raman spectroscopy, nuclear medicine imaging, terahertz imaging, photoacoustic imaging, hyperspectral imaging and pH measurement. The majority of the 134 studies were in early developmental stages and none of the techniques distinguished itself from the others by demonstrating both high feasibility and high diagnostic accuracy.

Some are just sensing devices founded on the difference in some physical properties of the tissue (e.g. rigidity) that could discriminate between normal and cancerous areas but with little to no interpretable feedback to the clinician [17]. For example, MarginProbe [19], a handheld probe based on radiofrequency spectroscopy, senses the reflection of radio waves, detecting subtle electromagnetic differences between cells by comparing the returned signal with known tissue signatures. In spite of its 70% sensitivity and specificity, it is the only margin assessment device approved by the FDA. However, even if easy to use, the little feedback they give to the surgeon makes sensing devices difficult to adopt.

OCT *vs.* Confocal Microscopy If we were to compare with the specifications of gold-standard histology, where the slide thickness is $5\,\mu\mathrm{m}$ and resolution up to $250\,\mathrm{nm}$, we can narrow down the imaging techniques suitable for rapid diagnosis to non-destructive *optical slicing*[5] techniques that offer the highest resolution: confocal microscopy [20] and OCT [21, 22]. Confocal provides a similar spatial resolution while allowing for optical slicing, however, the contrast needs to be enhanced with fluorescence dyes, while OCT allows for further in-depth focus completely free of preparation at the expense of losing a few hundreds nanometers in spatial resolution. The penetration depth of 1–$2\,\mathrm{mm}$ and lateral resolution of about $10\,\mu\mathrm{m}$ makes OCT well suited for in-vivo study: it is routinely used in the clinical setting for imaging the layers composing the back of the eye to diagnose a wide range of pathologies (e.g. macular degeneration, glaucoma, retinal detachment and diabetic retinal disease), or endoscopic

[4] All reported metrics in this paragraph are based on the review [17] which cites various studies on intraoperative breast cancer assessment; thus, the numbers should be taken as indicative (rather than absolute) values, given the samples size discrepancy and different protocols between studies.

[5] In a widefield microscope, the entire focal volume is illuminated, but that creates blur from areas out of focus above and below the image plane, that is the reason why histology samples are a few microns thin; on the other hand, optical slicing can be achieved by suitably designed microscopes that can produce clear images of focal planes deep within a thick sample.

imaging [3] of arteries, esophagus, etc. On the other hand, confocal microscopy offers a slightly better lateral resolution (sub-micron), but an insufficient axial resolution which does not give enough penetration (less than $100\,\mu\mathrm{m}$) for executing a complete optical slicing of a biopsied specimen[6] that can be comparable to the mechanical slicing in histology.

Full-Field Optical Coherence Tomography (FFOCT) [23] comes to fill this gap by balancing frontal imaging with sufficient penetration depth and an improved lateral resolution with one order of magnitude higher than for standard commercially available OCT systems. In a recent clinical study [24] two surgeons obtain an average sensitivity and specificity of 87% in diagnosing breast malignancy in FFOCT images. Moreover, Dynamic Full-Field Optical Coherence Tomography (DFFOCT) [25] also known as *Dynamic Cell Imaging* (DCI) takes the technique one step further by revealing complementary information related to live cellular structures thanks to endogenous contrast derived from cellular activity, allowing for an improved sensitivity of 90% and specificity of 96%, according to the same study [24]. It is on these two techniques, detailed in Chapter II, that we shall focus in this work.

Digital Pathology Going back to the universally accepted and applied paradigms in pathology, there is currently an ongoing effort to shift towards digitizing the field of pathology [26]. The prepared glass slides can be digitized either simply with a camera mounted on a manual microscope or with high throughput slide scanners which are essentially robotized microscopes able to image hundreds of slides at a time. This approach is covered under the name of digital pathology and a virtual slide goes by the name of Whole Slide Image (WSI). It has been found that there is no loss in diagnostic accuracy when analyzing a digital slide as compared to a physical one [27], not to mention the added advantages of easier visualization (virtual microscopy), sharing (telepathology) and storage[7] - reasons why WSIs are being used for diagnostic, educational, and research purposes. Still, there is one bottleneck represented by the numerous vendor-specific formats for storing the WSIs and patient metadata. However, there is an effort towards adopting the universal DICOM standard at a larger scale [28]. DICOM stands for "Digital imaging and communications in medicine" and was created from the need to remove vendor embargo in the field of radiology - the pioneer of digitized medical imaging, in an effort to have a unified format readable across multiple PACS[8] of different hospitals, and has extended to multiple disciplines since.

Regardless of its format, a WSI consists of large multi-resolution multi-planar images, stored in a pyramidal fashion from the highest magnification at the base to the thumbnail at the apex. This is necessary for the ease of access and seamless visualization, given the important image size, the limited rapid access

[6] Core needle biopsies are generally performed with a larger-gauge needle, ranging from 14-gauge to 20-gauge corresponding to an outer diameter of 2.1–$0.91\,\mathrm{mm}$.

[7] The directives in most countries impose the storage of physical histology slides for a period of at least 10 years.

[8] PACS (picture archiving and communication system) is a medical imaging technology used primarily in healthcare organizations to securely store and digitally transmit electronic images and clinically-relevant reports.

memory (RAM) of common computers and it is also dictated by the navigation pattern of pathologists (i.e. repeated panning and zooming, frequent switching between multiple magnifications). WSIs are gigapixel images, a typical glass slide may be $20\,\mathrm{mm} \times 15\,\mathrm{mm}$ in size and may be digitized with a resolution of $0.25\,\mu\mathrm{m}$ per pixel (corresponding to a 40X magnification), resulting in a 4.8 gigapixel image, taking up to 15GB of storage (when captured with 24-bit color). Therefore, the pyramidal data-structure allows for accessing sub-region of the image without loading large amounts of data, leading to faster viewing of the image.

In the same effort of democratizing digital pathology, the research community has brought open-source platforms dedicated to WSIs for quantitative analysis - QuPath [29], or collaborative annotation - Cytomine [30], but also a lower level library - OpenSlide [31], dedicated to programmers, allowing to read WSIs in C, Python or Java code.

Computational Pathology A natural development, derived from the spread of slide digitization concomitantly with the rise of big-data approaches, is the conjunction of WSI and mathematics, informatics and statistics materialized in the field of computational pathology [32]. It covers a plethora of methods: from the classic image analysis approaches to advanced deep learning models, and applications: from cell counting, nuclei detection to more complex diagnosis, even predicting disease progression and treatment outcome.

The models can either mimic pathologist analysis or even discover new criteria. Histology images can also be combined with supplementary information extracted from pathology reports, clinical data (i.e. demography, patient history etc.) or even genomics data in order to build comprehensive methods. The leading applications are based on artificial intelligence / machine learning [33] algorithms.

I.2.2 Computer Aided Diagnosis

A Brief History of Computation, Machine Intelligence and Automated Diagnosis

Artificial intelligence (AI) is defined as the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making and translation between languages. The idea of a programmable computer dates back from the 19$^{\text{th}}$ century, when Ada Lovelace theorized a set of instructions (i.e. algorithm) to be followed by "The Analytical Engine" - the first computer - designed by Charles Babbage in order to execute complex calculations. The groundwork for AI was urged when diplomatic history met history of science with Alan Turing's automaton[9] ("The Turing Machine") cracking the Enigma, a machine used by the German

[9] An automaton (plural: automata) is a relatively self-operating machine, or control mechanism designed to automatically follow a sequence of operations, or respond to predetermined instructions.

armed forces to send encrypted messages securely in World War II. In spite of the major evolution since then, Turing's question *"Can machines think?"* [34] still guides research in machine intelligence.

Another pioneer in the field of computer science, whose work influenced nearly every major branch of mathematics, was John von Neumann who came up with the design paradigm of storing both the program and the data in a shared memory, which allowed for faster more complex computations; von Neumann architecture is still used in modern computers. He also touched to artificial intelligence ideas, namely on the relations between the human brain and computers [35]. Von Neumann defined life as an organism which can reproduce itself and simulate a Turing machine, idea which inspired John Conway's Game of Life [36] in 1970, a simulation mathematical model describing a two-dimensional cellular automaton that evolves under a simple set of rules, the only variable being the initial state. Regardless of the complexity of the grid, one can predict the state of each cell in the next timestamp based on the rules, still, modern neural networks are found to not straightforwardly converge to learn these hidden rules [37].

In the 1960's **dedicated programming languages** started to emerge, John McCarthy, scientist who also coined the term "artificial intelligence"[10], created LISP (list processing) one year after the appearance of FORTRAN - the oldest high-level programming language. PROLOG (*programmation en logique*), a language based on formal logic, was developed in 1973 at the University of Marseille. Both programming languages belong to the declarative programming paradigm, meaning they expresses the logic of a computation without describing its control flow, as opposed to the imperative paradigm (e.g. FORTRAN, C/C++, etc.). Even if these two legacy programming languages are still used nowadays, the current preferred choice for developing AI algorithms is Python.

LISP, PROLOG or similar derivatives were used to power the first AI applications which were mainly focusing on textual data. In 1966 MIT's AI Laboratory developed ELIZA, an early natural language processing (NLP) chat-bot simulating a human therapist. The predecessors of automated medical diagnosis systems were implemented via **expert systems**. An expert system can be used to solve problems within a specialized domain that usually requires human expertise. It relies on two components: a knowledge base and an inference engine. A knowledge base is an organized collection of facts about the system's domain gathered from the human experts. An inference engine interprets and evaluates the facts in the knowledge base in order to provide an answer through *if-else* clauses, known as production rules. Typical tasks for expert systems involve classification, diagnosis, monitoring, design, scheduling, and planning for specialized endeavors. In 1972 work began on MYCIN at Stanford University for treating blood infections based on reported symptoms and medical test results. The program could request

[10] John McCarthy coined the term "artificial intelligence" at a conference held on the campus of Dartmouth College in 1956 which is considered the birthplace of AI as those who participated became leaders in the field.

further information concerning the patient, as well as suggest additional laboratory tests, to arrive at a probable diagnosis, after which it would recommend a course of treatment. MYCIN could also explain the reasoning behind its response. Using about 500 production rules, MYCIN operated at roughly the same level of competence as human specialists in blood infections and better than general practitioners. In the 1980's the ONCOCIN [38] expert system was used by Stanford's Oncology Clinic to help choose the chemotherapy plan with the best chance of cure and the least chance of side effects, based on an extensive history of cancer cases.

The most popular early breakthrough of AI happened in 1997, when *DeepBlue*, developed by IBM, beat at chess the world champion, by testing the outcome of all possible moves, therefore leveraging more brute force than "intelligence", marking a revival of the field powered by hardware advances.

Another paradigm in AI, which touches more closely to the idea of "intelligence" by introducing the concept of learning, is **machine learning** (ML). Unlike classical algorithms that explicitly describe a solution to a well-defined problem that can be translated in computer language under the form of specific instructions, ML is destined for tasks that are ambiguous to define in such a way; ironically enough, those tasks are the ones that "come naturally" to humans, like visual or speech related tasks. ML algorithms are based on statistics, probabilities and optimization and can solve problems by experience, studying an extensive set of examples and finding patterns in them, then use that prior training on new data to make predictions. Tom Mitchell sums it up as "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." [39].

There are two families of ML methods: supervised and unsupervised. In a **supervised** problem, the input data is accompanied by the desired results and the algorithm has to learn a generalized rule that maps inputs to outputs; if the outputs are categories then it is a *classification* problem, and if the outputs are continuous then we call it a *regression* problem. In **unsupervised** learning, patterns in the data are found with no guidance (no given outputs), the main problem being *clustering* - splitting data into groups. Some popular ML methods are: random forests [40, 41], support vector machines [42] (SVM) (supervised), K-means [43] (unsupervised).

Concomitantly with ML, the **computer vision** field arose, based on an important discovery in neuropsychology on the hierarchical nature of mammalian vision. In the 1960s Hubel and Wiesel [44] observed that the neurons of the visual cortex each respond to a small region in the visual field i.e. receptive field, moreover, the neurons are linked in a cascading way, they detect basic features like edges then feed into more features like shapes to ultimately create more complex visual representations. Inspired by this concept, computer scientists developed pattern recognition and connectionism embodied by artificial neural networks.

Deep learning has its roots in 1957 with the invention of the **perceptron** [45], which served as a linear classifier, but it is basically the prototype of the modern artificial neuron. However, back then neural networks did not manage to perform better than existing machine learning algorithms, so research in the domain almost stagnated. In 1986 a major discovery was made, the **backpropagation** [46] algorithm which was used for adjusting the weights proportionally with their influence in the error. Thanks to it, in 1997, LeCun [47] managed to develop the first large-scale practical application with convolutional neural networks ***LeNet***: handwritten digits recognition on the MNIST database.

The field of deep learning developed slowly (and networks got deeper) together with the improvement of **GPUs** and the creation of big annotated datasets, like ImageNet [48], a set of 1.3 million high-resolution natural images belonging to 1000 classes, curated in 2009 by Fei-Fei Li's team. Until it led to the major "boom" of 2011 when Cireșan and colleagues [49] achieved superhuman performance in several image recognition challenges with deep neural networks, AlexNet [50] surpassed classical machine learning approaches by a significant percentile, winning the ImageNet competition. Furthermore, in the medical domain Cireșan's neural networks [51] also outperformed other approaches, in the ICPR'12 and MICCAI'13 challenges for breast cancer (mitosis) detection in large histology images.

The bet was always integrating research prototypes into clinical practice, which required outstanding robustness. Unfortunately, examples are scarce. Historically, the first computational prototype approved by the FDA dates back from 1995, AutoPap QC was destined for assisting the reading of Pap smear cytological samples. In 1998 FDA approved the first commercial CAD system for mammography screening ImageChecker. The latest news in the field, and a true milestone in computational pathology, is the FDA clearance in September 2021 of an aid to diagnosis for prostate cancer developed by Paige [52], it assists the pathologist by highlighting suspicious areas in a WSI, making it the first AI-based pathology product.

I.3 Outline

The present manuscript is structured as follows:

- **Chapter I** is the introductory chapter where we set the context of the work, namely by reviewing the evolution of imaging techniques involved in cancer therapy, from the invention of the optical microscope to the development of the computational pathology field, we also take a brief snapshot of the evolution of informatics and artificial intelligence leading towards aid-to-diagnosis methods.
- **Chapter II** dives deeper into the imaging techniques studied, Full-Field Optical Coherence Tomography (FFOCT) and Dynamic Cell Imaging (DCI). Here we present the theoretical implications involved in understanding the mechanisms behind them, from a brief introduction to optics

and how light interacts with biological matter to explaining the technical specifications of the techniques, like the components of the setup or the image formation.

- **Chapter III** focuses on the data. Here we describe the curation steps of the 3 datasets built and exploited in the present work. We introduce an original contribution of a generic texture-aware image sampling method, applicable on a wide set of image modalities and problems. We also enumerate the challenges of the present work related to the imaging and datasets which drive the methodological choices from the following chapters.

- **Chapter IV** introduces some indispensable notions of Deep Learning, with a special focus on Convolutional Neural Networks design, training and validation.

- **Chapter V** is a rich exploratory chapter covering multiple aspects of the FFOCT / DCI imaging, governed by the common objective of Fully Supervised Classification of cancerous *vs.* healthy samples. We employ multiple feature extraction techniques on FFOCT images, DCI raw signal and DCI processed images, together with a variety of classification strategies like: handcrafting a CNN architecture, training tree-based classifiers on source-separation features, fine-tuning a state-of-the-art architecture and decoding its learned feature etc.

- **Chapter VI** touches to a real-world problem, namely learning from data acquired in the clinical setting without special curation and lack of expert annotations, the ground truth being extracted from the readily available pathology reports. In this regard, we develop a training pipeline for Multiple Instance Learning classification of malignant *vs* benign biopsies; the model benefits of a transparent definition allowing to access the predicted diagnosis of image sub-parts.

- **Chapter VII** exploits the duality of FFOCT / DCI imaging, here we apply a Contrastive Learning approach to overcome the artifacts in DCI with the robustness of FFOCT. We develop a method for robust characterization of fibers from DCI imaging by using corresponding FFOCT images as a guide i.e. minimizing the cosine distance between matching image pairs in a common latent space learned via a Siamese Neural Network. Moreover, we also give special attention on validation, by defining the identity and symmetry error metrics.

- **Chapter VIII** is the concluding chapter where we summarize the contributions of the present work and propose some ideas for future developments.

Chapter II

Static & Dynamic Full-Field Optical Coherence Tomography

The gold standard tissue assessment technique i.e. histology suffers from multiple difficulties caused by the complex sample preparation, associated with sample deformation and degradation and a long processing time between sample excision and diagnosis delivery. This delay hinders fast clinical decision making - necessary at the time of surgery, for example. Optical imaging techniques have been investigated as a more efficient alternative. These techniques are based on the properties of light which have been studied extensively for centuries and they are supported by a strong theoretical background on newly emerging light-based techniques.

In this chapter we will first compile some basic notions about optics, notably the interaction between light and biological matter, but we will mainly focus on static and dynamic full-field optical coherence tomography (FFOCT and DFFOCT / DCI) by also introducing the technique they are based on, i.e. classical OCT.

Contents

II.1 Optics 101: Theoretical Foundations

II.1.1 Light

Theory of light has evolved greatly from the ancient times when it was believed that our eyes were producing light that allowed us to see, to theories enabled by the scientific revolution and superior mathematics (by Huygens, Young, Fresnel, to name a few) which were later unified by Maxwell in 1865 by defining light as an electromagnetic wave, it was experimentally confirmed by Hertz later in 1888. In the 20$^{\text{th}}$ century, quantum mechanics and the theory of relativity were major deviations from the classical theory of Newtonian physics which led to dramatically revising our understanding of the nature of light. Einstein in 1905 proved the **speed of light** in vacuum to be a constant $c = 3 \times 10^8\,\mathrm{m/s}$ and he also contributed together with Plank to formalize the dual nature of light as both wave and particle (i.e. photon). [53]

Some phenomena are explainable by the particle nature of light, like the photoelectric effect, while others by the wave behavior which explains how light bends (or diffracts) around an object, a relevant example is light interferometry principle.

Electromagnetic waves are created as a result of vibrations between an electric field and a magnetic field, i.e they are composed of oscillating magnetic and electric fields, and they can have natural (e.g. the sun) or artificial (e.g. the light bulb) origin. Electromagnetic waves can be described by their most basic properties: amplitude A, wavelength λ and frequency f. **Amplitude** is the height of a wave as measured from the highest point on the wave (peak) to the rest point and can be interpreted as the intensity of oscillation. **Wavelength** refers to the length of a wave from one peak to the next and is inversely proportional with frequency $c = \lambda \times f$ where c is the speed of light constant, f is the frequency in Hz or s^{-1} and λ is the wavelength measured in m. **Frequency** corresponds to cycles per second and can be interpreted as the rate of oscillation. Longer wavelengths will have lower frequencies, and shorter wavelengths will have higher frequencies. The electromagnetic spectrum encompasses all of the electromagnetic radiation that occurs in our environment and includes (from high to low frequencies) gamma rays, x-rays, ultraviolet light, visible light, infrared light, microwaves, and radio waves. The visible spectrum in humans is associated with wavelengths that range from $380\,\mathrm{nm}$ - blue light to $740\,\mathrm{nm}$ - red light (See Figure II.1).

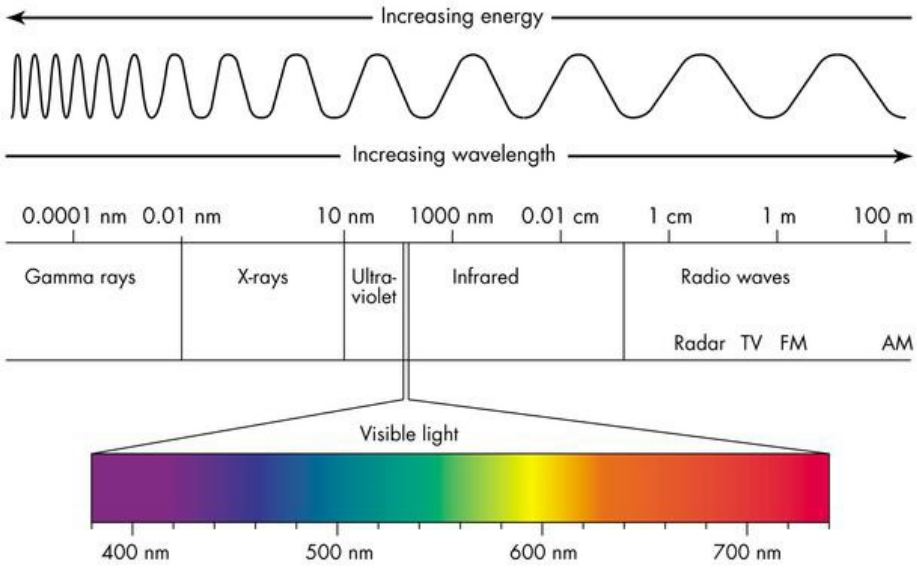


Figure II.1: The electromagnetic spectrum.

The speed of light in a transmitting medium is less than that of the speed of light in a vacuum. The **index of refraction** n is the ratio of the speed of light in a vacuum to the speed of light in a medium and is an important physical property of different materials that needs to be considered in optical setups. To give a few examples: $n_{\text{air}} \approx n_{\text{vacuum}} = 1$, $n_{\text{tissue}} \approx n_{\text{water}} \approx 1.3$ and $n_{\text{glass}} \approx n_{\text{oil}} \approx 1.4$. The changes in the refractive indexes along the light path cause it to bend. In the next section we will develop more on the phenomena taking place between light and matter.

II.1.2 Light - Matter Interactions

Light is indispensable to life on earth, the primordial light source - the sun - produces energy sustaining both animal and plant life, directly or indirectly. It directly sustains plant life by a series of biochemical reactions in *photosynthesis* that converts light energy captured by light-absorbing pigments (e.g. chlorophyll) into chemical energy in the form of ATP [1].

Interactions between light and matter determine the appearance of everything we see through four basic phenomena - emission, absorption, transmission and reflection (or scattering) - usually taking place concomitantly at different extents.

Emission

Light can be produced by matter which is in an excited state and excitation can come from a variety of sources, ranging from electric current (e.g. incandescent light bulb, LEDs etc.), chemical reactions or

[1] Adenosine triphosphate (ATP) is an energy-carrying organic compound that drives many processes in living cells.

even biological sources, etc. Some of those phenomena are leveraged for marking molecules of interest in biological imaging.

Bioluminescence is light resulting from biochemical reaction by a living organism: luciferin is a substrate that reacts with oxygen in the presence of a luciferase (an enzyme) to release energy in the form of light (e.g. fireflies). Transferring luciferase-expressing genes into animal models (e.g. mice) allows their in-vivo study through bioluminescence imaging.

Fluorescence is the emission of light by a substance that has first *absorbed* light or other electromagnetic radiation. Fluorophores are fluorescent chemical compounds that can re-emit light upon light excitation, they are notably used to stain tissues, cells, or materials in fluorescent imaging (e.g. DAPI, diamidino-phenylindole, stain binds to DNA so it is used in fluorescent microscopy to reveal cell nuclei, when DAPI absorbs UV light ($350\,\mathrm{nm}$), it emits blue ($460\,\mathrm{nm}$)).

Absorption, Transmission, Scattering

All living or inorganic things have a color, which means that they selectively reflect, absorb and/or transmit light. Pigments are selective color absorption substances that absorb certain wavelengths while reflecting others, therefore the reflected wavelengths will be perceived by the observer as the object's color. An example of such substance is *chlorophyll*, which absorbs the blue and red colors of the spectrum and reflects green therefore leaves appear green. In particular, white and black colors reflect and, respectively, absorb all wavelengths.

The quality of transmission of light of a material - *transmittance* - is simply translated by its transparency and it is usually quantified as the percentage of the incident light that can move all the way though the material.

Particles scatter light, this is a fundamental fact and something we all encounter on a daily basis, the sky is blue. This is caused by stronger light scattering of blue light by atmospheric particles than red light. The angle, wavelength(s) and intensity of the scattered light depend upon the particle size. We note two main theories: *Rayleigh scattering* for particles smaller than a tenth of the wavelength (here $r \leq \frac{\lambda_0}{10} \approx 50\,\mathrm{nm}$) which is distinguished by a stronger backscatteringthan *Mie scattering* (for larger particles $\frac{\lambda_0}{10} \leq r \leq \lambda_0$) in which case most of the light is transmitted forward, in the direction of the incident light (see Figure II.2). For even larger particles, the forward scattering is even more focused and amplified.

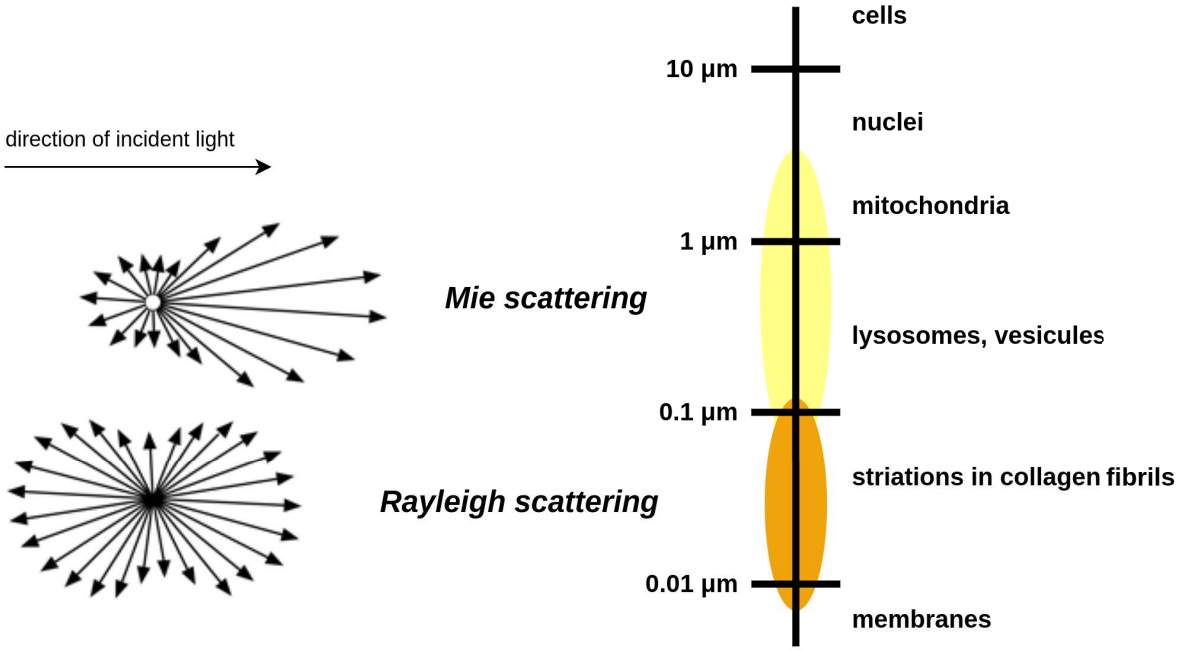


Figure II.2: Relationship between scattering pattern (Mie or Rayleigh) and particle size as a hierarchy of biological ultrastuctures found in animal tissue (scale adapted from [54]).

II.1.3 Optical Properties of Biological Tissue

Some important *chromophores*[2] responsible for visible light absorption in animal tissue are: hemoglobin (blood), melanin (skin, hair, iris of the eye), nucleic acids (cell nuclei), water, etc.

As back-scattering depends upon smaller than wavelength particles, it is attributed to the ultrastructure of tissue (see Figure II.2), e.g. the density of lipid membranes in the cells, the size of nuclei, the presence of collagen fibers, the level of hydration in the tissue, etc. Some notable *biological scatterers* are: collagen fibers, lysosomes, mitochondria.

Collagen Fibers

Collagen fibers are the main component of connective tissue in the body, acting as a glue[3] to hold tissues together, they provide structure and support throughout the body, being present in different proportions depending on the organ.

Collagen fibers (about 2-3 µm in diameter) are composed of bundles of smaller collagen fibrils about 0.3 µm in diameter. Mie scattering from collagen fibers dominates scattering in the infrared wavelength range. On the ultrastructural level, fibrils are composed of entwined tropocollagen molecules. The fibrils present a banded pattern of striations with 70 nm periodicity due to the staggered alignment

[2]Chromophores are molecules in a given material that absorb particular wavelengths of light, and in doing so confer color on the material.

[3]The name "collagen" comes from the Greek *kólla* meaning "glue" and suffix *-gen*, denoting "producing".

of those tropocollagen molecules. The periodic fluctuations in refractive index on this ultrastructural level appear to contribute a Rayleigh scattering component that dominates the visible and ultraviolet wavelength ranges [55].

Mitochondria

Mitochondria[4] are intracellular organelles about 1–10 µm in length, that make up to 25% of the cell volume, counting 1000-2500 mitochondria per cell.

Their function is crucial for cell homeostasis as they generate most of the cell's energy i.e. ATP, which gave its nickname "powerhouse of the cell". Even though their textbook depiction is of as bean-like structures, they form a highly *dynamic* network in the majority of cells where they constantly undergo *fission* and *fusion*.

Mitochondria are composed of many folded internal lipid membranes. The basic lipid bilayer membrane is about 9 nm in width. The refractive index mismatch between lipid and the surrounding aqueous medium causes strong scattering of light. Folding of lipid membranes presents larger size lipid structures which affect longer wavelengths of light. The density of lipid/water interfaces within the mitochondria make them especially strong scatterers of light.

Measurements of isolated organelles indicate that mitochondria and other similarly sized organelles are responsible for scattering at large angles, whereas nuclei are responsible for small-angle scattering [56]. In a later study [57], the same team states that the nucleus is responsible for less than half the scattered light in the cell, and scattering intensity is increased in cells with a higher number of mitochondria. In [58], it has been shown experimentally that mitochondria are responsible for 80% of the Mie scattering in cell suspensions.

II.1.4 Light Coherence Principle

Coherent light is a focused beam of light which consists of only one single frequency (or very few frequencies i.e. monochromatic) that are also in-phase, it is produced by lasers. On the other hand, incoherent light contains multiple wavelengths that are also out of phase.

Spatial coherence describes the correlation between waves at different points in space and it is illustrated in Young's double slit experiment. While temporal coherence describes the correlation between waves observed at different moments in time, it is demonstrated by the Michelson interferometer.

The presence of coherent light in an optical system enables the generation of interference fringes. Interference is an optical phenomenon that appears when two waves superpose in a coherent fashion, creating

[4]The name "mitochondrion" (plural mitochondria) comes from Greek *mitos* "thread" and *khondrion* (diminutive of *khondros* "granule").

a wave with a different pattern, which contains periodical regions of greater amplitude (constructive interference) and regions of low amplitude (destructive interference).

The property of wave interference involves the phenomenon of self-interference (a wave interferes with itself) due to multiple forward and backward scattering of light, conducting to the degradation of the signal, leading to a specific type of noise, known as speckle.

We can understand wave interference in the simplest case, by considering the superposition of two sine waves which have the same wavelength, same amplitude, but with a a phase difference of φ: $W_1(x,t) = Acos(kx - \omega t)$ and $W_2(x,t) = A\cos(kx - \omega t + \varphi)$. Their resulting wave is reduced to $W_1 + W_2 = 2A\cos(\frac{\varphi}{2})\cos(kx - \omega t + \frac{\varphi}{2})$. For $\varphi = 0$ (and multiples of 2π): $W_1 + W_2 = 2A$ (constructive interference) while for $\varphi = \pi$ (and multiples of π): $W_1 + W_2 = 0$ (destructive interference).

Note that low coherence sources are desired in exploiting interferometry, this quality of a wave is quantified by the coherence length. Several microscopy techniques use the information contained in the interference pattern, among which is optical coherence tomography (OCT).

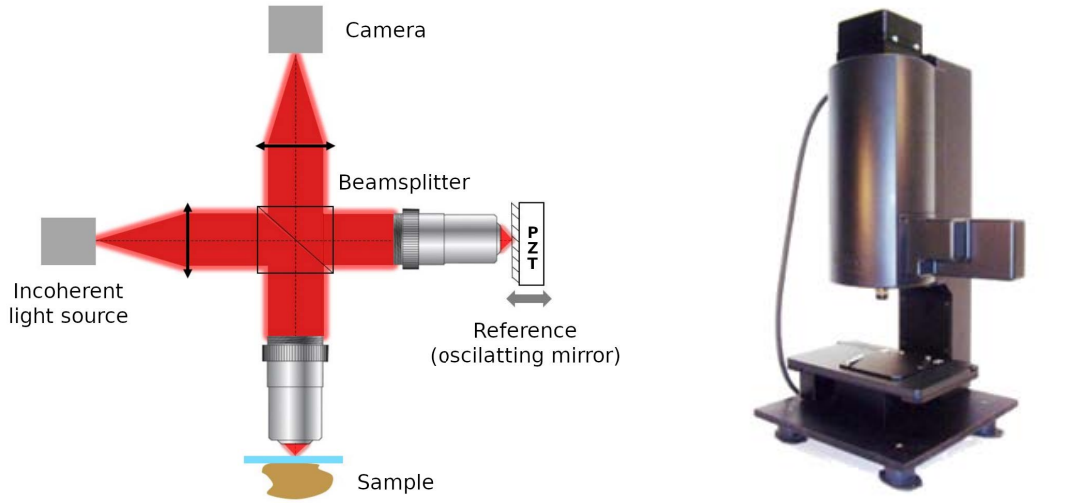
II.2 Optical Coherence Tomography

Proposed in 1991 by Huang et al [59], optical coherence tomography (OCT) is, as the name suggests, an imaging (*-graphy*) technique which uses visible light (*optical*) and exploits its *coherence* quality in order to perform virtual sectioning (*tomo-*).

The idea behind its functioning principle can be compared to ultrasound, which uses sound waves to image inside the body by capturing the time it takes for the sound wave to return (*echo*) to the sensor, revealing the internal anatomy. However the speed of light is "infinitely" higher than the speed of sound by an order of one million and no sensor can capture the respective arrival time of photons. Therefore, low-coherence interferometry measures the echo time delay and intensity of backscattered light by comparing it to light that has traveled a known reference pathlength and time delay.

OCT is a non-invasive, non-toxic, in-vivo imaging technique that allows the observation of precise structure and tissue information at the micrometer scale on biological samples that are a few millimeters deep. In practice, OCT lays at the frontier between medical in-vivo and ex-vivo imaging. To date, OCT has had the largest clinical impact in ophthalmology, being used routinely in clinical exams, for studying the integrity of the layers forming the retina.

For 30 years, this technique has been studied, developed, extended [60], to become one of today's most promising alternative to surgery-based histology. Modern OCT images, with higher resolution, give access to more details on the morphological and tissue information, broadening the image analysis



(a) Linnik interferometer (adapted from [61])

(b) LightCT™ scanner (LLTech, France)

Figure II.3: Optical Setup for FFOCT / DCI Imaging

possibilities in fields like optical biopsy and virtual histology. What is more, the classical scanning approach that produced cross-sectional images is now extended to *en face* imaging, more suitable for histology-like assessment.

II.3 Full Field Optical Coherence Tomography

Full Field Optical Coherence Tomography (FFOCT), developed [1] and perfected [23] by the ESPCI team of Pr. Claude Boccara and made commercially available by LLTech in 2011, comes to fill a gap between classical OCT and confocal microscopy. It is used for medical and research purposes in analyzing biological tissue morphology and function, without any preparation (e.g. dying), offering a resolution of $\approx 1\,\mu\mathrm{m}$ in all 3 dimensions. To give an intuition on the order of resolution, organelles (e.g. mitochondria) vary in size from $1 - 10\mu m$, animal cells measure on the order of tens of microns up to $100\mu m$.

To perform optical sectioning, i.e. the selection of a signal coming from a confined area in the Z plane, FFOCT relies on the same principle as OCT, low coherence light interferometry.

As opposed to classical OCT, which performs a single-point raster scan producing a cross-section XZ image, FFOCT produces "en face" XY images thanks to 2D illumination and array detector i.e. megapixel camera (CCD or CMOS). This justifies the name "full field", because it captures the entire frontal plane at once, instead of point by point.

The commercial FFOCT / DCI system, developed by LLTech offers a resolution of 1 micron in 3D with a supported sample size of maximum 2.7cm in diameter and 5 mm in height. An individual field of

view measures $1.3mm \times 1.3mm$ and its acquisition takes $67ms$, its processing $30ms$ which ensures a speed of scanning of 1 minute per cm^2. The penetration depths it can reach depends on the type of tissue examined, it can vary between $100\,\mu\mathrm{m}$ and even $500\,\mu\mathrm{m}$ for more transparent tissues. It has a simple setup, the scanner has the size of a regular microscope and is connected to a computer with its dedicated control and acquisition software installed on it.

II.3.1 Interferometer Configuration

The LightCT™ system, which is the LLTech FFOCT system, is based on a Linnik interferometer (a variation of the more common Michelson interferometer, with the only distinction of having objectives in both arms) whose optical setup is presented in Figure II.3. The light from a low coherent broadband source is divided by a beam splitter into two arms: (i) a sample arm pointing to the examined sample and (ii) a reference arm containing a reference mirror.

After the two beams are reflected respectively by the sample and reference mirror, their recombination (i.e. interference) is acquired by the camera. The final image is obtained by subtracting multiple images acquired with a phase shift so as to remove the incoherent part of the signal and reveal the backscattered light in the focus plane. By moving the optical block towards the sample holder and matching the same displacement in the reference arm, one can move the focus plane to image the sample at different depths, this explains the z-sectioning ability of the system, with z resolution increasing as coherence length decreases.

The main components of the setup are:

- **illumination arm**: low-coherence broadband light source with the central wavelength $\lambda_0 = 565\,\mathrm{nm}$ and the bandwidth $\Delta\lambda = 104\,\mathrm{nm}$, ensuring a sectioning ability of $1\,\mu\mathrm{m}$;
- **beam splitter**: ensures both the splitting of the two sister beams and then their regrouping into the detector;
- **reference arm** with a **reference mirror** attached to a **piezoelectric translation** (PZT) in order to apply phase-shifts (i.e. fine displacements on the order of nanometers) to demodulate the interference signals;
- **sample arm**: equipped with the sample holder;
- water-immersion **microscope objectives** with a numerical aperture of $NA \approx 0.3$ and magnification $\mathcal{M} = 10\times$ are present in both reference and sample arms;

- x**,** y**,** z **translation stages**: both reference and sample arms are mounted on mechanical translation stages for optical path length matching and scanning the sample in 3D, which are piloted by high precision **motors**;

- **detector**: CMOS camera with a 2MP resolution corresponding to 1440×1440 pixels of size $12\,\mu\mathrm{m}$ with high full well capacity (FWC)[5], and up to $700\,\mathrm{Hz}$ framerate.

II.3.2 Image Formation

The image captured by the camera is the sum of: continuous background composed by the continuous contributions of the light reflected on the mirror and on the sample and the incoherent light coming from the sample's multiple scattering and out of focus planes, and the weaker modulated interference signal: $I = I_0 + A\cos\Phi$. In order to extract the relevant interference signal i.e. the second term of the equation, another acquisition with a phase modulation of π is introduced through a Piezo inducing the mirror to displace with $\lambda/2$ where λ is the coherence length. This method called 2 phase-shift method is detailed in equation (II.1). In order to extract both the amplitude A and the phase Φ terms separately 4 acquisitions are needed i.e. 4 phase shifts of $\pi/2$ steps corresponding to displacements of $\lambda/4$), the method is detailed in eq. (II.2).

$$\begin{cases} I_1(x,y) = I_0(x,y) + A(x,y)\cos[\Phi(x,y)] \\ I_2(x,y) = I_0(x,y) + A(x,y)\cos[\Phi(x,y)+\pi] \end{cases} \Rightarrow A\cos\Phi = \frac{I_1 - I_2}{2} \tag{II.1}$$

$$\begin{cases} I_1(x,y) = I_0(x,y) + A(x,y)cos[\Phi(x,y)] \\ I_2(x,y) = I_0(x,y) + A(x,y)cos[\Phi(x,y)+\frac{\pi}{2}] \\ I_3(x,y) = I_0(x,y) + A(x,y)cos[\Phi(x,y)+\pi] \\ I_4(x,y) = I_0(x,y) + A(x,y)cos[\Phi(x,y)+\frac{3\pi}{2}] \end{cases} \Rightarrow \begin{array}{l} A = \frac{1}{2}\sqrt{(I_1-I_3)^2-(I_4-I_2)^2} \\ \Phi = \arctan\frac{(I_4-I_2)}{(I_1-I_3)} \end{array} \tag{II.2}$$

The obtained FFOCT image represent the amplitude A of the interference, and it captures the optical path difference which is a result of the optical properties of the tissue under investigation, such as differences in refractive indexes, scattering variations or differences in absorption. See Figure II.4a) for reference. The result is a gray scale image where highly backscattering elements, mostly fibrous structures (e.g. collagen), appear white while weakly backscattering content like cells appears dark gray or black.

[5]Larger pixels collect more light and are crucial for the FFOCT technique where, by its nature, the camera works at a near saturation regime.

II.3.3 Technical Specifications

We define resolution as the minimum physical distance between two points at which these points appear to be separated in an image. Now we define the axial (cross-sectional or Z-axis) and lateral (spatial or XY-axes) resolution of the FFOCT system.

Axial resolution

Axial resolution is given by the **temporal coherence length** L_c, which is inversely proportional with the source spectral width. Therefore, it is the low coherence light coming from broadband source that ensures fine optical slicing. The light source used has a central wavelength of $\lambda_0 = 565\,\mathrm{nm}$ and a bandwidth of $\Delta\lambda = 104\,\mathrm{nm}$.

$$\Delta L_c = \frac{2\ln 2}{\pi} \times \frac{\lambda_0^2}{\Delta\lambda} \approx 2.7\,\mu\mathrm{m} \tag{II.3}$$

The imaging medium is defined by $n_{\text{coverslip}} \approx n_{\text{oil}} \approx 1.4$. Assuming the source has a Gaussian broadband spectrum, the axial resolution is:

$$\Delta z = \frac{L_c}{2n} \approx 0.96\,\mu\mathrm{m} \tag{II.4}$$

Transverse resolution

One of the main limitations of optical microscopy is the **diffraction limit**. Due to the wave-like nature of light, the objects present on the light-path of the microscope induce a modification of the waveform of the incident beam, due to self-interference of the wave. As a consequence, when looking through a microscope, a spot is seen rather as a small disc pattern, known as the Airy disc. Hence, if two spots are too close to each other, the microscope will not be able to determine from where exactly is coming the light, and these two spots will be seen as a single Airy disc.

The transverse (spatial) resolution is given by the limit distance at which two Airy discs can be resolved as two separate spots and it is defined as the Rayleigh limit. It is dependent upon the properties of the light source (i.e. wavelength λ but also the properties of the objective lens that is characterized by its numerical aperture $NA = n\sin\alpha$ which is a measurement of the refractive index of the medium between the specimen and the coverslip n and angle α of the light cone. In this case the immersion medium is oil (whose refractive index is $n_{\text{oil}} \approx 1.4$) and it is uses a water-immersion objective with NA $= 0.3$ in water and NA ≈ 0.32 in oil that allows for a $10\times$ magnification, therefore:

$$\Delta x = \Delta y = \frac{1.22\lambda_0}{2\text{NA}} \approx 1.08\,\mu\mathrm{m} \tag{II.5}$$

Sensitivity

Sensitivity is defined as the minimum reflectivity R_{min} measurable by the system. This can be translated into the limit where the signal is at noise level, i.e. SNR $= 1$. It is dependent upon the reflectivity of the mirror in the reference arm R_{ref} and the saturation level of the camera i.e. the full well capacity, FWC $= 2 \cdot 10^6$ electrons. R_{ref} is measured as the percentage of reflected light per the total of the incident light and for our setup including a silicone coverslip and silicone oil coating (to reproduce the imaging conditions of the sample arm) the reflectivity is 23%.

$$R_{\text{min}} = \frac{R_{\text{ref}}}{16 \cdot \text{FWC}} \approx 7 \cdot 10^{-9} \tag{II.6}$$

Penetration depth

The penetration depth is strongly dependent on the sensibility of the system and the type of tissue, as light is gradually attenuated due to absorption and multiple scatterings. Generally, we can image at a maximal depth of $200\,\mu\text{m}$ down to $500\,\mu\text{m}$ in more transparent media.

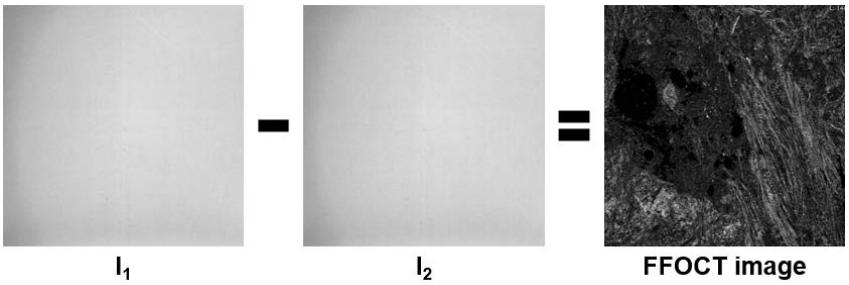
Acquisition speed

The imaging speed of one FOV is dependent upon camera framerate, the phase shift algorithm (here 4-phase Eq. II.2), processing time, image accumulation rate for increased sensitivity or signal to noise ratio. For large-field acquisition, when scanning is required, it is also added the time necessary for the mechanical stage displacement and a short pause for setup stabilization in order to image the next FOV.

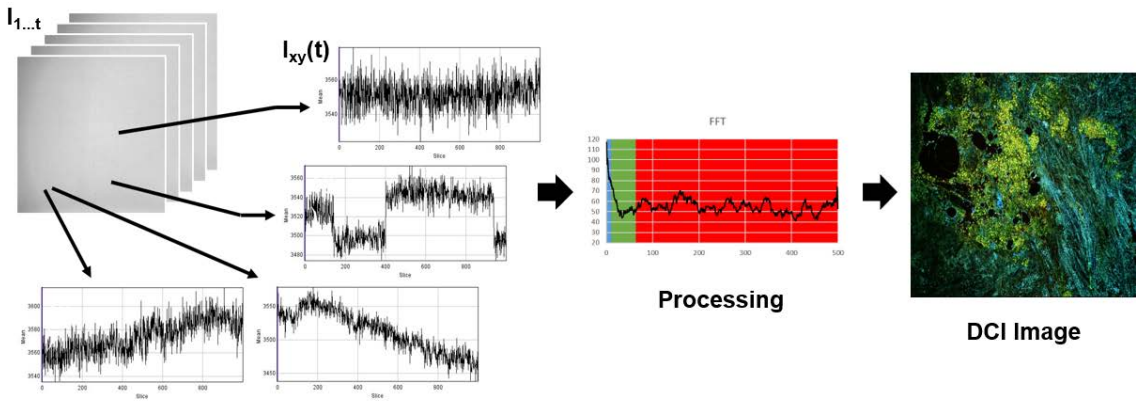
In the case of our systems, the acquisition speed is around $1\,\text{cm}^2/\text{min}$ with a FOV overlap of 10% on each axis, a waiting time of $50\,\text{ms}$ between each acquisition, 5 accumulations of images and at a frequency of acquisition of $300\,\text{Hz}$.

II.4 Dynamic Full Field Optical Coherence Tomography

Dynamic Full Field Optical Coherence Tomography (DFFOCT) or, by its commercial denomination, Dynamic Cell Imaging (DCI) has been developed recently by Pr. Boccara's team [62]. It works on the same setup as static FFOCT, the main difference consists in acquiring multiple interferometric figures over time, at fixed reference arm (piezo off), which allows to quantify microscopic movements of the backscaterrers. It comes to solve a major drawback of FFOCT namely that highly scattering structures like fibers and membranes can mask less reflective ones like cells (see Figure II.2). Since cells have higher intrinsic activity than collagen fibers they become apparent in DCI. Therefore, it is only suitable



(a) Simplified FFOCT image formation using the 2 phase-shift method, obtained as the difference of two interferometric frames acquired with a $\lambda/2$ phase difference, which results in conserving the interference signal only.



(b) DCI image formation from the acquisition of a stack of interferometric frames to processing of the signal in frequency domain and composing the RGB image.

Figure II.4: Image formation schematic illustrations.

for imaging fresh tissue, before cells would experience apoptosis. It brings information on the metabolic activity at an intra-cellular scale that is complementary to the static information given by FFOCT.

II.4.1 Technical Specifications

By analyzing both contributions to the dynamic signal we can show that axial (z) sensitivity to displacement is larger than the transverse (x, y) sensitivity. Indeed, the minimal z motion corresponding to a change from the minimum to the maximum of the $cos(\varphi)$ is:

$$\Delta z_{\text{dynamic}} = \frac{\lambda}{8n} = 50\,\text{nm} \tag{II.7}$$

In the x and y directions however, a scatterer has to move from one voxel to its neighbor to affect the signal, typically corresponding to half the point spread function (PSF) width:

$$\Delta xy_{\text{dynamic}} = \frac{\Delta x}{2} = 575\,\text{nm} \tag{II.8}$$

No phase-modulation is needed because the motion information is extracted from time series analysis. The meaningful information is extracted from the raw signal either by calculating the standard deviation of the signal in time or by Fourier analysis. At 1000 frames acquired and a camera framerate of 300Hz, we can observe oscillations of down to $3\,\mathrm{ms}$ and up to a few seconds, meaning that no long displacements (e.g. migration) or very fast displacements (e.g. Brownian motion), are observed.

The major drawback of this method is its high sensitivity to vibrations, either exterior noise or the implicit vibrations induced by the motors that drives the translations in x, y or z axes, for exploring the entire sample. When an undesired movement happens it induces a displacement that acts as a phase modulation which artificially increases the dynamic signal. This creates signal artifacts that are more visible on the fiber which normally gives poor dynamic signal. In order to attenuate the influence of vibrations, a pause time of $5\,\mathrm{s}$ is introduced between acquisitions to help the setup stabilize.

II.4.2 Image Formation

The data cube resulting from a DCI acquisition of 1000 frames acquired with a 300Hz frame-rate in a $1.3\,\mathrm{mm}^2$ field of view (FOV) is quite significant: $1440 \times 1440 \times 1000$ pixels ($\sim$4 GB). For visualization purpose the 3D data is transformed to an RGB image according to the image formation algorithm [62], which consists in performing a Fast Fourier Transform (FFT) and averaging the amplitudes over 3 sets of frequency bands, resulting in 3 channels coded in RGB colors:

In DCI images both spatial and temporal information are superimposed in one image, in color this time, each color channel (RGB) depicting a range of oscillation speeds of the scatterers. Starting with the steps of image formation are:

1) normalize frames to constant mean value, close to captor saturation, to remove frame-to-frame inconsistencies;
2) average the frames by group of 4 obtaining 250 frames pseudo-acquired at 75Hz;
3) pass in frequency domain with Fast Fourier Transform (FFT) obtaining 125 variation maps with a step of 0.3Hz;
4) normalize FFT by its norm l_0;
5) define color channels: blue is the 0.3Hz frequency map, green is the average of the maps from 0.6Hz to 5.1Hz, red is the average of the maps from 5.4Hz to 24Hz;
6) apply fixed multiplication gains to each channel to balance their contributions;
7) contrast stretching with fixed thresholds for each color channel.

This processing method, illustrated in Figure II.4b), allows a contrasted visualization which is consistent[6] between fields of view. However, it illustrates only vaguely the physical properties of the tissue, the intuition being that highly saturated pixels reveal oscillations of high amplitudes, while the colors suggest the frequency of oscillation (from lower in blue to faster in red). Accordingly, fibers usually appear in blue as they are static, while live cells in yellow as they have highly active sub-components confined in their volume.

[6] A better contrast might be obtained by employing adaptive processing methods, at the expense of losing tractability and reproducibility between acquisitions.

Chapter III

From Clinical Data to Computational Input

Having presented the FFOCT / DCI imaging modalities lying at the core of this work, we shall now plunge deeper into the collected datasets. In this chapter, we first set the grounds for data curation, by introducing the steps needed to accommodate the data "from patient to processor". Then, we present our working datasets by showing the particularities of the studied organs and their adjacent pathologies, as well as the strategies for preparing the data, annotations, etc. An important contribution of the present chapter involves building the datasets by appropriately mining images and annotations, for example, by developing the texture-aware sampling method *Sample Optimally with SLIC* (*SoSleek*).

Contents

III.1 Data Curation

Data is the world's most valuable resource today; *"Data is the new oil"*, as the mathematician Clive Humby said in 2006. In 2020 tech giants like GAFAM[1] are capitalizing on data-driven services worth 4.5 trillion dollars. Since Google and Facebook offer completely free services and their only revenues come from advertisements, their worth comes from the detail profiling of their users to whom they feed personally targeted ads. Similarly, 35% of Amazon's revenue is generated by informed marketing through its recommendation engine and likewise 80% of the content streamed on Netflix is thanks to their recommendations, which brings them over $1 billion a year in value from customer retention.

Data Centralization

It is indisputable that big data analytics are already well-established in our daily life consumption, but for the sector of healthcare it is still in the developing process due to the higher risks and matters at stake, like privacy, but also due to the lack of *data digitization and centralization*. However, the COVID-19 pandemic proved that rapid advances can be achieved in the medical field (and not only) with sufficient deployment of human and material resources and inter-nations collaboration. This was not only demonstrated by the fast development of vaccines, but also for creating a common gateway (in the form of the digital vaccine passport) as a universal way of monitoring immunizations while ensuring data protection. Another advance driven by the pandemic is the creation of one of the largest (in terms of number of patients) unified database: The National COVID Cohort Collaborative (N3C) comprising clinical data from 8 million persons out of which almost 3 million are pathological cases gathered from 65 sites across the U.S.

Data Digitization

In order to create databases, more important than centralization is data digitization, but the good news is that the global acceleration of digitization adoption driven also by the COVID-19 pandemic can be quantified as a fast-forward of 7+ years in the growth of the share of online products and services according to a McKinsey Survey. The people's growing need of a continuous information stream and the health-oriented trends are pushing the "smart" health trackers forward like fitness watches or even FDA or CE approved medical devices as connected blood glucose monitor sensors, not to mention the ambitious Neuralink project developing implantable brain–machine interfaces. Nonetheless, in the lab of the histopathologist, the shift from glass slides to WSI is slow.

Regardless, some important digitization and centralization efforts were deployed to create medical databases primarily focusing on genomics, but containing also the WSIs of the specimens: The Human Protein Atlas [63], The Cancer Genome Atlas (TCGA) [64], some smaller more standardized databases

[1] Google, Apple, Facebook, Amazon, Microsoft

released through multiple inscription-based automatic diagnosis grand challenges: breast cancer metastasis in lymph nodes from histology - CAMELYON [65] and BCNB [66], breast cancer from histology - TUPAC [67] and BACH [68], glioblastoma from MRI - BRATS [69], colon cancer from histology - GLAS [70], to name a few.

Data Storage

The volume of data multiplied exponentially over the last years reaching 64.2 zettabytes (zetta- = 10^{21}) in 2020. By the year 2025 global data creation is projected to grow to more than 180 zettabytes. However, when it comes to data storage, just 2% of the data produced and consumed in 2020 was saved and retained into 2021, as the storage capacity reached 6.7 zettabytes, according to a Statista Survey. This tells us that besides data acquisition, there must be carefully crafted methods to help bypass the storage bottleneck by compressing unstructured data into relevant, comprehensive information that can be then re-used to answer new questions. Timeliness is an important quality of data. Taking the example of longitudinal analysis which is threatened by recency bias manifested by giving more importance to most recent patterns just because there is more data available for recent times. In order to accomplish a successful data curation campaign and prioritize quality over quantity one needs to: conceptualize, acquire, select and clean. In this section we will discuss the steps data goes through to become a suitable input for data analysis methods.

III.1.1 Data Acquisition

As our imaging does not belong to routine clinical practices, the image acquisition campaigns are conducted through clinical studies. Clinical studies are organized under strict conditions as an agreement between the sponsor and hospital ethics committee, they must approve the clinical procedures implied by the study (which are bound to alter the standard procedures), the research hypothesis as well as the cohort size and quality required to get statistically significant results. Then, with each procedure eligible for the study, the patient is informed about the aim of the study and use of their data and they need to give their consent of participation in order to be included in the study.

For the clinical studies conducted by LLTech to test the feasibility of the LighCT™ scanner, the direct risk to the patients is non-existent, as the device does not come into contact with the patient, but with an excised piece of tissue. Moreover, most of the time, there is no need for a supplementary tissue sampling especially destined for the LighCT™ scanner, as the FFOCT / DCI imaging procedure does not alter the sample in any way so it can be then analyzed following the standard protocol by the histopathologist, without a risk of having an impact on the diagnosis.

The scanner has to be located in the proximity of the place of excision, e.g. operating room (for surgical excisions) or imaging room (for biopsies), firstly because that is its intended use in clinical practice i.e.

at point-of-care analysis, but also because the nature of DCI imaging itself, as it captures phenomena occurring in living cells only, the image acquisition has to be done shortly after tissue excision to ensure sample integrity and excellent image contrast. Fortunately, anyone with a minimum prior training can perform the image acquisition protocol, be it non-medical personnel, nurses or even the clinician performing the excision act, as it requires less than 5 minutes to image a biopsy, for examples.

An important step that differs from the standard protocol is the storing medium of the sample, in the routine procedure, the sample is placed directly into a formalin solution to stop the biological processes leading to the degradation of the tissue, but that would also stop the processes that allow for DCI imaging. Therefore, the sample needs to be placed in a saline solution instead, to prevent tissue dehydration without killing the cells. If the sample will undergo histopathology analysis then it needs to be moved to a formalin solution after being imaged with DCI. The medical staff involved in the study needs to be aware of this extra step as to not jeopardize the usability of a specimen.

III.1.2 Data Annotation

Regardless of its abundance, unlabeled data is most often useless. State-of-the-art diagnosis algorithms are supervised, meaning they learn from labeled data and even for the case of unsupervised methods there needs to be a ground truth for validation. However for natural images, finding the category it belongs to or performing object selection can be accurately done by any non-expert human agent or even by web crawling. To give an example, the ImageNet [48] database was created starting from a hierarchical words database, then ~ 1000 images per concept were gathered by querying search engines, followed by a crowdsourced manual validation phase on the automatically collected images, resulting in 14 million annotated images. On the other hand, medical data is not only expensive to collect, but also to annotate, as only domain experts can perform this time-consuming cumbersome task. To assist the annotation process there are dedicated software tools for labeling with bounding boxes, polygons or pixel annotations, like ilastik [71] or Cytomine [30] and Icy [72] tailored for bio-medical imaging.

In our case, annotation is ever more challenging as there is no expert in both the FFOCT / DCI techniques and all the pathologies it is applied on. Still, annotation is usually done by histopathologists as they have the most knowledge domain in tissue micro-architecture, but insights can be collected from the main intended users of the technique, i.e. surgeons or radiologists. Anyhow, every medical expert annotator undergoes a prior training in FFOCT / DCI supported by an appropriate image atlas for the tissue type.

Nonetheless, the annotator turns to the corresponding H&E slide for reference, but image correlation is not straightforward. Even when the same specimen is being imaged with FFOCT or DCI and then processed with H&E, image registration presents itself with other challenges induced by morphology variations due to: differences in the optical and physical slicing plane and orientation; deformations

caused by the particular states of the tissue once fresh and compressed inside the sample holder box, respectively fixed, dyed and cut (moreover, even different fixation methods can induce a distinct inter-cellular space [73]). Another aspect is the difference in the amount of information present, as a histology slice has a thickness of around $5\,\mu\mathrm{m}$ while a FFOCT / DCI slice has a "thickness" of $1\,\mu\mathrm{m}$, therefore, we would probably see less cells, especially for the cell types with a small diameter, e.g. immune cells. For a manual approximate matching, the natural approach is to start from a low resolution and find correlations by progressively zooming. Since the overall shape of the sample varies a lot from one modality to another, it makes it difficult to even find the correct rotation transformation; matching the structures is far more viable but quite an exhaustive work due to the highly resolved images.

Given the cumbersome annotation process, not all instances get annotated due to insufficient confidence caused most often by a failed correlation with histology. That is why data post-processing is usually necessary to harmonize the dataset.

III.1.3 Data Sampling

Given the important resolution of medical imaging in general and DCI and FFOCT images in particular and the limited computational resources available for performing image analysis, image subsampling is a crucial step. This is particularly true in the case of deep learning and convolutional neural network architectures since the input size influences exponentially the dimensions of the computed tensors.

III.1.3.1 Regular Grid Sampling

The naive approach for patch sampling, which is also the most used, is using a regular grid with or without overlap. The undebatable advantages of this method are the implementation simplicity, lack of parameters and fast runtime (order of micro seconds) make it sufficient for dense splitting into small patches of convex shapes. However, as regular grid sampling is agnostic to the image content, in the case of more delicate datasets where instance-level annotations are not available and multiple structures are present in the original image, the regular grid sampling would inevitably partition object instances which could further impact the analysis. In the case of small datasets data quality is of uttermost importance to reduce the problem complexity. In this respect, it is possible to opt for texture-aware sampling to ensure the presence of similar structures in a patch, at the expense of computational complexity.

III.1.3.2 Texture Aware Sampling with *SoSleek* Method

In the field of image analysis we often draw inspiration from the human visual system, now we turn to the incremental grouping theory for answers, it states that the visual cortex tends to group similar entities and perceives them as a whole object [74], this assumption naturally extends to the way a pathologist

analyses the tissue so we try to mimic this texture-aware part-based decomposition in our sampling strategy.

Early image segmentation methods introduced the notion of superpixels which are groups (or clusters) of pixels that share some common features like intensity and proximity. One simple - as the name suggest - method Simple Linear Iterative Clustering (SLIC) introduced in [75] gathered more than 7 thousands paper citations and was used in at least 90 patents from different imaging fields and application types. Its state of the art quality is proved not only by its user base, but also by comparison [76] with similar methods [77–80].

SLIC is based on the K-means [81] algorithm to cluster the image pixels which are each described by a feature vector consisting of its intensity value(s) and its spatial coordinates. At initialization, the K cluster centroids are set on a regular grid, therefore the parameter K is necessary for this method and it represents the approximate number expected clusters i.e. superpixels. Pixels are assigned to a cluster based on a similarity measure which weights both the distance in the color and spatial domain. Furthermore, the method allow finer parametrization for controlling the size, and shape and connectivity of the superpixels.

In computer aided tissue assessment, superpixels are used as a preliminary step for ROIs detection [82–85] or specific structures (i.e. cell nuclei [86]) from which features are extracted for further analysis, mostly classification. However, our analysis becomes agnostic to the properties of the generated superpixels, like shape or connectivity, instead we use the approach to ensure elegant tissue sampling, in order to employ classic patch-based algorithms.

An updated SLIC variant [87] restricts the segmentation to a previously given region (i.e. mask) and is advertised as having medical imaging as its main application field.

In order to adapt SLIC segmentation for the task at hand i.e. optimally sample big images into texture-aware patches, the parameters have to be set accordingly. For SLIC, the main parameter is the (approximate) number of superpixels to generate, corresponding to the number of clusters from the K-means algorithm; on the other hand, for patchification, the main parameters are the patch size and patch overlap. However, the path overlap is an approximate measure of sampling density given in pixels, it only controls the initial grid spacing. The stride of the regular grid used to initialize will therefore be $S = \text{patch size} - \text{patch overlap}$, and given the total number of pixels in the image N, for classical SLIC: $K = N/S^2$ and for maskSLIC: $K = \alpha \times N_{mask}/S^2$, considering N_{mask} to be the number of pixels in the mask and $\alpha \geq 1$ an allowance compensating for shape irregularities. Given the big dimensionality of the images, it is desirable to run the algorithm on the downscaled image without any quality loss of the expected output, also the image can be converted to grayscale, or other colorspace, e.g. CIELAB used in the original paper.

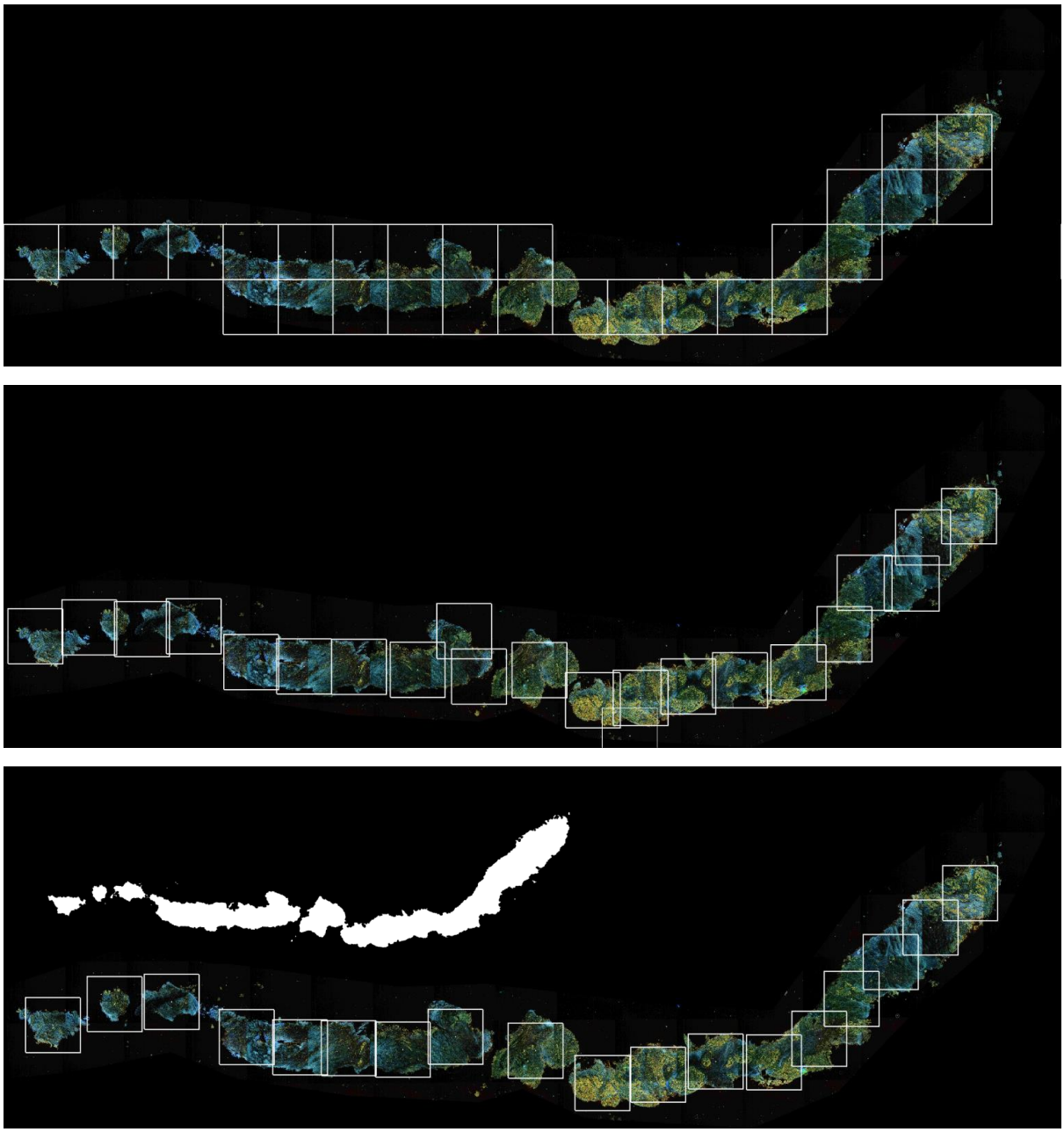


Figure III.1: Sampling example of a breast core needle biopsy (top to bottom): regular grid, *SoSleek* using classical SLIC, *SoSleek* using maskSLIC (mask was obtained with bi-modal thresholding followed by morphological operations); image size is $19\,920 \times 6\,805$, patch size is 1024×1024 with zero overlap.

In order to ensure an adapted superpixel segmentation for our patchification task we tune the SLIC parameters, the ones bearing most influence on the outcome are:

- *min / max factor size*: control superpixel size; parameters given as proportions of the size of the square superpixel initialization;
- *compactness*: balances color proximity and space proximity in the distance metric, higher values give more weight to space proximity, making superpixel shapes more square, which is desirable for our patchification task.

When a foreground mask is not available, the developed method allows removing the background superpixels by performing a threshold on the average intensity of the superpixels, the choice of the thresholding method is dependent on the data distribution and desired results, so more methods are available, like Otsu [88], Isodata [89], Triangle [90], Li [91], Quantile, etc. The resulted sampling is obtained by extracting the patches centered at the center of mass of each foreground superpixel.

Algorithm 1 SLIC texture aware sampling (without mask)

Require: image I, patch size p, overlap o, scale s, ...

$I \leftarrow \text{grayscale}(I)$ ▷ Transform image to grayscale
$K = \frac{I.width \times I.height}{(p-o)^2}$ ▷ Estimate number of superpixels
$I \leftarrow \text{downscale}(I, s)$ ▷ Downscale image for faster execution
$S \leftarrow \text{SLIC}(I, K, ...)$ ▷ Apply SLIC algorithm
$S \leftarrow \text{threshold}(S, method)$ ▷ Keep foreground superpixels only
$S \leftarrow \text{upscale}(S, s)$ ▷ Return to original image coordinates
$C \leftarrow \text{center\_of\_mass}(S)$ ▷ Get superpixel centroids
$P \leftarrow \text{extract}(I, C, p)$ ▷ Get image patches
return patches P, centers C ▷ Return list of patches and their centroid coordinates

See Figure III.1 for a comparison of regular grid and SLIC sampling. See Algorithm 1 for a concise depiction of the pipeline. Moreover, the code developed in the present work can be found at `https://github.com/dmandache/sleek-patch` in the form of a Pyhton package together with a running demo. The code uses the scikit-image SLIC implementation.

III.1.4 Data Balancing

Real world data is highly imbalanced. In perspective, the worldwide healthy human population is far bigger than the one suffering from the most prevalent form of cancer, i.e. breast cancer, which is of 0.21% [92]. However, the ratio changes dramatically in a different population slice - for example in the case of people undergoing breast biopsy (so already having a suspicious mammography) the chance of finding cancer sufferers increases dramatically as 10% of neoplasms are malignant.

However, in mathematical modeling and especially in AI, all such bias is undesirable as it introduces a favoritism for predicting the majority class and thus the accurate discrimination between classes becomes difficult. This problem is highly studied as it touches all domains related to big data [93] and it is either handled at the data level or at the algorithm level. In this section, we will briefly introduce some of the most used data-level approaches like: random resampling, informed oversampling or data augmentation with transformations. As for the methods playing upon algorithmic particularities, they rely mostly on sample weighting. However, for huge data discrepancies weighting accordingly could introduce numerical instability so a joint approach with resampling is advised. Another aspect of predictive algorithms that is impacted by class imbalance is the evaluation of performance, in this regard the choice of suitable metrics is discussed in Section IV.3.

Random resampling [94] provides a naive yet effective [95] technique for rebalancing the class distribution for an imbalanced dataset. Random oversampling involves randomly selecting examples from the minority class, with duplication, and adding them to the training dataset. Random undersampling involves randomly selecting examples from the majority class and removing them from the training dataset. However, one drawback of the former approach is overfitting, while for the latter, it risks to not capture true class distribution by discarding vast quantities of data.

A more data-driven approach is Synthetic Minority Oversampling Technique [96] (SMOTE) which can be seen as an informed oversampling or data augmentation for the minority class. The process consists in creating a new sample by interpolating two samples from the minority class; more specifically, a random example from the minority class is chosen together with one randomly selected neighbor (from its k nearest neighbors [97, 98]) and a synthetic example is created at a randomly selected point between the two examples in feature space. Although the authors suggest SMOTE is suitable for low dimensional data, in [99] is proposed a so called ImageSMOTE algorithm which does nothing else but add Gaussian noise to the randomly oversampled images, nevertheless the method improves performance by more than 30% in detecting glioblastoma from MRI scans. Regardless, there is little evidence in favor of successfully using SMOTE for images, however, in the next section we will talk about more suitable methods to enrich an imaging training set via data augmentation. The imblearn Python package provides functionalities for the aforementioned approaches.

To conclude this section, class imbalance is a crucial yet inevitable problem in ML applications and it needs to be addressed at all stages of a method development: data pre-processing, algorithm design and performance evaluation, therefore we will return upon this problem along the present manuscript.

III.1.5 Data Augmentation

In stricto sensu data is abundant and omnipresent, however, as we have already mentioned in the previous sections, structured data suitable for training ML algorithms is utterly scarce due to the laborious and expensive process of collecting and annotating it. In other words, *data is cheap, but information is expensive*. These aspects are ever more severe in the clinical setting where many ethical questions need to be addressed and also the highest expertise is needed for annotation. Data augmentation [100] is a way of maximally capitalizing on already available data and is now an indispensable practice in data driven applications. It consists of artificially enriching training sets by reproducing relevant and naturally occurring transforms on the given data, thus trying to capture the true data variation.

The image processing field offers a plethora of techniques for data augmentation, from simple to more complex, through various transformations:

- rigid transforms: flipping, rotation, translation, scaling, cropping, etc.;
- elastic transforms: warping etc.;
- color altering: changing brightness, contrast levels, etc.
- adding noise or blur.

Apart from these classical approaches, the more recent DL advances can be exploited to create new synthetic data, like variational autoencoders (VAE) or generative adversarial networks (GANs). The choice of the adapted strategies has to be made with respect to the properties of the original images and applications at hand. For example, CNNs are translation invariant by definition thanks to the convolution operation and they can become rotation invariant thanks to the input they were trained on (in the case rotations were used as an augmentation transformation). However, the choice of data augmentation strategies needs to be consciously made; taking the example of rotation, natural images are normally upright, while biological images are inherently unoriented.

III.2 Working Datasets

Data lies at the center of the new ML paradigm applications including the present work. In this section there will be presented all three datasets built for the present work together with their particularities that drove the choice of pre-processing methods introduced in the previous section.

The two organs studied are skin and breast, they present different grades of pathological complexity and histological appearance diversity. What is worth mentioning here is that one of the main histological features for diagnosis skin cancer is the organization of collagen fibers in the tumor adjacent stroma, which is best captured by FFOCT. On the other hand, breast cancer exhibits a wider scope of morphological

features which revolve mainly around cell shape, size and organization, making DCI a more suitable imaging modality for diagnosing it.

The following sections will detail the context of the imaging data obtained including: the original aim of the cross-sectional clinical study it originates from, the figures of the cohorts included, annotation types and the strategies for building the corresponding training sets.

III.2.1 Skin Cancer: Basal Cell Carcinoma Clinical Study

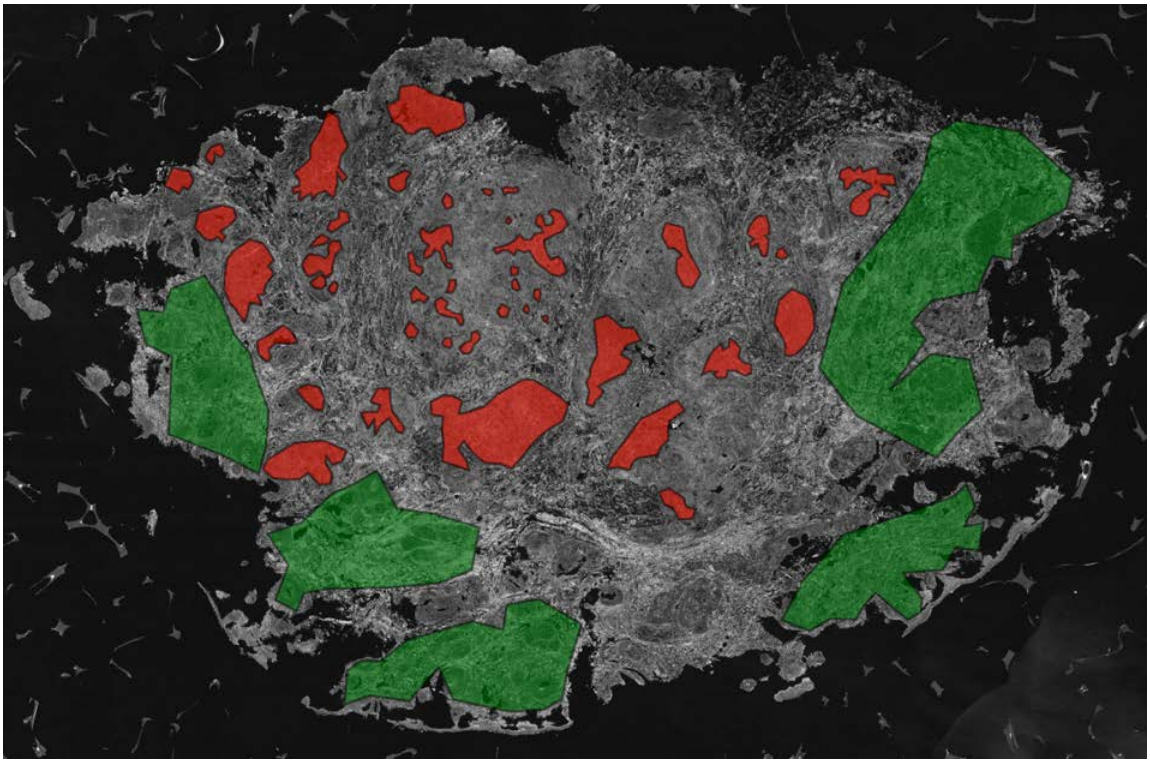
Skin cancer is the most common human malignancy, predominantly represented by non-melanoma types with 5.4 million cases per year, 80% of which are Basal Cell Carcinomas (BCC) with the majority the remaining being Squamous Cell Carcinomas (SCC) [101]. The gold standard procedure for treating non-melanoma skin cancer in high risk areas is Mohs Surgery [102]. The technique involves the consecutive removal of thin layers of skin, followed by histological preparation and microscopical examination for tumor clearance. This process can take up to an hour and guides further tissue extraction. For this clinical study conducted in partnership with *Drexel Medicine* it was investigated the feasibility of using FFOCT imaging, together with an automated diagnosis of the cancerous areas, which would lead to speeding up the procedure, consequently, improving patient comfort and physician throughput.

Cohort

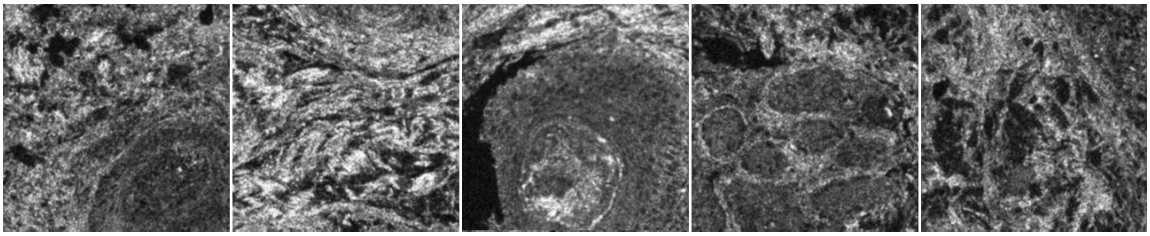
Our data set consists of 40 FFOCT images (out of which 10 were cancerous) of tissue excisions obtained from Mohs surgery, biopsies and conventional excisions, which were then imaged using the LightCT^{TM} scanner. Each image is a 2D transverse slice of a unique tissue sample imaged at $20\,\mu\mathrm{m}$ below the surface. Samples measure between 2-2.5cm^2 which gives high-resolution images of around 200 Megapixels.

Annotation

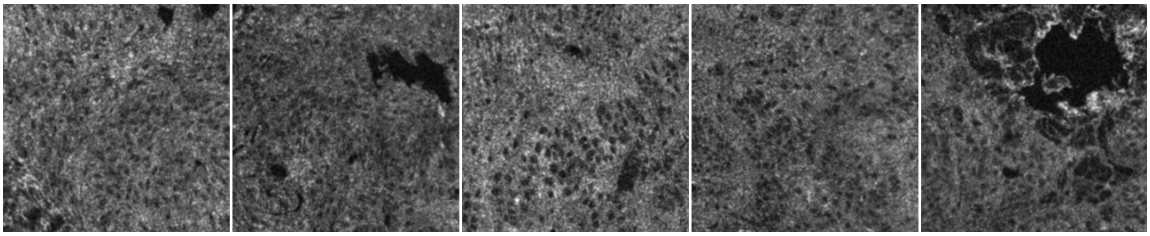
As shown in Figure III.2 the images were manually segmented and diagnosed by a dermatopathologist with experience with this modality and who could access the gold standard H&E frozen sections of the specimen for validation. The data was gathered and annotated using the Cytomine [30] platform. 26% of the total imaged area was segmented, the rest (unlabeled areas) being background or abnormal tissue that sometimes surrounds the tumors but its appearance is not relevant for either class, it should be treated separately. The images are preponderantly annotated with the *normal* label and only 10 images present some cancerous areas, annotated as *BCC*, more precisely, only 9.5% of the annotated data is pathological. Therefore, as it is the case of most of the medical applications, we face with the class imbalance problem, which we try to solve by oversampling the minority class.



a) Annotated FFOCT image of cancerous skin sample (*normal* in green, *BCC*in red)



b) Normal patches: collagen, hair follicles, glands



c) BCC patches: aggregates of cancer cells, retraction artifacts

Figure III.2: An example of skin sample imaged with FFOCT ($11\,808 \times 8\,352$) annotated in Cytomine and some patches (256×256) extracted from its annotated areas.

Sampling

The scanner produces 16-bit DICOM images, but only 10 to 12 bits are actually used; they were converted to 8-bit JPEG for convenience, so they can be tested with out of the box pre-trained architectures that only accept this depth.

Speckle noise is strongly present in FFOCT images, but proper denoising algorithms are too computationally costly (of the order of hours) therefore, since one of the requirements of our application is speed, we applied a 3×3 Gaussian filter to provide some smoothing while preserving the structures (e.g. cancerous cell nuclei appear as dark blobs with 10 pixels in diameter).

A constraint imposed by the computational power needed to train artificial neural networks is its number of parameters. This is a function of the depth (number of filters, layers) and width (input size of the layers) of the network. To satisfy this constraint while capturing enough context to discern normal skin structures from the cancerous cell organization, we split the images in patches of 256×256 pixels. With the aim of augmenting and also balancing the data set, we oversampled the patches with different step values for the two classes: 170px for the normal class, while BCC patches overlap more, with a stride of only 40px. This produces 108 082 patches: 59 112 normal and 48 970 BCC; 80% of which serve as a training set and the rest is used for measuring the performance.

Among the popular practices in deep learning is data standardization (zero centering + normalization) which translates into imposing the data to follow a normal distribution. This influences the robustness of the algorithm to variations in the images caused by the acquisition conditions, for example, and it also ensures a better convergence of the learning process. Data standardization is done by subtracting the mean intensity value over the training set and dividing by their standard deviation. Note that the same preprocessing has to be applied on the test data for consistency, but with the statistics of the trainset. Furthermore, some basic data augmentation was performed, which led to doubling the training set, by adding synthetically generated images obtained through horizontal and vertical flipping, slight rotations and shifts.

III.2.2 Breast Cancer

Breast cancer is the most frequent cancer in women worldwide, representing almost 25% of all cancers in women, it is also the second most deadly (15.4% of deaths) after lung cancer. Standard treatment involves the surgical removal of the tumor, with partial or total breast ablation. Even after heavy surgery, the risk of recurrence after 5 years is above 10%, suggesting that an imperfect removal of the tumor was performed during surgery. Hence, there is a crucial need to improve real-time intraoperative characterization of the tumor margins, in order to reduce the ablation of healthy tissue, the surgery time, and the risk of additional surgery and cancer resurgence.

Standard diagnosis procedure consists in mammography screening, where tissue of abnormal density can be suspected through X-rays and is then biopsied and analyzed at microscopic scale to be actually diagnosed. Localizing and reaching the lesion is not a trivial task. The common protocol implies multiple samplings (minimum 5 [103]) of the suspicious mass to ensure correct probing for reliable diagnosis. Nevertheless, the false negative rate caused by poor sampling is still currently up to 2% [104].

The breast anatomy consists of adipose and fibrous tissue together with its specific structures: lobules and ducts which play the role in lactation, namely to produce and transport the milk, respectively. The lobules have a "grape-like" appearance and the ducts are tubes of different sizes that link the lobules to the nipple. The size and number of lobules greatly differs from one individual to another and they also change with time, notably with menarche, pregnancy and menopause. The appearance of ducts depends on the slicing orientation, therefore for transverse cuts they would appear rounded and can be easily confounded with a lobule.

This branching ductal network is composed of two *epithelial cell* types: an inner layer of polarized luminal epithelial cells (where cancer arises most often) and an outer layer of myoepithelial cells, separated from the collagenous stroma by a basement membrane. Other cell types that help sustain the function of the mammary glands are adipose, fibroblasts, immune, lymphatic and vascular cells.

The most frequent histological type of breast cancer is invasive ductal carcinoma (IDC). IDC accounts for 80% of all invasive breast cancers in women which is also reflected in our datasets. IDC evolves from ductal carcinoma in-situ (DCIS), their correspondents are invasive lobular carcinoma (ILC) and lobular carcinoma in-situ (LCIS). It is worth mentioning that in-situ appearance can also be localized in the samples diagnosed as invasive or infiltrating. Apart from this 4 main categories, there are more sub-classifications, in total there are 21 histological types of cancer and 5 main molecular types [105]. Figure III.3 illustrates the complexity of breast cancer morphology.

The gold standard diagnosis method is the ultrasound guided core needle biopsy procedure followed by histological analysis of the samples. In the eventuality of a positive diagnosis, the standard treatment involves the surgical removal of the tumor, with partial or total breast ablation. Even after heavy surgery, the risk of recurrence after 5 years is above 10%, suggesting that an imperfect removal of the tumor was performed during surgery. Hence, there is a crucial need to encourage real-time intraoperative assessment of the tumor margins, in order to reduce the ablation of healthy tissue, the surgery time, and the risk of resurgence.

Both clinical acts that could benefit from rapid at point-of-care diagnosis, biopsy and surgery, are tackled in this work.

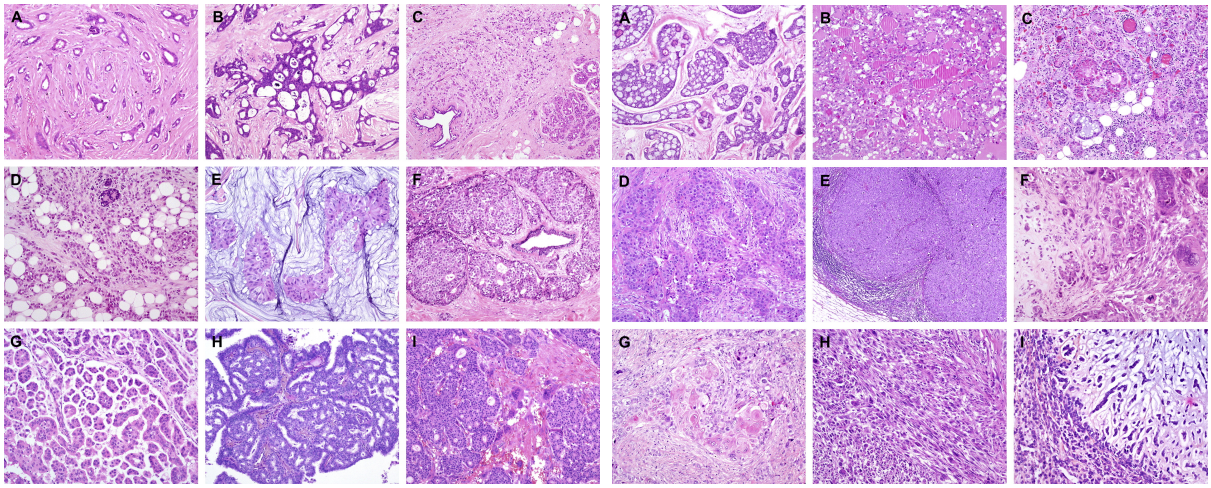


Figure III.3: Breast cancer morphological variation in H&E staining adapted from [106]).

III.2.2.1 Surgical Excisions Pilot Study

Cohort

An early pilot study was conducted by LLTech together with *Gustave Roussy Institute* in order to test the feasibility of using FFOCT and DCI in discriminating breast tumors from healthy breast structures. The tissue imaged comes from surgical waste resulted from mastectomy surgeries. The dataset contains a total of 47 samples coming from 33 patients, 11 of these samples are normal and 36 cancerous, according to the diagnosis set by a pathologist based on the corresponding H&E stained histology slides. For each sample there are available three image modalities:

- gold-standard histology slide digitized with Hamamatsu scanner with a 20X objective, pixel resolution $0.452\mu m$, taking up around 5GB of space, in *.ndpi* format;
- large-field FFOCT acquisition on the whole sample extent, average size roughly $20k \times 18k$ px dependant on sample size, with a pixel resolution of $0.9\mu m$, occupying a few hundreds of MB, in *.dicom* format;
- multiple randomly distributed unit DCI fields of view, chosen by the LLTech engineer who performed the acquisitions, with a resolution of 1440×1440 px corresponding to an area of $1.3mm \times 1.3mm$; ranging from 3 to 16 fields of view (FOV) per sample, with a median of 10 FOVs/sample;
  - the processed DCI images use less than $10\,\mathrm{MB}$ of storage space and are in *.tiff* format;
  - the raw interferometric stack (1000 images at $300\,\mathrm{Hz}$) is also saved in case post-processing is required, they take $4.15\,\mathrm{GB}$ and are in *.hdr / .img* format.

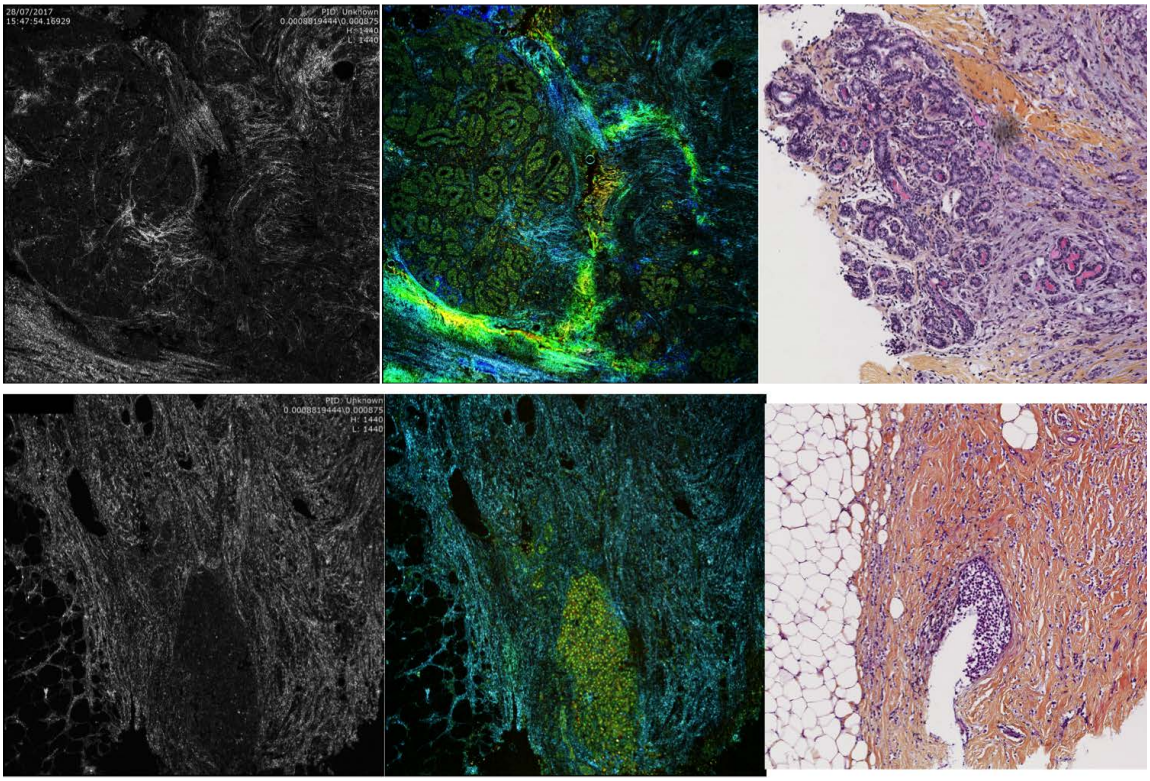


Figure III.4: Breast tissue FOVs acquired in (from left to right) FFOCT, DCI and corresponding area on H&E stained slide: normal lobule (top), ductal carcinoma in-situ (bottom)

Annotation

The ground truth diagnosis per FOV was set by a pathologist using the correlation with the corresponding area on the H&E slide. Despite the fact that the same structures are visible in both DCI and histology, finding the corresponding areas in both imaging modalities is a challenging task. One of the main difficulties is the fact that only a limited number of DCI FOVs are available. This difficulty is specific to the presented dataset. At the time of the study, the DCI acquisition was not fully integrated in the LightCT™ scanner and was too slow to comply with the time constraints of the clinical process. Consequently, the choice was made of limiting the areas of acquisition and keeping the raw files (which is also time-consuming).

Given these challenges, the labels of each FOV are set by the pathologist according to the spatial correlation with histology and confirmed by the expected FFOCT / DCI appearance, some uncorrelated FOVs have a low confidence diagnosis based on either modality, while some are left un-annotated. The per-FOV diagnosis is given in unstructured text form and contains information about the structures present: tumor, stroma, inflammation, fibrosis, normal lobules or ducts. In some cases there is extra information about the type of cancer (invasive or in-situ) and the presence of isolated cancer cells or lymphocytes. However, we have decided to analyze only the presence or absence of tumor in a binary fashion, since

the imaging technique is meant for fast intraoperative detection of cancer cells, not a detailed diagnosis and also given the limited dataset compared with the high complexity of the pathology.

Sampling

Considering the already fragmented quality of the data, the entire FOVs are considered for analysis.

III.2.2.2 Mammary Biopsies Clinical Study

Cohort

In the light of the findings of the aforementioned pilot study where it has been confirmed that DCI imaging is suitable for revealing the cell organization and tissue micro-architecture that help in differentiating healthy from tumoral breast tissue in the operating room, another clinical study was conducted to check the feasibility of using DCI in the radiologist's office at the time of the biopsy. On these grounds LLTech teamed up with the *Pitié Salpêtrière Hospital* and designed a cross-sectional clinical study with the aim to image biopsies coming from 204 specimens of either breast nodules or sentinel lymph nodes.

One gets to have a biopsy after a suspicious dense mass is discovered through a breast radiography (i.e. mammography) which becomes a standard screening procedure after a certain age. The standard procedure consists of minimally invasive imaging-guided core needle biopsy, i.e. punctuation of the breast tissue with a hollow needle that will extract a small tissue cylinder with a diameter of around 1.5 millimeters (14-gauge needle) and up to a couple of centimeters long, resulting in 17 to 20 mg of sampled tissue. To ensure a good localization of the targeted mass there are 2 approaches: stereotactic or X-ray (see Figure III.5a) and ultrasound (see Figure III.5b) guidance. Furthermore, to ensure correct sampling, an extra step might be taken in the act of radiographying the excised specimens, to look for microcalcifications, a sign of suspicious tissue. Nonetheless, multiple (usually 5) biopsying gestures are conducted to ensure the success of the of the act, then the specimens would be transferred to an anatomical pathologist who will perform a labor intensive and time consuming histopathology protocol and finally microscopic analysis (see Figure III.5e). By this means the patient will receive the diagnosis in 2 to 5 days.

Taking the example of the U.S., roughly 37 million mammograms are done every year and more than 1 million women have breast biopsies every year. In France there are performed 400 thousand biopsies per year due to systematic screening; at the *Pitié Salpêtrière Hospital* there are carried out around 500 biopsies annually. We can safely say that breast biopsies are a common procedure in the hospital setting. However, in France less than 50% of the targeted women respond the screening calls despite it being free of charge for the patient, one possible explanation for this is the stress and anxiety induced by the waiting time, as quantified in a study [107] measuring increasing cortisol (i.e. the stress hormone) levels

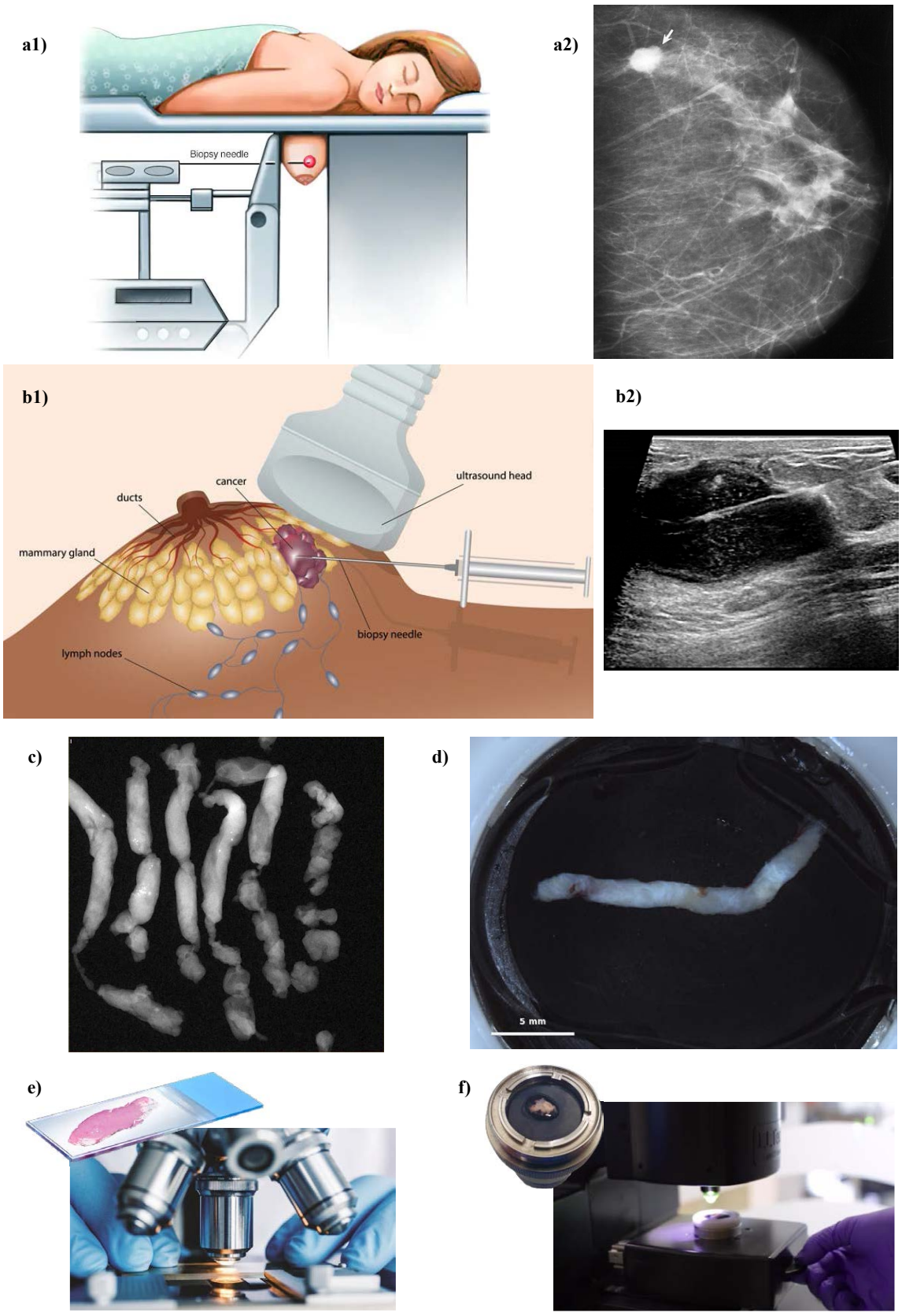


Figure III.5: Breast core-needle biopsy protocols and analysis: **a1)** stereotactic guided biopsy protocol **a2)** mammogram showing abnormal dense mass **b1)** ultrasound guided biopsy protocol **b2)** ultrasound image of breast with cyst penetrated by biopsy needle **c)** radiographs of core-needle biopsy specimens showing microcalcifications **d)** macro image of specimen acquired with the LightCT™ scanner **e)** analysis of prepared sample on glass slide under optical microscope **f)** analysis of fresh unprepared sample with LightCT™ scanner.

along the waiting period. Hence, an immediate diagnosis could ease the patient's psychological strain and lead to an improved medical care overall.

All things considered, the aim of the study to verify if a non-pathologist, particularly a radiologist can diagnose the biopsies based on DCI imaging and secondly, to prove the feasibility of using automated aid-to-diagnosis algorithms to help with the task; with the end result that the patient would be spared the stress caused by waiting for the - more often negative - diagnosis.

A major advantage of the DCI technique is that it can be seamlessly integrated in the existing protocol because the sample is not altered in any way and can then undergo normal histological preparation and analysis. What is more, the radiologists can easily manipulate the scanner without the need of technical assistance on the long run and he or she would only need to spare at most 5 minutes per sample.

At the moment the work presented in this manuscript was done, the study had not yet been concluded, hence roughly 120 (of the 204 goal) biopsied nodules are considered here, 86 coming from breast tissue and 33 from sentinel lymph nodes.

For the sake of clarity in the follow-up we need to define the following terms:

- *inclusion* = a breast nodule or lymph node; note that there could be multiple nodules per breast or per patient, but there is no patient information included, so each nodule is treated independently here;
- *sample* = a location of the biopsying gesture (1 to 4 imaged samples per inclusion); as mentioned before, multiple biopsies are acquired of the same nodule;
- *fragment* = a tissue entity imaged at once, however, there are rarely multiple tissue fragments from the same biopsied location.

Consequently, each image would be uniquely defined by the sum of these three identifications: inclusion, sample and fragment. The resulting collection of data will finally be comprised of:

- per fragment:
  - a macro image (Figure III.5 d) of the specimen used at the image acquisition stage for an aware tissue exploration;
  - a large-field processed DCI and its corresponding large-field FFOCT image for each fragment (see Figure III.6 for an example); there is a total of 276 fragments imaged and for the majority of cases (i.e. per inclusion) there are between 2 and 4 fragments imaged; images are ranging from $6\,\mathrm{MP}$ up to $200\,\mathrm{MP}$ with a median of $40\,\mathrm{MP}$ resolution, with the same pixel resolution of $0.9\,\mu\mathrm{m}$.

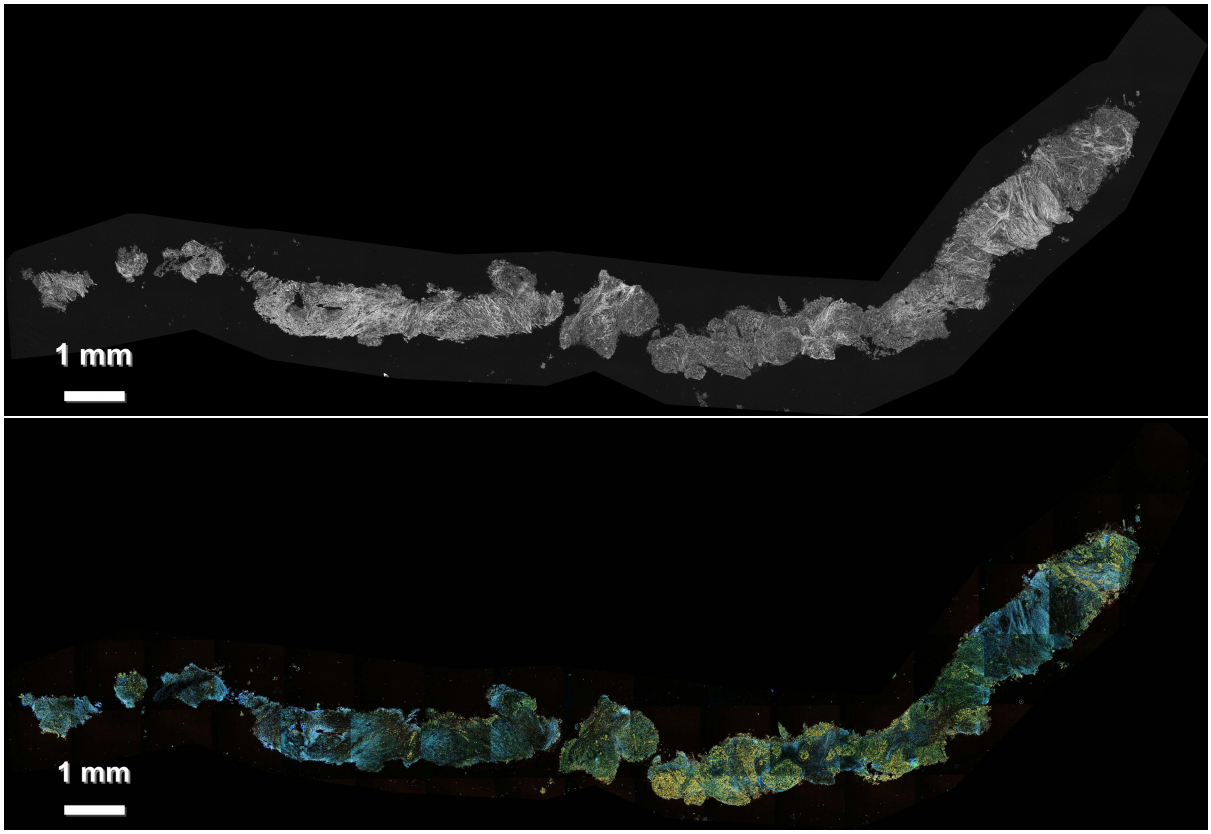


Figure III.6: Breast biopsy example in FFOCT (top) and DCI (bottom). Malignant, invasive ductal carcinoma of high grade.

- per inclusion:
  - a gold-standard H&E stained histopathology slide; note that all fragments from an inclusion are prepared on the same slide so an exact correlation might be burdened;
  - a pathology report containing the diagnosis.

Annotation

For this study annotations are directly extracted from the pathology reports issued by a qualified medical professional after analyzing the histology slides, so totally agnostic to the DCI acquisitions. In this scope, a laborious effort was dedicated to interpret the pathology reports and translated them from unstructured text of specialized lexicon into machine interpretable structured information through pertinent knowledge representation.

From correlating literature review on breast pathology [105, 108–111] with the terms in the 98 reports processed up to this point, we have extracted the following main attributes (number of cases between brackets):

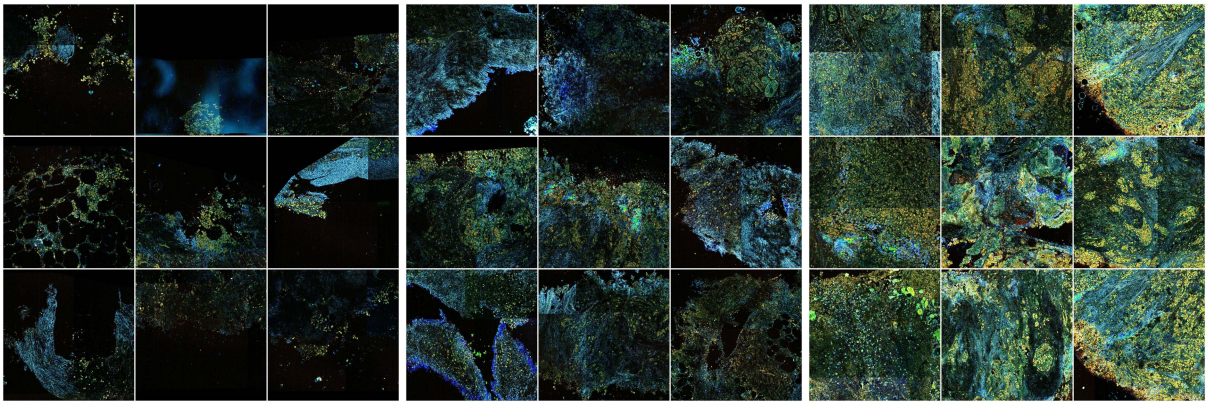
- **location**: breast (86) or sentinel lymph node (33);
- **malignancy**: malignant (33), benign (48) and normal (9), note that due to the nature of the medical act only lymph nodes were normal;

* for *malignant* tumors:
  - **proliferation grade**: low (6), intermediate (15), high (10), using the Elston-Ellis grading system [112] that sums up gradings on three criteria for **architectural disorganization**: low (3), intermediate (4), high (24), **nuclear pleomorphism**: low (0), intermediate (16), high (14) and **mitotic count**: low (18), intermediate (6), high (5);
  - **origin**: ductal (25) or lobular (4);
  - **proliferation extent**: in-situ (1), invasive (31), both (9);
  - **external invasion**: vascular (1) or lymphatic (5).

* for *benign* tumors or lesions:
  - **type**: fibroadenoma (17), adenosis (9) and papilloma (5) which are usually in conjunction with ductal hyperplasia, mastosis (6), hamartoma (2), inflammation (2) for breast and histiocytosis (5) for lymph nodes, characterized by an abnormal increase of immune cells;
  - **risk of progression**: in an attempt to group the multiple types of lesions found under a more compact categorization we defined the risk of disease progression, from low to high as follows: inflammation (7), non-proliferating (25), proliferating without atypia (12), proliferating with atypia (1).

It is important to observe that there is missing information in the reports, for example not all carcinomas are graded, or that some specific pathologies are not well represented to consider them for analysis separately, like the sub-types of benign tumors. Another obstacle in the annotation process is the fact that there is one diagnostic given per inclusion and the totality of specimens (i.e. fragments) from one inclusion were grouped together on the glass slide for the analysis, so there is no diagnosis per fragment.

Given the current state of the study with the incomplete cohort, the noisy annotations coming from the lack of correlation between the histology slides and DCI images and relying on our interpretation of the pathologist reports, we decide to further exploit this dataset for automated diagnosis in a straightforward manner. First, we separate the data coming from breast tissue and lymph nodes, as they are associated with two very different problems. We shall focus on the breast biopsies first as both the radiologist and pathologist who did a precursory assessment of the images deem metastasis detection from lymph node biopsies very challenging. Therefore, in the preliminary automated diagnosis application we shall



(a) low entropy **(b)** median entropy **(c)** high entropy

Figure III.7: Examples of extracted 1024×1024 px patches which all contain cells but have different entropy levels.

consider breast samples in a binary fashion as malignant vs benign. Lastly, we could treat benign and malignant in a regressive fashion by ranking their severity: from the benign tumors with the lowest progression risk up to malignant tumors of high proliferation grade (this way we could smartly train for high risk benign tumors which is sometimes advised to resect). With more data collection, the complexity of the automated tasks could increase, like cell counting, cell type detection, segmentation etc.

Sampling

As in the majority of successful aid-to-diagnosis applications on WSI, there is a need of breaking down the image into smaller entities i.e. patches, mostly due to the constraints that come with the leading methods in the field based on Deep Learning [113]. They are usually small in size 256 to 512 px^2, but this choice is strongly dependent on the problem formulation and nature of the data and annotation. For example, on our skin cancer dataset (Section III.2.1) we opted for small and densely sampled patches due to the important perk of having pixel-level annotation, meaning we could then label each patch with maximum confidence. On the other hand, for the other study on breast (Section III.2.2.1) the imaged patches (i.e. FOVs) were carefully selected to ensure they are informative for the diagnosis and then individually interpreted by a pathologist, so individually annotated.

For the task at hand, there is no special image curation, moreover, there is a global diagnosis for groups of images which does not mean it can be always extended to each image, therefore we try to keep to a minimum the further fragmentation of the images. In this regard, we choose the patch size in concordance with the width of the biopsying needle i.e. 1024×1024 px and we try to capture the entire surface of the imaged tissue with minimally overlapping patches. What is more, we opted for texture-aware patchification so we don't split up coherent morphological features, for that we used the

SoSleek method described in Section III.1.3, there is also a running example extracted from this dataset in Figure III.1. As a result we obtained a total of $2\,542$ patches from the breast biopsies.

In addition, we went a step further in trying to compensate for the lack of dense labels and did a manual assessment of the extracted patches as to indicate if they contain cells or not. This can be done by a non-medical expert who is familiar with the DCI modality, not to mention that cell presence is also a very important information for the diagnosis. Out of the $2\,542$ patches, in 842 or 33% cells were highly visible, in $1\,185$ or 47% there were mostly fibers, fat cells and sometimes few isolated and/or low-contrast cells, while 516 or 20% patches contained imaging artifact and were discarded. There is a median of 6 patches containing cells and 7 patches containing other tissue structures in each sample, while the maximal numbers of patches per sample with and without cells are 25 and 32, respectively. In hindsight, only 66.67% of the samples contain highly visible cells, at the global inclusion-level the statistics are more promising with 84.72%. Not distinguishing cells in an image is particularly problematic when it comes to detecting malignancy as the cell morphology and organization are the key bio-markers, 81.03% malignant samples coming from 92.86% inclusions contain cells.

Moreover, in order to define finer ranking between patches we turn to information theory and compute *Shannon's entropy* [114] for each patch. This is a measure which quantifies the amount of information in a sequence, it can also be interpreted as the number of bits needed to encode a piece of information. For an an 8-bit grayscale image x, Shannon's entropy is formulated as:

$$H(x) = -\sum_{k=0}^{255} p_k \log_2(p_k) \tag{III.1}$$

where p_k is the frequency (or probability) of pixels having intensity k from the patch x (converted from RGB to grayscale). Looking at Figure III.7 we notice that indeed the higher the patch entropy the more information they bear towards a successful diagnosis, as they have a higher number of cells and a better contrast.

To sum up, the current figures of the breast biopsies dataset (with images and annotations extracted from the correlated pathologist reports) are 150 samples (60 or 40% malignant) coming from 72 inclusions (28 or 39% malignant). From those samples there were extracted $2\,027$ minimally overlapping patches of size 1024×1024 px using texture aware sampling and annotated for the presence or absence of cells. The overall malignant to benign ratio is just slightly unbalanced and evaluates to 2:3.

III.3 Challenges

To sum up, in this chapter we gave the keys to building insightful and structured datasets from raw clinical information which led to building three datasets.

It is worth noting that the order of presentation of the datasets corresponds to the chronological order of acquisition, which also gives a snapshot on the evolution of the imaging technique: from the more "historic" FFOCT, to the pass to DCI when it was possible to image only some FOVs in the beginning, then to finally acquiring entire biopsies in DCI. Moreover, the granularity of labels associated with each datasets, from pixel level annotations, to patient level, is pointing to a tendency of diminution in supervision level as we increase in complexity and scale up the datasets.

These datasets, summarized in the following Table III.1, shall serve to build data analysis pipelines in the purpose of gaining precious insights about this revolutionary imaging technique that could change clinical practice for the better.

Table III.1: Working datasets overview.

Clinical Study	Modality	Image scale	Annotation level	Patch size	# patches	# annotations
Skin Cancer	FFOCT	whole-slide	pixel	256	100K	100K
Breast Cancer Surgical Margins	FFOCT DCI	patch	patch	1440	400	400
Breast Cancer Biopsies	FFOCT DCI	whole-slide	image group	1024	2K	150

Even if we are still in the context of big data, we are facing with the challenge of **limited data**, ubiquitous in the medical field due to patient data privacy regulations and expensive expert annotations; in addition, there are multiple unknowns related to the novelty aspect of the technique and the imaging nature itself:

- **single-center data**: as data is collected through targeted clinical studies as opposed to routine practice, for each application data comes from a single center and consequently, compared to the medical dataset that power state-of-the-art methods, we face all the more data scarcity;
- **label noise**: as of now there is no medical expert who can diagnose confidently based only on FFOCT / DCI images; image annotation resulted either from the collaboration between an imaging expert and a medical expert (i.e. pathologist) or directly extracted from the pathology report based on classical H&E histology preparation of the same sample. Therefore, the annotation procedure risks to introduce some noise in the labels due to the possible difference in tissue composition

visible in the mechanically cut histology slide as opposed to the preparation-free FFOCT / DCI optical slicing;

- **imaging artifacts**: the dynamic nature of DCI gives rise to image aberrations related to physical instabilities (external vibrations like air conditioning or a tap on the scanner's table) which can hamper the correct interpretation of the images, moreover, this noise is stochastic by nature and it is not known at this point how to model and by consequence filter it out;
- **undefined captured biological pathways**: the source of the DCI signal is not yet fully characterized, which implies that the criteria of cell appearance are not biologically validated; for example, it is still unclear if what we perceive as "cell" corresponds to the nucleus or the entire cell i.e. the cytoplasm, or how the intensity measured inside the cell perimeter which is believed to be representative for the "cell activity" could be correlated to biological processes.

In order to tackle these unavoidable challenges, and delve into the data in the most meaningful way, we took care of some aspects at the dataset curation step, especially via the adapted data sampling methods. Moreover, the methods to be developed - be they exploratory analysis or aid-to-diagnosis prototyping - need to follow some consequent **requirements**: firstly, the methods need to remain in the scope of *interpretability* in order to counter the unknowns about the data itself or even try to answer to some of these fundamental questions and secondly, methods need to be *versatile* and *extendable* to new applications and problematics as the technique evolves and its adoptability increases, a concrete example is the capacity to seamlessly scale up in model complexity with more data collection.

Given the enumerated points, it is clear that *data driven* (rather than model driven) approaches are suitable to accommodate all the underlying unknowns, therefore we should look at the powerful machine learning algorithms family dedicated to computer vision in order to explore the dynamic signal, the processed FFOCT / DCI images, their multimodal aspect and ultimately build effective aid-to-diagnosis solutions.

Chapter IV

Fundamentals of Convolutional Neural Networks

In the previous chapters we exposed the real-world conditions of data curation together with the particularities of the data which exposed multiple challenges. With the aim to surpass the difficulties and achieve adapted automated aid-to-diagnosis methods we will exploit various approaches from the field of *deep learning* which is considered the silver bullet for hard-to-define problems, by modeling the relationships between input and output of a system without much information about that system itself.

In this chapter we give some theoretical foundations on artificial neural network design and training dedicated to computer vision. There are brilliant reviews [115] and books [116] on deep learning theory, but we shall still introduce some elementary methodological aspects that concern our methods presented in the later chapters. Accordingly, in this chapter we introduce some basics of DL theory in the context of supervised image classification, with a deeper focus on *Convolutional Neural Network* (CNN) based architectures, as they represent one of the main pillars of the computer vision field. Moreover, as this work regards medical application, we bear special attention to model validation. Hereof, we present some model acceptability strategies meant to both safeguard the model performance and also bring interpretability - in an effort to overcome the black-box nature of neural networks.

Contents

IV.1 Computer Vision

A computer only "understands" numbers, therefore, images need a numerical representation. A grayscale image can be stored as a matrix, where each cell represents one pixel of the image and the cell value represents the gray level i.e. intensity of the pixel. In computer vision, it is not very effective to consider the attributes of the data to be the raw pixels, but to define some **features** which encode the data; this can also be seen as a dimensionality reduction problem or representation learning and stands at the core of all computer vision applications.

The classic approach for image classification is **bag-of-features** [117]:

1) extract salient keypoints (SIFT [118], SURF [119], HOG [120]) from a set of images;
2) split those keypoints in k groups by K-means clustering;
3) each cluster is defined by its center, all the cluster centers (which are nothing else but patches of the images: can be textures, small object parts etc) form a vocabulary;
4) encode each image in the dataset using the vocabulary;
5) train a classifier (e.g. SVM) in the space of the vocabulary.

Another approach is to have pre-defined so-called **filter banks** [121], independent from the dataset (e.g. Gabor filters, Wavelets basis, etc.). A filter convolved with an image gives a feature map which brings lower level information like the presence of edges with a certain orientation. Here learning consists in discriminating the data according to their responses to the filters.

Those paradigms are not scalable, so they have mostly been dropped in favor of artificial neural network based approaches which are able to encode more complexity, namely convolutional neural networks (CNN) which are dominating the computer vision field.

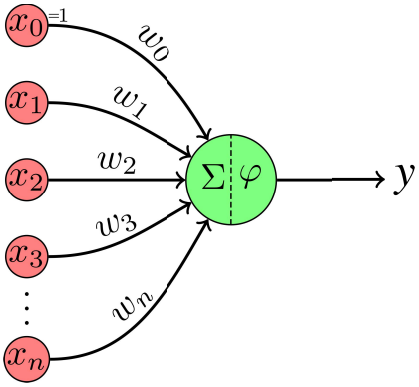


Figure IV.1: The perceptron a.k.a. the artificial neuron (source: Creative Commons).

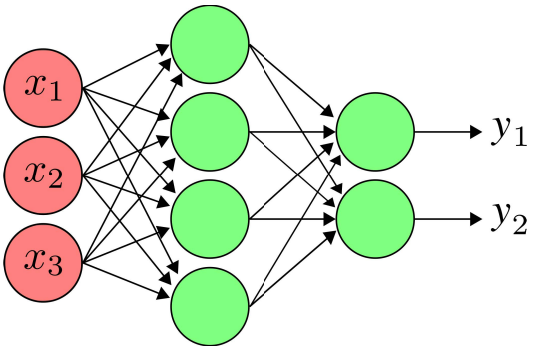


Figure IV.2: A multi-layer perceptron (MLP) with one hidden layer (adapted from Creative Commons).

IV.2 Model Implementation

Deep learning is synonym with deep **neural networks**, which represent a biologically inspired computational model that consists of interconnected layers of artificial neurons. The output of one neuron represents the input of one or more neuron from the next layer and there are no connections between neurons of the same layer, so a neural network is an acyclic graph. A neuron can serve as a stand alone weak classifier, so a network can be seen as a combination of weak classifiers that form a strong one. The *intelligence* of a network resides in the weights of its neurons, which also represent its adjustable parameters. To get to correctly solve a task, the network adjusts its parameters through a learning mechanism that is based on the *trial and error* principle.

IV.2.1 The Artificial Neuron

The single layer **perceptron** [45] is an algorithm based on an artificial neuron, i.e. the building block of any neural network architecture regardless of its complexity. An artificial neuron by itself works as a linear binary classifier. Consider a feature vector $x \in \mathbb{R}^n$ that is used to predict the probability y of occurrence of a certain event. Each input value x_i is scaled up or down according to its corresponding weight w_i and then the sum of the products is fed to an activation function φ which mimics the biological excitation of a neuron, to finally produce the output y (see Figure IV.1). In practice a bias b is added to the weighted sum to better fit the data.

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{IV.1}$$

The role of the **activation function** is to work as an on-off switch (i.e. the term "activation") which is best modeled by the *Heaviside a.k.a* step function which is 1 if its input value is positive and 0 if else, however, in nature, the change of state does not happen instantly so the *Sigmoid a.k.a* soft step function

was introduced $\varphi(z) = \frac{1}{1+e^{-z}}$ to mimic natural synapses. In modern architectures, *rectified linear logic unit* (ReLU) is generally used because it accelerates the convergence of the learning mechanism to a factor of 6 [50] and because of its simpler definition $\varphi(z) = max(0, z)$, so it acts as a threshold that removes weak signals. Another very popular activation function used in multiple output networks is the *Softmax* function which outputs the probability of the result belonging to a certain set of classes, defined as: $\varphi(z_i) = \frac{e^{z_i}}{\sum_j^m e^{z_j}}$ for $i = 1, ..., m$ and $z \in \mathbb{R}^m$. In mathematics, the Softmax or normalized exponential function is a generalization of the logistic function that squashes a m-dimensional vector of arbitrary real values to a m-dimensional vector of real values in the range [0, 1] that add up to 1. In probability theory, the output of the Softmax function represents a probability distribution over m different outcomes.

Initially, the weights and bias are randomly initialized and the prediction is calculated given the input values, if the predicted output is the same as the desired output, then the performance is considered satisfactory and no changes to the weights are made. However, if the output does not match the desired output, then the weights need to be changed to reduce the **error** ε, defined as their difference, for example. This is done using repeated updates (bounded by a fixed number of iterations or a convergence criterion on the maximal accepted error for example. Updates are done $w \leftarrow w + \eta \cdot \varepsilon \cdot x$ in steps of magnitude η, representing the **learning rate**, which is set prior to training. A learning rate that is too small leads to extremely slow convergence, while a learning rate that is too large can hinder convergence and cause the loss function to fluctuate around the minimum or even to diverge.

Single layer perceptrons are only capable of learning linearly separable patterns, therefore, even functions as simple as the logical XOR cannot be represented by one neuron only [122].

IV.2.2 Training Artificial Neural Networks

By stacking multiple perceptrons on top of one another i.e. multilayer perceptron (MLP), more complex functions could be encoded, moreover, according to Cybenko's theorem [123] MLPs are universal function approximators.

Neurons are interconnected forming neural networks, the output of one transfers to the input of another through **forward propagation**. This configuration (see Figure IV.2) would not be effective without the mechanism of propagating the error from the output through all the neurons in the hierarchy, i.e. **backpropagation** [46], a generalization of the least mean squares algorithm in the single layer perceptron. Backpropagation distributes the prediction error to all interconnected neurons in a path, proportionally to their contribution to the error.

As opposed to the perceptron, as we are dealing with more complex and diverse problem formulations, the error between the true expected output y and the predicted output $\hat{y}$ can be computed using a variety

of functions i.e. **loss functions**. The choice strongly depends on the problem objective, for regression mean square error or absolute error are used, while for classification cross entropy loss a.k.a. log loss is the most used. Cross-entropy loss measures the performance of a classification model whose output is a probability value between 0 and 1, the loss increases as the predicted probability $\hat{y}$ diverges from the actual label y, penalizing especially those predictions that are confident and wrong.

$$L(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{IV.2}$$

A perfect model would have a null loss, in order to minimize the loss L, backpropagation is used to adjust the weights accordingly. Mathematically, it implies computing the gradient (or derivative) of the loss function with respect to the weights of a multilayer stack of neurons $\nabla_W L(W)$ by applying the *chain rule* for derivatives: for $y = f(h(x))$, $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial h}\frac{\partial h}{\partial x}$. Then, with the derivative of the loss computed with respect to each weight in the network, the weights are adjusted in the negative direction of the gradient (or derivative) with a scaled step modulated by the learning rate: $w_i \leftarrow w_i - \eta \frac{\partial L}{\partial w_i}$, i.e. **gradient descent**.

There are several **optimization algorithms** [124] used for implementing different gradient descent strategies. First of all, the standard strategy by computing the loss on the entirety of the dataset is almost impossible in the big data context. At the other end lays the *stochastic gradient descent* (SGD) which implies computing the loss and adjusting the weights for one example at a time, however, this leads to noisy training and potentially converging to a suboptimal local minimum or not at all. A compromise between computing the true gradient and the gradient at a single sample is to compute the gradient against a subgroup of training samples (called a "mini-batch") at each step. Because it is so widely used, the naming convention accepts the term stochastic gradient descent as actually describing the mini-batch gradient descent. A variant of SGD implies adding *momentum* [125] to accelerate training. The weights are modified through a momentum term, which is calculated as the exponentially decaying moving average of past gradients, making it conceptually equal to adding velocity. The momentum term β can be seen as air resistance or friction which decays the momentum proportionally. β can take values between 0 and 1, in practice higher values are preferred $\beta \geq 0.8$, while for $\beta = 0$ the formulation is identical to classical SGD.

$$\begin{aligned} \Delta w_i &:= \beta \Delta w_{i-1} + \eta \frac{\partial L}{\partial w_i} \text{ , where } \beta \in [0, 1] \\ w_{i+1} &:= w_i - \Delta w_i \end{aligned} \tag{IV.3}$$

Adaptive Moment Estimation (Adam) [126] is based on the idea of tailoring the learning rate for each parameter during training, i.e. increase η if the descent has constant direction and decrease it when

descent direction is changed. Adam keeps an exponentially decaying average of past gradients similar to momentum m_i and an exponentially decaying average of past squared gradients v_i. m_i and v_i are estimates of the first moment (the mean) and the second moment (the uncentered variance) of the gradients respectively. Adam takes two parameters, the exponential decay rates for the two moments, β_1 and β_2 which have the textbook values: $\beta_1 = 0.9$ and $\beta_2 = 0.999$. There is also a numerical stability term $\epsilon = 10^{-8}$.

$$\begin{aligned} m_i &:= \beta_1 m_{i-1} + (1-\beta_1)\frac{\partial L}{\partial w_i} \\ v_i &:= \beta_2 v_{i-1} + (1-\beta_2)\frac{\partial L}{\partial w_i}^2 \\ w_{i+1} &:= w_i - \eta \frac{m_i}{\sqrt{v_i} + \epsilon} \end{aligned} \tag{IV.4}$$

Interestingly, multiple works [127, 128] argue that although Adam converges faster, SGD generalizes better and thus results in improved final performance. In practice, Adam is the first-hand choice when determining the training strategy. Then, if Adam fails, SGD and variants are explored, together with other optimization "tricks" like learning rate decay and schedulers.

Thus, the components of a neural network model i.e the activation function, loss function and optimization algorithm play a very important role in efficiently and effectively training a model and produce accurate results. Different tasks require a different set of such functions along with a suitable network architecture that controls the information flow to give the most optimum results.

IV.2.3 Convolutional Neural Networks

MLPs are so-called fully connected networks, meaning that all input values are connected to each neuron in the hidden layer and all neurons in consecutive layers are interconnected, respectively. However, for the case of images, which are most often high dimensional, this implementation is hard to grasp. Moreover, this approach would be suboptimal for image inputs as MLP incorporates the pixel position in its logic, but for image recognition what is important is the presence of a feature, rather than its exact position. Convolutional Neural Networks (CNNs) are a class of network architectures designed for matrix inputs, especially images, as they are shift invariant by definition and require less parameters than an MLP would. Their connectivity pattern of neurons is inspired by the organization of the visual cortex, every neuron responding to a small area of the field of view called receptive field. In the scope of neural networks, this is implemented via sparse connections, meaning each neuron is connected to a small number o neurons (for example a 3×3 squared region) from the previous layer. In addition, the weights and biases are shared between adjacent nodes, property which makes CNN invariant to

translation. CNNs are a successful example of incorporating domain-specific information (in this case, for vision tasks) into the architecture design of a neural network.

CNNs get their name from the main operation in the network, the **convolution**. In image processing, a kernel, convolution matrix, or filter is a small matrix used for blurring, sharpening, edge detection, etc. This is accomplished by doing a convolution between the specialized kernel and an image which results in another image which is called the *feature map*. Convolution is performed by adding each element of the image to its local neighbors, weighted by the kernel. This is related to a form of mathematical convolution similar to cross-correlation. The result of an input image I convolved with a kernel K is computed as:

$$\begin{gathered} O = I * K \text{ with } I \in \mathbb{R}^{w \times h},\ K \in \mathbb{R}^{k \times k} \text{ s.t. } O \in \mathbb{R}^{(w-k+1)\times(h-k+1)} \\ O_{(i,j)} = \sum_{a=0}^{k-1} \sum_{b=0}^{k-1} K_{(a,b)} \cdot I_{(i+a,j+b)} \end{gathered} \tag{IV.5}$$

A **convolutional layer** is composed of more such kernels i.e. **filters**, so more features are learned simultaneously (in the literature the number of filters per layer can vary between 32 up to 512). If at the first layer the filters are applied directly on the image (which can have one or more channels, e.g. 3 channels for RGB images), for deeper convolutional layers, the input being represented by the previously obtained feature maps (which therefore have much more channels, equal to the number of filters in the previous layer), their filters will have multiple dimensions. Feature maps represent the response of the input after being convolved with a filter, a filter defines a feature and the map signals its presence in the input image.

After a convolutional layer, there is usually a **pooling layer**, which does not do any learning per se but downsamples the feature maps by either taking the maximum value (max pooling) or taking the mean value (average pooling) in an area. Therefore, a pooling has as hyperparameters the function and size of the area to apply it. The motivation behind downsampling is to reduce the number of parameters of the network and reduce the representation size of data which makes the network less sensitive to noise and helps generalization.

After several convolutions and downsamples which serve as feature extractors, one ore more **fully connected layers** are added. As the name suggests, their neurons are connected to every neuron from the previous layer and to the next layer. They can learn non-linear combinations of the features discovered before and they are charged with the high-level reasoning in the neural network.

CNNs exploit the property that images are compositional hierarchies, hence lower layers learn basic features (e.g. edges, blobs) and higher layers learn a combination of the features from previous layers (i.e. textures, object parts).

IV.2.4 Design Principles for CNN Architectures

IV.2.4.1 The Convolutional Kernel Hyperparameters

When designing CNN architectures, special attention needs to be paid when defining the properties of the convolutional layers. The choice of the number of layers and the number of filters per layer is intuitively governed by the complexity needed for mapping the data to the problem objective. However, the properties of the kernels, often overlooked in the literature, are in line with the important structures found in the input data. The kernel is defined by its size k, stride s, padding p which together with the properties of the previous neurons in the hierarchy define the receptive field size in the input r. It is by controlling the size of the receptive field that we could indirectly inject information about the size of the objects present in the input that would influence towards the final prediction.

Every kernel "looks" at a certain area of the input image, performs multiplications, then moves a set number of pixels (stride) and repeats. While its default is usually 1, a stride of 2 can be used to simultaneously downsample an image, similar to pooling. Padding is used to capture the patterns at the edges of the input, a padded convolution will also keep the spatial output dimensions equal to the input. In the above equation there were used the default values $s = 1$ and $p = 0$. This area in the input space that a particular CNN's kernel is looking at is called *receptive field*. Since CNNs are deep, meaning they stack multiple convolutional layers, the receptive field for each layer is different. If the layer is deeper in the architecture then its receptive field will be larger because its input space is represented by feature maps from previous layers, i.e. already downsampled input image.

Intuitively, for the first convolutional layer $l = 1$ the size of the receptive field r_1 is equal to the size of the kernel k_1, but for each i-th convolutional layer corresponding to a kernel with size k_i, stride s_i and padding p_i, the receptive field size r_i is computed iteratively with respect to the previous layer, as follows: $r_i = r_{i-1} + (k_i - 1)j_{i-1}$, with the cumulated stride (or jump) $j_i = j_{i-1} \cdot s_i$. By merging those two expressions, the size of the receptive field of the i-th layer can also be written as:

$$r_i = r_{i-1} + (k_i - 1) \prod_{l=1}^{i-1} s_l \tag{IV.6}$$

Moreover, the size n of the output feature map for each i-th layer is obtained with the formula:

$$n_i = \frac{n_{i-1} + 2p_i - k_i}{s_i} + 1 \tag{IV.7}$$

For computing the receptive fields of a neural network, there is an open-source library developed by Google Labs, with its mathematical foundations explained in [129].

A trick for increasing the receptive field size while keeping the same number of parameters is the dilated or *atrous* convolutional kernel [130]. Dilated convolutions introduce another parameter to convolutional

layers called the dilation rate. This defines a spacing between the values in a kernel. A 3×3 kernel with a dilation rate of 2 will have the same field of view as a 5×5 kernel, while only using 9 parameters.

IV.2.4.2 Controlling Information Flow

Another key aspect in designing CNNs is setting up the actual connections between layers which define the information flow in the network. Legacy single-path feed-forward networks share the same design principles, i.e. alternating convolutional with pooling layers and tailed with fully connected layers, but newer models increase in complexity to accommodate richer datasets, from MNIST - numeric digits, to ImageNet - thousands of natural images of various scenes. In 1998 what is considered to be the first CNN - *LeNet* [47], with its two convolutional layers, was introduced for handwritten digit recognition. In 2012 appears *AlexNet* [50] which is considered to be the first deep CNN with its 5 convolutional and 3 pooling layers. In 2014 *VGG* [131] introduces the idea of smaller consecutive kernels rather, favoring 3 stacked convolutional layers with 3×3 kernels each, which manage to capture an area of 7×7 in the input, as opposed to a single neuron with a 11×11 kernel like AlexNet. The VGG approach allegedly adds more reasoning capacity due to the adjacent activation functions of the neurons. VGG consists of 5 blocks of pairs or triplets of convolutional layers and one pooling layer summing up to 13 convolutional layers and 5 pooling layers.

It seems like going deeper solves the need for accommodating more complexity, however, there are mathematical limitations to this, as the deeper the network, the more difficult becomes the training, because the problems of vanishing or exploding gradients could arise. In this regard, networks tend to get "wider" instead of deeper, *Inception* [132] architectures introduce the concept of parallelism and sub-blocks acting as a "network in network". The inception block consists of multiple kernels (usually of different sizes) applied to the input, then, the resulting outputs are concatenated and sent to the next layer. Another type of sub-block, implemented in *ResNet* [133], is the residual block which introduces skip connections, i.e. the activation of a layer is fast-forwarded to a deeper layer in the neural network. Another family of CNN blocks are Squeeze and Excitation blocks, introduced by *SE-Net* [134], they enable the network to perform dynamic channel-wise feature recalibration by reweighting the channels of each layer with respect to their average activation to achieve a lightweight attention model.

Other approaches to control the information flow in CNNs are embodied by weight sharing, multi-stream architectures taking multiple inputs (e.g. siamese networks [135]), encoder / decoder architectures for image segmentation (e.g. *Unet* [136], *MaskRCNN* [130]), or generator / discriminator generative adversarial networks (GAN) [137] for image synthesizing, etc.

In this section we have merely scratched the surface of neural networks design and training, however, we presented the main theoretical building blocks and design principles which guide the development of the applications making the subject of this manuscript.

IV.3 Model Validation

IV.3.1 Quantitative

IV.3.1.1 Classification Metrics

Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. The **confusion matrix** provides a more insightful picture which is not only the performance of a predictive model, but also which classes are being predicted correctly and incorrectly, and what type of errors are being made. To illustrate, we can see how the 4 classification metrics are calculated (TP, FP, FN, TN) based on the predicted value compared to the actual value in a confusion matrix is clearly presented in the confusion matrix (see Figure IV.3).

- **true positives** (TP): correctly diagnosed pathological cases;
- **true negatives** (TN): correctly diagnosed healthy cases;
- **false positives** (FP) - type I error: healthy cases predicted as pathological;
- **false negatives** (FN) - type II error: missed pathological cases.

For diagnostic tests in the medical field type II errors are more problematic than type I error, in other words, the cost of FN is higher than the cost of FP. Taking the example of cancer screenings, a false positive would surely lead to further tests (e.g. biopsy) that would eventually infirm the pathology, but a false negative case would be dismissed with a strong chance the pathology silently progresses until a belated examination, greatly jeopardizing patient's outcome.

Accuracy is the most common evaluation metric in classification problems, that is the total number of correct predictions divided by the total number of predictions made for a dataset.

$$\text{Accuracy} = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN} \tag{IV.8}$$

However, accuracy is only useful when the target class is well balanced but is not a good choice with unbalanced classes. Say we had 99 images of a healthy tissue and only 1 image of a pathological case in our training data, our model would most likely always predict healthy, and therefore we would get

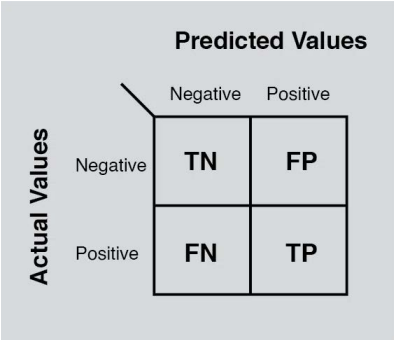


Figure IV.3: Confusion matrix definition.

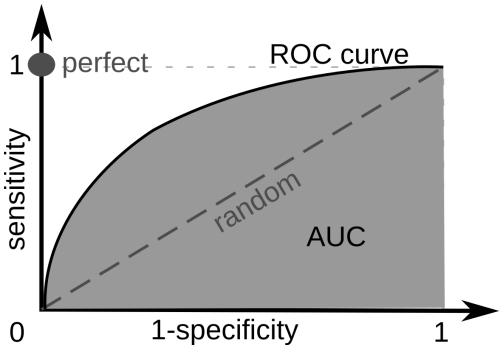


Figure IV.4: ROC curve definition.

99% accuracy, which is not reflecting the true model performance. Data is always imbalanced in reality, especially in medical diagnostics. Hence, if we want to have a full picture of the model evaluation, other metrics should also be considered such as **recall** and **precision**, which are metrics of relevance. Recall is simply the complement of the type II error rate and it represents the fraction of the positive cases that are successfully classified, while precision is the fraction of the predicted cases that are actually correct.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{(IV.9)} \qquad\qquad \text{Recall} = \frac{TP}{TP + FN} \qquad \text{(IV.10)}$$

Precision and recall are not useful when used individually. For instance, it is possible to have perfect recall by simply retrieving every single item. Likewise, it is possible to have near-perfect precision by selecting only a very small number of extremely likely items. One metric that combines the two together, capturing the global performance better than accuracy - as it is less sensitive to class imbalance, is **F1 score**.

$$\text{F 1 score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} = \frac{2TP}{2TP + FP + FN} \qquad \text{(IV.11)}$$

Nevertheless, in the medical literature we always find another pair of metrics, namely **sensitivity** and **specificity**. Sensitivity is reflecting a test's ability to correctly identify all people who have a condition and it is equivalent with recall or true positive rate (TPR); while specificity or true negative rate (TNR) is reflecting a test's ability to correctly identify all people who do not have a condition. For a diagnosis test to be useful in the clinical scope then sensitivity + specificity should be at least 1.5 [138], i.e. halfway between 1 - which is considered useless, and 2 - which is perfect.

$$\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN} \qquad \text{(IV.12)} \qquad\qquad \text{Specificity} = \frac{TN}{TN + FP} \qquad \text{(IV.13)}$$

Other such pair of metrics consists in **positive predictive value** (PPV) and **negative predictive value** (NPV). PPV is the probability that people with a positive screening test result indeed do have the condition of interest and is also equivalent with precision. In turn, NPV is the probability that people with a negative screening test result indeed do not have the condition of interest.

$$\text{PPV} = \text{Precision} = \frac{TP}{TP + FP} \qquad \text{(IV.14)} \qquad\qquad \text{NPV} = \frac{TN}{TN + FN} \qquad \text{(IV.15)}$$

Applying a classification model results is a mapping of instances to certain classes or categories. For most algorithms, the output of the classifier is an arbitrary real value (continuous output), its output is most often a probability value. Therefore, the said continuous output needs be translated into a categorical output. The predilection is towards **thresholding** at 50% to define the appurtenance to a certain class or not. However, the model can be tuned by choosing to interpret the probabilities using different thresholds that allow the operator of the model to trade-off concerns in the errors made by the model, such as the number of false positives compared to the number of false negatives. This is required when using models where the cost of one error outweighs the cost of other types of errors, i.e. type I error *vs.* type II error. Naturally, the best cut-off has the highest true positive rate together with the lowest false positive rate, but ultimately the choice depends on the clinical stakes.

A receiver operating characteristic curve, or **ROC curve**, is a graphical plot (see Figure IV.4) that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The method was originally developed for operators of military radar receivers during World War II (which led to its name) and was soon introduced to psychology to account for perceptual detection of stimuli. ROC analysis since then has been used in medicine, radiology, biometrics, model performance assessment and other areas for many decades and is increasingly used in machine learning and data mining research. The ROC curve is created by plotting the **true positive rate** (TPR) against the **false positive rate** (FPR) at various *threshold* settings. The true-positive rate is also known as sensitivity, recall or probability of detection, while the false-positive rate is also known as probability of false alarm and can be calculated as FPR $= 1 -$ TNR. Therefore a perfect classifier would have TPR $= 1$ and FPR $= 0$ at all threshold values, while a random classifier would have its ROC curve plotted as the second diagonal, therefore what lies below that line is considered a poorly performing model and what lies above is an acceptable model.

However, comparing plots is not a straightforward way to measure and confront the performance of multiple classifiers, therefore a scalar measure acting as a summary of the ROC plot is welcome. Computing the **area under the curve** (AUC) as the integral of the ROC curve is a robust metric widely used

in machine learning applications. An AUC of 1 corresponds to the perfect classifier, 0.5 to a random one, and in practice 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding [138].

IV.3.1.2 Cross-Validation

It goes without saying that in order to have a clear picture of the actual performance of a trained model, the aforementioned metrics should be computed on a different data set than the one used for training the model. Based on the models performance on unseen data it can be verified that the model is under-fitting, over-fitting or well generalized. In standard practice, a minority proportion (e.g. 20%) of the data set is held-out for testing, while the rest (e.g. 80%) is used for training. The split can be done in a complete random or a stratified fashion, meaning that the same class distribution is enforced on the two sets, this is useful with unbalanced data. However, if the entire data set is not big enough to ensure that all cases are well represented, this split might not be sufficient.

To ensure that every observation from the original dataset has the chance of appearing in training and test set, multiple models are trained on various data splits, i.e. **cross-validation** (CV) [139]. Depending on the size n of the datasets, there can be adopted either leave-one-out or K-fold split strategy. Leave-one out implies training n models on $n-1$ samples and testing on the remaining sample, however, this introduces a lot of computational overhead and the n models risk to be redundant and similar to a single model trained on the entire dataset. K-fold CV divides all the samples n in groups of k samples, called folds (if $n = k$, this is equivalent to the leave-one-out strategy), of equal sizes (if possible). The prediction function is learned using $k-1$ folds, and the fold left out is used for testing. The overall performance of a method is illustrated by the aggregated performance of the k models through the average and variation of the desired evaluation metrics. As a rule of thumb, $k = 5$ is often employed in practice which keeps consistent with the usual 80 / 20 train / test split.

IV.3.2 Qualitative

As opposed to classical step-based algorithms, where obtaining the expected output without any running errors is a sufficient indicator of a good execution, in the case of neural networks, they are failing silently. Given their black-box nature, it is difficult to understand the "steps" taken by the model to obtain a certain result, therefore, the only way to gain some intuition about the reasoning behind an obtained results is to employ some qualitative validation methods that "look under the hood", like visualizing and assessing the learned filters or feature maps.

IV.3.2.1 Learned Filters

Visualizing what the convolutional filters learned can be done by simply visualizing the weights, but this is only usually feasible for the first CONV layer, where the filters are applied directly on the raw pixel data, therefore, those filters belong to the input space and share its dimensionality. The shallow filters often take the shape of edge detectors (i.e. Gabor filters) when trained on natural images [50]. It is worth noting that those first filters were particularly informative in the case of big kernel sizes used historically, like 11×11 in AlexNet, but all current state-of-the-art CNN architectures have 3×3 kernels. Moreover, the filters coming from deeper CONV layers in the hierarchy, have multiple channels - on the order of tens or hundreds, making them impossible to interpret as is. Regardless, it is possible to also show a similar representation in the input space for the filters belonging to layers deeper in the network. One option is to iteratively test out which images in the dataset produce the highest activation of a certain filter and deduce the common triggering pattern by empirical observations and correlations on the retrieved images. In [140] they visualize the internal representations for an image via deconvolution, the input representation of each layer is projected back to the pixel space for understanding what information is kept.

Erhan et al. [141] proposes a filter activation maximization method that is completely agnostic to any input data. It produces synthetic images that would maximize the filters, making it a very powerful and straightforward qualitative interpretation method. Let Φ denote the neural network model and let $f_{ul}(\Phi, x))$ be the activation of a given unit u from a given layer l in the network; therefore, f_{ul} is a function of both Φ and the input sample x. Φ is fixed as it represents the trained network, the problem can be formulated as an optimization problem, we are looking for the (synthetic) input x^* that maximizes the activation.

$$x^* = \arg\max_x f_{ul}(\Phi, x) \text{ with } ||x|| = 1 \tag{IV.16}$$

The procedure implies that for a given filter u from a given layer l, we first randomly initialize x to an image of size (w, h), then at every iteration, we compute the gradient of the activation of the unit with respect to x and make a step η in the gradient direction: $x \leftarrow x + \eta \frac{\partial f_{ul}(\Phi, x)}{\partial x}$. The gradient updates are continued until convergence, i.e. until the activation function does not increase by much anymore. Note that after each gradient update, the current estimate is re-normalized to avoid very small and very large gradients and ensures a smooth gradient ascent process $x \leftarrow \frac{x}{||x||}$. The process is then repeated for all the filters of all layers in the network, to obtain a complete picture of the trained network.

Learned filters are useful to visualize because well-trained networks usually display nice and smooth filters without any noisy patterns. Noisy patterns can be an indicator of a network that has not been

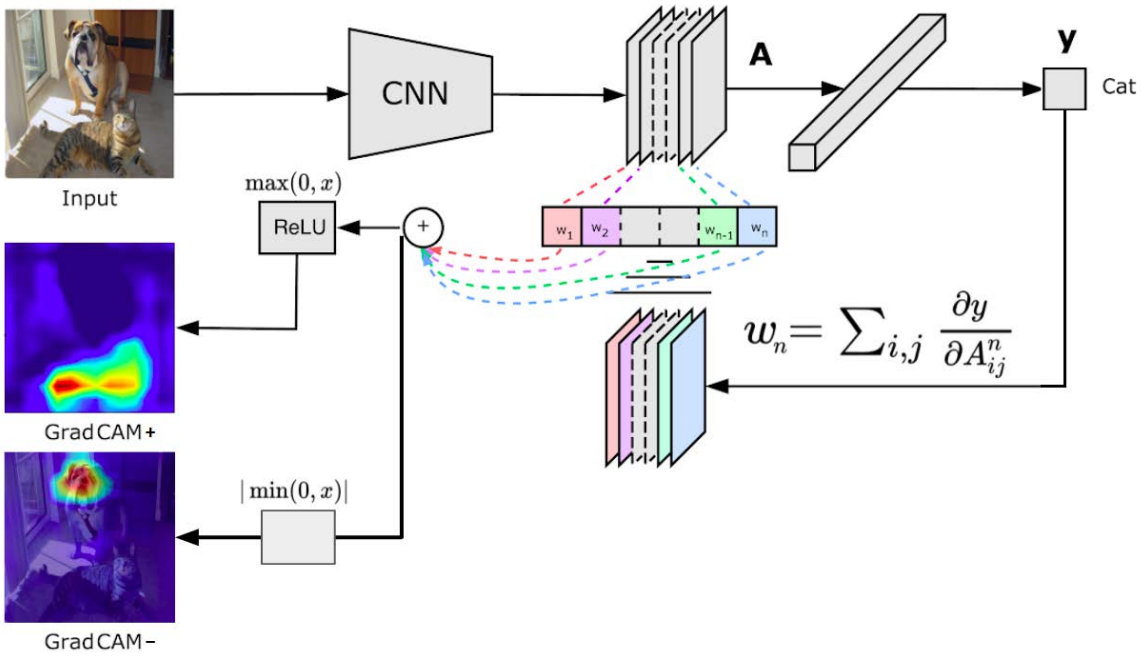


Figure IV.5: Grad-CAM algorithm (adapted from [142] by adding the negative branch and showing a binary classification case, rather than multi-class).

trained for long enough, or possibly a very low regularization strength that may have led to overfitting. Various visual examples of the learned filters obtained via this activation maximization method will be seen across this manuscript as we have used this framework systematically for model validation and disambiguisation.

IV.3.2.2 Attention Maps

Feature maps, or activation maps, are resulting from applying a certain convolutional filter on an input image. While they indeed reveal the localized abundance of the features encoded by the filter, they do not tell us anything about the importance of that feature in the final prediction. In order to make this connection, the relation between the filter and the output needs to be established; this is done by combining the activation map with the gradient flowing into it from the class output, implemented via the method *Gradient Weighted Class Attention Map* [142] (Grad-CAM). It produces a heatmap corresponding to the aggregated filters in a convolutional layer highlighting the areas contributing the most to a certain output. In the classification context, those areas should correspond to class-discriminative features for a well-trained model.

The method implies first obtaining the activation map cube A of a convolutional layer with n filters via the forward pass on a chosen input image through the network. Then, the gradients flowing back from the prediction of the output class y are computed with respect to each feature map activation A^n in the given layer $\frac{\partial y}{\partial A^n}$. These gradients are global average pooled over the spatial dimensions (indexed by i

and j) to obtain the filter importance weight w_n. Finally, the class activation map is obtained from the weighted combination of the forward activation maps A^n and their importance w_n.

$$w_n = \sum_{ij} \frac{\partial y}{\partial A^n_{ij}} \tag{IV.17}$$

$$M^+ = \max\left(0, \sum_n w_n A^n\right) \tag{IV.18}$$

$$M^- = |\min\left(0, \sum_n w_n A^n\right)| \tag{IV.19}$$

In standard practice, a ReLU function, i.e. $max(0, \cdot)$, is applied to the linear combination of maps to consider only the features which have a positive influence on the class of interest, i.e. pixels whose intensity should be increased in order to increase y. However, by performing the inverse operation we could obtain the strongest evidence *against* the target class y. See Figure IV.5 for a depiction of the method.

The result in a coarse heatmap of the same size as the convolutional feature maps, therefore a much reduced resolution as compared with the one of the original image, e.g. by a factor of 16 for the last layer of a VGG-16 architecture.

In the present work, we shall use attention maps computed via the Grad-CAM method to apprehend both positive evidence M^+ and negative evidence M^- to gain intuition about the trained models and ensure they are unbiased by localizing the structures of interest.

Chapter V

Healthy *vs.* Malignant Classification with Dense Label Supervision

After having presented the working datasets and briefly introducing the theoretical aspects of DL methods, together with the challenges and requirements imposed by both the imaging and the methodology, we shall dive into their application.

In this chapter we are touching to multiple aspects of the FFOCT and DCI imaging - both the processed images and the raw dynamic signal in the common purpose of cancer detection, formulated as binary classification between healthy and malignant instances. Being in the well-posed setting of dense annotations, we can leverage various fully supervised learning approaches.

Contents

V.1 Normal vs. Basal Cell Carcinoma from FFOCT Images

Correlated with the chronology of the technological developments of the imaging technique, we present a pioneering application which acts as a proof of concept on the feasibility of using automated aid-to-diagnosis for our unique imaging. It corresponds to the detection of basal cell carcinoma (BCC), a sub-type of non-melanoma skin cancer, which is homogeneous and easy to discriminate from normal tissue, especially in FFOCT. In Section III.2.1 we have expanded on pathology and the motivation of using rapid FFOCT imaging during the clinical act of removing skin tumors (i.e. Mohs surgery), as well as details about the collected dataset itself and the processing steps. In this section we discuss the details about the method chosen to detect BCC from normal skin in FFOCT images as well as the results obtained.

In the field of dermatology, most automated diagnosis applications are on macroscopic mole-like lesions, however, unique imagining techniques are used more and more.

In the last decade, neural networks conquered this field, with [143], which classifies cancerous lesions from macro images of the skin surface with an above-human performance. Still, to our knowledge, at the point of our published work [Mandache2018] detailed in this section, there was almost no research in automatic diagnosing for the FFOCT modality, let alone using deep learning methods.

V.1.1 Architecture

We started by experimenting with some popular architectures like VGG-16 [131] or InceptionV3 [132] which already power many imaging applications in various fields. What is more, we use the models which have been pre-trained on the huge ImageNet [48] database, as it is supposed to help improve the results due to the amount of information already encoded in the weights, offering a more stable starting point for learning, as opposed to random initialization. By fine-tuning these state-of-the-art architectures we obtained an accuracy of 89.30% and 90.79%, respectively, which we deem to be unsatisfying results

given the reductive problem formulation (very dense and high confidence annotations, small size input i.e. 256×256 patches) and given the early overfitting phenomenon. Overfitting (*i.e.* "memorizing" training data, rather than learning to generalize) was quite important and was fast to appear, therefore, we inferred that those architectures were too deep and complex for our data distribution and that data over-specification caused overfitting. We thus decided to build a custom architecture, shallower (i.e. having less layers and therefore weights to train) and with adapted receptive fields (and kernel sizes in consequence) which we shall train from scratch, resulting in a model that is able to learn a generalized distribution of our data, with respect to our two classes, *normal* and *BCC*.

The proposed architecture follows the classical design of a multi-layer CNN while having a smaller number of parameters than state of the art architectures. Nevertheless, it takes advantage of the ideas employed by VGG or AlexNet, like: 1) convolutional blocks: consecutive convolutional layers to capture larger input with a spare of parameters; 2) pooling layers: to reduce dimensionality, reduce redundant information and enforce generalization; 3) dropout layers: randomly masking a fraction of neurons at training time to avoid overfitting by enforcing different "thinking pathways"; 4) rectified linear unit (ReLU): activation function used to speed up the computations.

We built a 10 layer CNN including: the feature extraction part, composed of 4 convolutional blocks (with two convolutional layers each) followed by max-pooling with 25% dropout and a classifier consisting of two fully-connected layers of 512 and 64 neurons, respectively, each followed by 50% dropout, lastly, there is one output neuron whose firing signals the classification of the input patch as *BCC* or *normal*, respectively. The layers from the first blocks have 32 filters and the rest have 64 filters each, the kernel sizes of the convolutions vary from 7×7 and 5×5 to 3×3 as we go deeper into the network. See Figure V.1 for the illustration of the chosen architecture.

The choice of the layers kernels size is made in accordance with their corresponding receptive fields in the input image. Firstly, the receptive fields of the first convolutional layer is equal to the size of the kernel, i.e. 7×7 px, this is bigger than in the standard architectures (which generally have layers with 3×3 kernels at all depths), but this is an educated choice based on the size of the cells (i.e. $\leq 10 \times 10$ px) and the smoothing filter applied in the preprocessing step to remove noise having a 3×3 kernel itself. Then, the receptive fields of the following pooling layers are 14×14, 32×32, 52×52 pixels and finally, the last layer detect features confined in an area of 92×92 pixels (as opposed to 212×212 pixels for VGG-16). Given the relatively homogeneous cell organization of BCC tumors and overall the size of the structures present, defining the image textures in a confined area is enough to differentiate normal vs. pathological aspect.

The embedding layer is obtained by flattening the activations of the last convolutional layer which are of size $16 \times 16 \times 64$, therefore, resulting in a $16\,384$-dimensional vector, hence a 4-fold dimensionality reduction given the input size $256 \times 256 \times 1 = 655\,536$.

V.1.2 Training

The network has in total $8\,654\,369$ parameters to train, more than 97.3% of the total parameters belong to the classifier part (*i.e.* fully connected layers), while only 2.7% represent the filters encoding the features, this corresponds to $232\,417$ meaning $60\times$ less than VGG-16. The weights are initialized using the Glorot method [144] which is based on the idea that the gradients of each layer should follow more or less the same distribution at the beginning of training and it is proven to converge faster and towards a "better" minimum. The training process consists in minimizing a weighted variant of the *Binary Cross Entropy* (BCE) [145] loss. When computing it, class weighting applies a higher penalization for misclassifying cancerous class with respect to the under representation of the minority class: $1 \div 1.2$ (there is 1 normal sample for 1.2 cancerous samples).

$$L(y, \hat{y}) = -w_p \cdot y \cdot \log(\hat{y}) - w_n \cdot (1 - y) \cdot \log(1 - \hat{y}) \tag{V.1}$$

Learning is possible using a gradient decent optimization algorithm, for our application *Adaptive Moment Estimation* (Adam) [126] worked best. Adam is one of the adaptive methods of gradient descent whose particularity is that they adapt the learning rate (*i.e.* step of the descent) to the parameters, performing larger updates for infrequent parameters (*i.e.* the ones which were rarely updated) and smaller updates for frequent ones. Adam also multiplies the learning rate by the momentum (*i.e.* average of the previous gradients) providing accelerated optimization. The mini-batch gradient descent approach is a trade-off between computational accuracy and convergence time, so between batch (entire dataset) and stochastic (one example at a time) gradient descent. We chose a mini-batch of 40 samples as it was the biggest size that respected the memory constraints.

Training time was about a day (25 hours and 17 minutes) on an Nvidia Tesla P100 GPU for 2000 epochs (45 seconds per epoch).

V.1.3 Results

We obtained a classification accuracy of 95.93%, corresponding to a sensitivity of 95.2% and 96.54% in specificity at patch level. Figure V.2 shows a comparison between the ground truth labeling and the patches classified with our method. We notice that the cancerous regions are coarsely detected and, interestingly, the abnormal tissue that was unlabeled (so unknown to the network during training) is

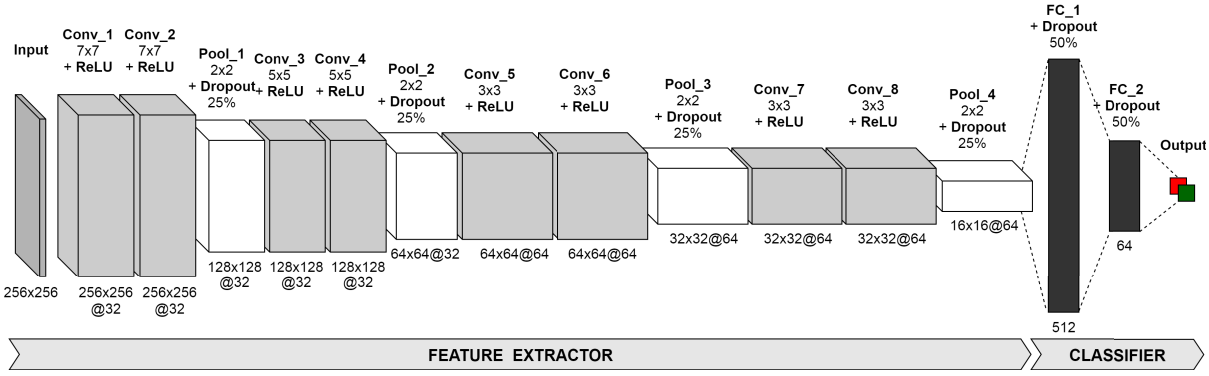


Figure V.1: Custom architecture for classifying skin cancer FFOCT patches.

classified as *BCC*. Note that background removal was not performed when testing, however, the method doesn't detect any abnormality outside the sample.

However, caution should be taken with the statistical results which can be misleading in interpreting the overall efficiency of the method and its behavior with different data. Since artificial neural networks are black-box models, gaining an intuition about the reasoning performed by the network is not straightforward. To do that we visualize what the network is learning. This is possible by viewing the weights of the neurons which correspond to the convolutional filters, but since they are very small, the textures encoded are not easily deductible. Still, to get the texture that a filter is responsive to, we can visualize the simulated input that would maximize the activation of its corresponding neurons. This is achieved by performing gradient ascent in the input space with respect to the filter activation loss (see Section IV.3.2 for more details on the method's formalism). In Figure V.3 are plotted the patterns learned by the 3rd convolutional filter. The patterns seem to represent different distributions of cells and orientation of collagen fibers. They could be reading criteria to make a diagnosis out of the images but this requires clinical confirmation from a pathologist.

V.1.4 Discussion

In this work we trained a CNN in the purpose of discriminating basal cell carcinoma from normal skin. We show preliminary results that open a promising research direction, which is analyzing FFOCT images with the powerful methods of deep learning. Developing computer-aided diagnosis tools could ease the integration of this novel "optical biopsy" in the clinical environment by assisting pathologists in their familiarization with the new modality and, ultimately, it could reduce the costs and duration of certain medical procedures, like Mohs surgery.

To improve our results we will firstly need a more consistent data set and also a better understanding of the decision flow of the pathologists in diagnosing the samples. This would allow us to translate the

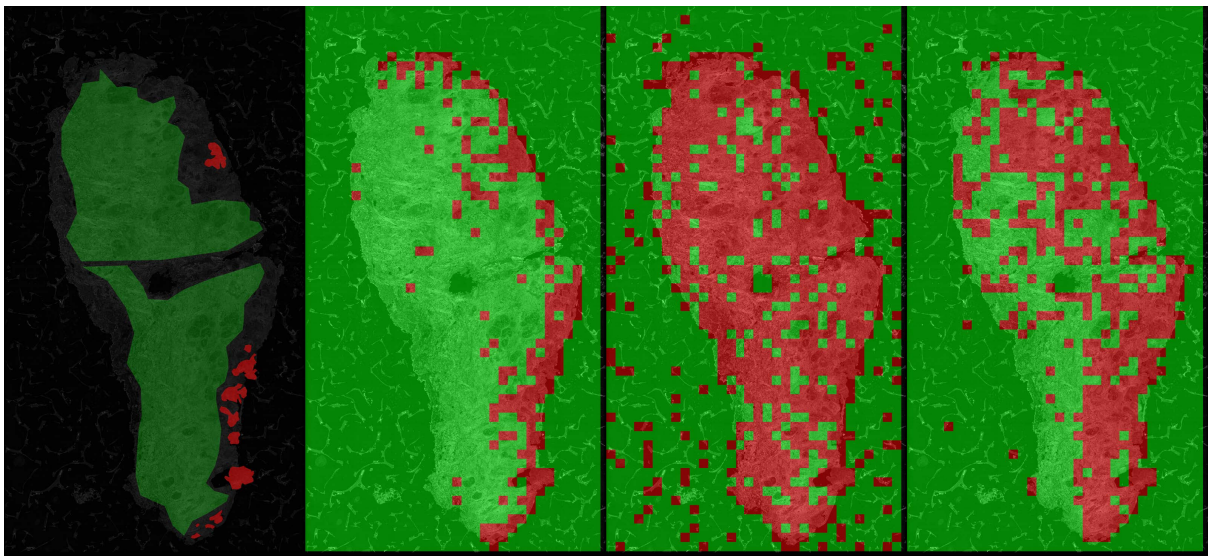


Figure V.2: Ground truth annotation *vs.* predictions obtained with **proposed** architecture, **InceptionV3** and **VGG16** (from left to right) on a skin sample imaged with FFOCT.

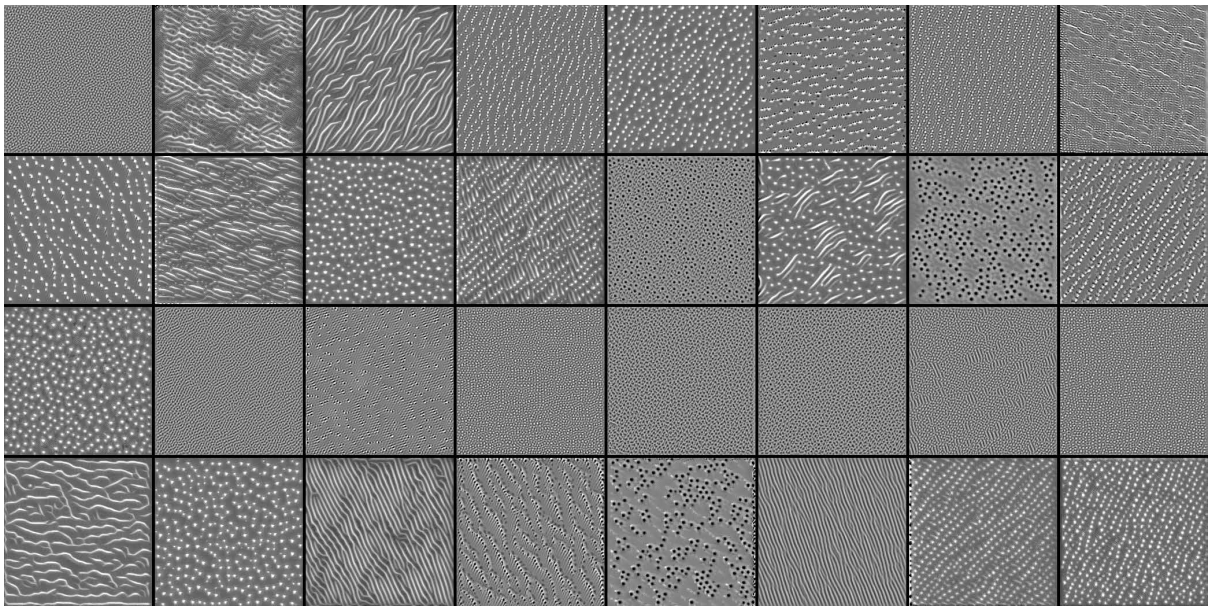


Figure V.3: Learned filters example.

knowledge of an expert to artificial neural networks design, still, in the present work we gave some hints on how to include a priori knowledge about the data and problem into the design of a CNN architecture.

Another introduced idea is that, in order to accurately assess the efficiency of such a model, we need to understand the reasoning learned by the machine. Therefore, we are ambitiously aiming towards demystifying artificial neural networks in the hope of also gaining knowledge about the data sets.

V.2 Normal vs. Breast Tumor from DCI Signal

Our analysis evolves together with the technique and, as DCI emerged as the counterpart of FFOCT, we are focusing on decrypting the DCI raw dynamic signal. In the scope of this work, we employ a source separation method in order to decouple the spatial and temporal information from the interferoemtric signal. However, since there is no ground truth at this level, we try to validate the decomposition by correlating with the known diagnostics on one of the datasets (Section III.2.2.1). In order to probe the importance of the metabolic signal revealed by DCI imaging, we will only take into account the dynamical profiles found in each FOV, and use them as a features towards classifying cancerous and normal tissue by exploring multiple tree-based models.

V.2.1 Feature Extraction

The motivation towards isolating different structures in the dynamic stack came from the prior intuition that there were multiple behaviors present: combined signals from the sample and perturbations (e.g. vibrations of the setup), multiple types of scatterers in the tissue (e.g. mitochondria and collagen), multiple sources of signal in one pixel (e.g. superposition of fiber and cell) or even at a lower scale, given the resolution of $1\,\mu\mathrm{m}$, different biological phenomena firing inside the cells at organelle level.

Therefore, a blind source separation approach is appropriate for tackling this problem. Suitably, we employed the Non-Negative Matrix Factorization (NMF) method for its highly interpretable results by virtue of its positivity constraint leading to part-based decomposition.

V.2.1.1 Dynamic Signal Representation

In order to extract the pertinent metabolic information and remove the incoherent part of the signal the raw interferometric (time) domain is transformed to the frequency domain. Starting with 1000 frames acquired at $150\,\mathrm{Hz}$ the next steps are performed per FOV:

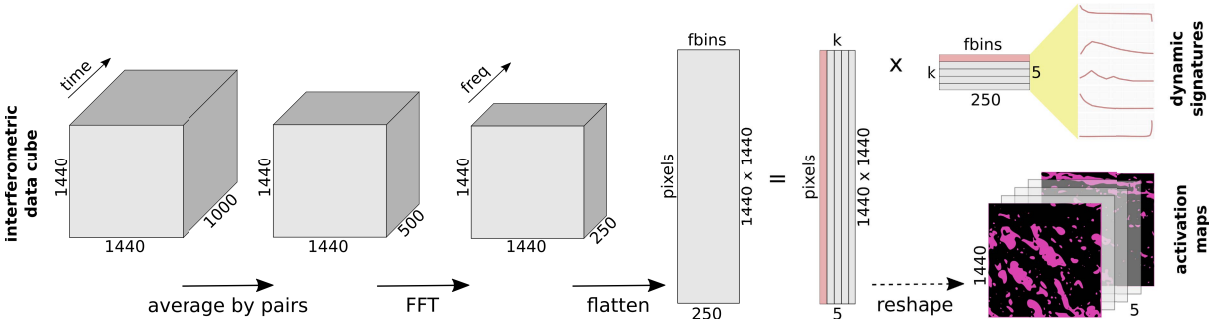


Figure V.4: Overview of pre-processing and decomposition algorithm.

1) normalize frame to constant energy to remove frame-to-frame inconsistencies introduced by the acquisition;
2) average the frames by groups of 2 to attenuate noise, obtaining 500 frames pseudo-acquired at 75 Hz;
3) pass to frequency domain with pixel-wise FFT obtaining 250 frequency maps with a step of 0.15 Hz (with respect to the Nyquist limit);
4) normalize FFT by its norm L_1;
5) pass to logarithmic scale to compensate the skewness of the amplitude towards low frequencies.

This results in a $1440 \times 1440 \times 250$ frequency stack holding both spatial and dynamical information which we will further decouple via source separation.

V.2.1.2 Non-negative Matrix Factorization

Introduced by Paatero *et al.* [146], and popularized by Lee *et al.* [147], NMF is successfully used in many domains [148]: hyperspectral imaging, audio source separation, topic modeling, face recognition, furthermore, biomedical domain where it gives excellent results in stain separation [149] and is used to segment cells in calcium imaging [150]. NMF formulates a feasible model for learning object parts, relevant to perception mechanism [151].

The purpose of NMF is factorizing a data matrix $X \in \mathbb{R}^{n \times d}$ into two low-rank positive matrices $H \in \mathbb{R}^{k \times d}$ and $W \in \mathbb{R}^{n \times k}$ representing the extracted feature basis and its corresponding activation, respectively: $X \approx W \cdot H$, where n is the number of data points, d the dimension of each data point and k the number of chosen components to split into. Finding the two composing matrices is achieved by minimizing the error (e.g. squared Frobenius norm - sum of squares) between the original data matrix and the result of the factorization: $\min_{W \geq 0, H \geq 0} \|X - W \cdot H\|_F^2$. Regularization can also be enforced on both matrices to introduce some prior knowledge into the model, like sparsity assumption e.g. 1) assuming there is a small number of components contributing to a data point through L_1 penalty on W or 2) piece-wise smoothness through L_2 regularization that would imply that neighboring data points

(pixels) are more likely to be characterized by the same components (if L_2 is applied to W). To solve this optimization problem the algorithm of multiplicative update [147] is used; it updates alternatively and iteratively for W and H in the direction of the gradient until convergence.

The NMF algorithm was applied individually on the flattened frequency cube of each DCI FOV, passing from $1440 \times 1440 \times 250$ to $2\,073\,600 \times 250$, so the spectrum of each pixel in the cube is treated as an individual data point, disregarding the spatial configuration. One drawback of NMF is the empirical choice of the rank of factorization k; it can be set using some prior knowledge about the data together with trial and error experiments. Given the lack of a validation metric, the optimal heuristic choice of rank $k = 5$ was based on qualitative assessment of activation maps and energy of frequency components. Accordingly, there were obtained frequency signatures $H \in \mathbb{R}^{5\times250}$ and their corresponding spatial activations $W \in \mathbb{R}^{1440\times1440\times5}$ (see Figure V.4 for the feature extraction pipeline). The revealed components correspond to:

- **baseline** signal: noise level of the signal, almost flat spectrum suggesting the FT of white Gaussian noise;
- **fibers**: high magnitude spectral components with its corresponding fiber-like structures in the spatial component;
- sampling induced **error**: the peak apparent in the last f_{bin} corresponds to the energy at the Nyquist frequency which seems to not be useful in practice;
- **cells**: cell shapes revealed in the spatial localization;
- motion **artifacts**: noisy frequency component with peaks in the higher part of the spectrum and spatial activation in the highly reflective fibers are clear indicators of a phase modulation of the DCI signal induced by external motion.

While they seem to offer proper signal separation further validation through biological experimentation needs to be conducted. Figure V.6 shows a representative example of the factorization with the obtained frequency components and their corresponding spatial activations.

To construct a **unified feature vector** for each FOV, the H components are ordered by their energy (area under curve) and the ones with the minimum and maximum energy are removed, since they correspond to the sampling induced error (H_2 in Figure V.6) and baseline component (H_0 in Figure V.6), respectively. Then the 3 remaining components are concatenated to form a single feature vector that will characterize each FOV. Note that ordering the components by their energy also ensures some consistency of the feature vector between FOVs.

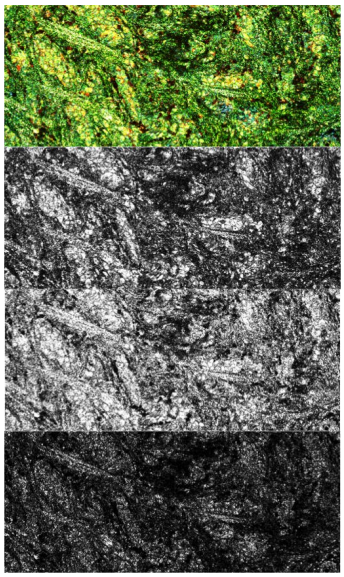


Figure V.5: DCI crop processed in RGB and the individual channels, showing poor signal separation.

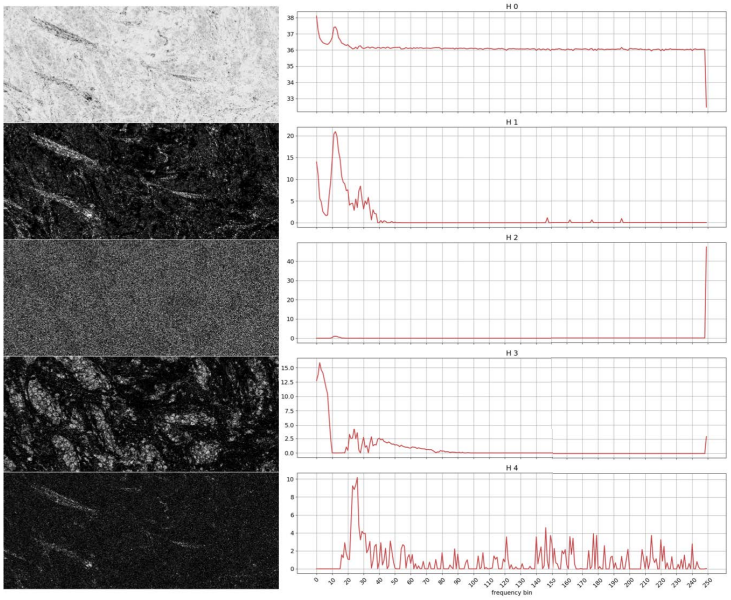


Figure V.6: NMF factorization results for $k = 5$: activations W (left) and signatures H (right) showing signal baseline, fibers, noise, cells, motion artifact (top to bottom).

V.2.2 Training

Training a single decision tree can be limiting in the sense that simple trees will have a large bias (over-simplification of the model - underfitting) while complex trees will display a large variance (lack of generalization - overfitting). The bias-variance trade-off is improved by ensemble methods, so combining multiple decision trees (weak classifier) towards building a stronger classifies. There are two main approaches: bagging [41], which consists in independently training multiple trees on random subsamples of data points and/or features and then aggregating their predictions by a voting mechanism and boosting [152] which incrementally trains trees on samples previously misclassified. In the proposed work, multiple tree-based classifiers were tested, from the simplest (single Decision Tree) to the more complex ensemble methods *i.e.* bagged trees (Random Forest, Extra Trees) or boosted (Adaptive Boosting, Gradient Boosting).

The splitting of the dataset into trainset and testset was done in a stratified manner, meaning that the class proportionality of the whole dataset was kept. Also, to tackle the class imbalance (thus, avoiding learning a biased model and also having clear interpretable performance metrics) an oversampling of the minority class (i.e. the normal class) was performed: for the train set we applied the SMOTE [96] algorithm which generates synthetic samples from interpolation and for the validation set we only applied random oversampling to avoid introducing any ambiguity in the performance metrics.

Table V.1: Normal vs Breast Cancer classification performance on NMF dynamic components: accuracy. sensitivity, specificity (mean percentage ± standard deviation).

Classifier	Train Accuracy	Test Accuracy	Test Sensitivity	Test Specificity
AdaBoost	90.95 ±1.39	**70.91** ±6.38	77.59 ±7.21	**64.22** ±10.37
XGBoost	91.38 ±5.66	70.69 ±5.21	82.33 ±5.5	59.05 ±8.48
Random Forest	98.13 ±0.32	65.73 ±7.93	**83.62** ±4.95	47.84 ±15.73
Extra Trees	96.77 ±1.81	65.52 ±4.00	78.45 ±5.52	52.59 ±10.73
Gradient Boosting	98.42 ±0.72	65.09 ±4.66	76.72 ±4.64	53.45 ±10.20
Decision Tree	99.93 ±0.12	57.54 ±3.41	64.66 ±3.11	50.43 ±4.46

V.2.3 Results

We trained multiple tree-based models using 4-fold cross validation: 75% of the samples for training (286 samples: 174 cancerous, 112 normal) and 25% for validation (96 samples: 58 cancerous, 38 normal). Only the lower half of spectrum (up to $f_{bin} = 120$) was considered since there was observed that the higher part of the spectrum has low SNR, hence overfitting on noise is avoided. Take note of some of the most important hyperparameters chosen: maximum tree depth = 10 (for bagging ensembles and simple decision tree) or 1 (for boosting models), number of trees in ensembles = 100. Table V.1 presents the classification metrics obtained. The model with the best generalization power proves to be AdaBoost, this can be deduced by the fact that it obtains the best accuracy, while also keeping consistency between train and test metrics. We also notice a lower specificity compared to the sensitivity which is due to the under-representation of the normal class in our dataset. The results are promising, being comparable with the other sensing-based non-invasive margin assessment techniques [17].

V.2.4 Feature Importance

As one of the main motivations for choosing this type of models was their semantic interpetability, we are looking at the most discriminating features as established by the best-performing algorithm (AdaBoost). They are highlighted in Figure V.7, plotted over the average components of the whole dataset. Feature importance is calculated for each attribute in a decision tree as the amount by which the split points over the considered attribute improve the performance measure. Then, for each feature, its importance is averaged over all the trees of the ensemble. In other words, for the given situation, feature importance is the ability and contritbution of an attribute (here frequency bin in a NMF component) to discriminate towards normal or cancerous class.

We observe the following frequencies appearing: the peak at frequency f=2.1 Hz corresponding to an oscillation time T=0.5 s for to the cell component, as well as the lower part of the spectrum corresponding

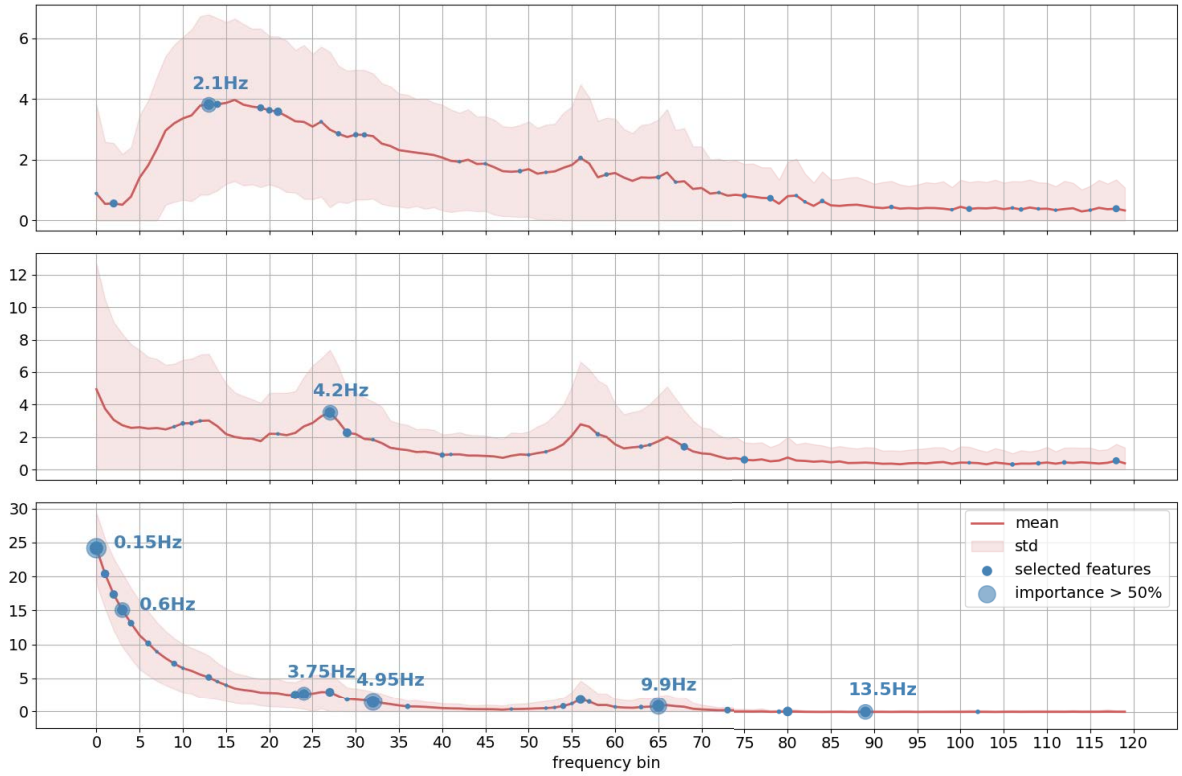


Figure V.7: Average and standard deviation of the features (3 NMF components) over the training set (orange) and the important f_{bins} selected by AdaBoost (blue).

to the more static fiber components f=0.15–0.6 Hz, T=2–6 s. However, for the peaks in the vicinity of f=4.2 Hz and 9 Hz we intend to further investigate their corresponding spatial maps to characterize their nature, but based on our preliminary observations we believe they correspond to vibration artifacts due to external nuisance.

V.2.5 Discussion

In this section, we test the feasibility of employing a blind source separation technique, namely NMF, to better extract the signal coming from different types of moving scatters in breast tissue imaged with the DCI technique in an interpretable and quantifiable way that can overcome the noise and motion artifacts. Here we used NMF decomposition to classify between cancerous and normal FOVs with 70.91% accuracy and we revealed some salient frequencies, but based on these results we conclude that the dynamic signal alone (therefore treating the scanner more like a sensing device than an imaging device) is not enough for diagnosis.

We believe NMF is a promising direction worth exploring for future research as it could help in better characterizing and understanding the dynamic signal. For example, applying a scaled-up NMF algorithm on more FOVs at a time could achieve separation between cell types by revealing some more meaningful

frequency signatures, as well as better defining the noise producing image artifacts in order to design useful filtering methods. This approach could also lead to an improved image formation algorithm. Another more ambitious application could be detecting the captured biological processes, however this would require a joint effort of data analysis and biological experimentation and hypothesis testing.

However, these directions are out of the scope of this thesis, not to mention that the raw signal is not stored in routine practice as it burdens the other processes which are already carefully optimized for leveraging efficiently the computation resources needed for fast acquisition. We will further focus on diagnosis based on the processed DCI images as they are currently used and collected during the clinical studies.

V.3 Normal vs. Breast Tumor from DCI Images

We are looking at the same problematic of breast cancer *vs.* normal tissue characterization as in the previous section, but this time based on the processed DCI images. In this purpose we explore, as in the earlier application on skin cancer detection from FFOCT images in Section V.1, a purely data-driven approach facilitated by the Deep Learning paradigm. However, for this application we are facing with a more complex problem due to the increased difficulty of the pathology itself (multiple breast cancer sub-types as opposed to one sub-type of skin cancer, i.e. BCC), but also of the imaging (DCI *vs.* FFOCT) and the dataset at hand, namely the annotation level (FOV or millimeter *vs.* pixel or micrometer).

To overcome these drawbacks we notice through multiple unsuccessful experiments that training from scratch is not possible in this case, therefore we shall leverage a pre-trained model. In [153] they have found through extensive literature review that transfer learning knew a surge in usage showing its efficiency on small datasets. Therefore, we decide to use ImageNet as pre-training database as it is uncomparably richer than any medical dataset. Moreover, initialization with ImageNet was proven to be about 15% more efficient compared to other specific datasets in the case of breast cancer H&E histology [154]. Thereupon we tested several state-of-the-art architectures (previously trained on ImageNet) as backbones like VGG16, InceptionV3, ResNet50, but also more complex information flow networks like SENet [134]. We prefer fine-tuning as opposed to just transfer learning in order to stay consistent with our secondary objective which is, besides diagnosis, trying to decode the imaging and consequently learning adapted features. Once again simplicity takes the stage with VGG16 performing best among the named architectures; in the next section we describe all the network design details and the reasoning behind.

V.3.1 Architecture

The CNN architecture is similar to the VGG16 [131] architecture with weights pre-trained on the ImageNet dataset [48], but with small modifications. We removed the classifier part and added a *Global Average Pooling* (GAP) layer followed by a fully-connected layer of 1024 neurons and an output neuron with sigmoid activation. The GAP layer allows network inputs of different sizes since it reduces each activation map of the last layer to a single value representing the mean excitation of the corresponding neural kernel over the entire input image, losing the spatial dimension. It results in a 512-dimensional (as the last convolutional layer has 512 filters) vector bottleneck between the feature extracting convolutional layers and the dense classifier layers. Using GAP in CNNs can be seen as the deep learning flavor of dictionary learning, where each learned filter encodes a concept (or visual word in the dictionary) and the GAP encodes the input representation in the dictionary space as the coefficients associated with each word i.e. marking the presence, absence or abundance. Another advantage of GAP is the reduction of network parameters, acting as a structural regularizer [155], which improves generalization and is particularly useful in the case of small scale data sets. Following the same reasoning we chose only one fully-connected layer. In terms of the receptive field size, the embedding layer covers features of 212×212 which is enough to enclose an entire lobule in cross-section [156].

With the presented configuration we obtain a network with approximately 15M parameters of which only 500K correspond to the classifier and the rest to the pre-trained weights, meaning that less than 4% parameters are trained from scratch, which could help convergence on the limited dataset.

V.3.2 Training

The network was trained on full resolution $1440 \times 1440 \times 3$ RGB fields of view with binary labels indicating the presence or absence of tumorous tissue, obtained from the pathologist's refined annotation per ROI. An important detail is that in tumorous ROIs, there might be portions of the image resembling healthy tissue.

The network was fine-tuned by minimizing the binary cross-entropy loss using the stochastic gradient descent (SGD) optimizer with a learning rate of $1\mathrm{e}{-4}$ and momentum of 0.8 on mini-batches of size 3 (due to memory constraints).

In order to validate the method and ensure model correctness, we ran a 5-fold cross-validation training with the same hyper-parameters and trained 5 models on partitions of $4/5$ of the samples and tested their performance on $1/5$ of the samples, respectively, hence keeping the same 80/20 train/test ratio at each run. The dataset partitioning into training and test sets was performed in a stratified manner with

respect to the global per-sample diagnosis i.e. ensuring the same distribution of classes in both sets ; in terms of size, 80% of the samples ($N^{train} = 37, N^{train}_{tumor} = 27, N^{train}_{normal} = 10$) used for training and the remaining 20% ($N^{test} = 10, N^{test}_{tumor} = 7, N^{test}_{normal} = 3$) for evaluating the performance. However, this stratification strategy does not ensure the exact same distribution of classes at ROI level, as each sample has a different number of acquired ROIs; nonetheless, taking the first fold as example, there are 286 fields of view for training (185 positive and 101 negative) and 87 for testing (60 positive and 27 negative). To compensate for the class imbalance, an importance penalization of the loss function was applied for each ROI, which is also known in the literature as **class weight**. In our case the healthy ROIs are less numerous so they will have a higher weight (i.e. 1.5 for healthy ROIs and 0.75 for cancerous ROIs).

In terms of data augmentation i.e. artificially increasing the number of data points by applying relevant transformations to the existing data, for this application we applied contrast stretching, with 3 look up tables per image, together with vertical and horizontal flips, which expanded the training set by up to 6 times.

Since we opted for fine-tuning, training time is significantly reduced compared to a training from scratch approach: consequently the phenomenon of overfitting occurs much faster. To avoid training beyond the optimal model, we have set two stopping conditions: (i) validation loss has not improved in the last 100 epochs or (ii) training accuracy has already reached 100%. With this, training lasts around 200 epochs and the optimal model is found somewhere between epoch 80 and 150 (depending on the fold i.e. the data split). Training time is around 6.5 minutes per epoch, and 20 hours per experiment. Thus conducting a 5-fold cross-validation experiment took around two and a half days (64 hours); note that in these delays we also included the lag introduced by logging performance metrics, as well as the overhead introduced by reading the image batches from the disk, and not only training (i.e. forward and backward propagation). Experiment tracking was made possible with the software *Neptune* (Neptune.ai), which helped organize and compare the performance of over 200 experiments conducted for this project, therefore allowing us to choose the optimal hyperparameters in an exploratory fashion.

V.3.3 Quantitative Results

ROI-level CV metrics

With a probability cutoff set to the standard threshold of 50% for ROI diagnosis, we obtained a per-ROI accuracy of $89 \pm 4\%$ which corresponds to $88 \pm 4\%$ sensitivity and $86 \pm 6\%$ specificity. Another metric that is worth mentioning, due to its lack of dependence on the probability threshold, is the area under the ROC curve (AUC), which is equal to 0.92 ± 0.02 (See Figure V.3.3).

Sample-level aggregation strategy

To aggregate the per-ROI predictions to a global per-sample diagnosis, assigning the maximum tumor probability prediction to the sample would be the most straightforward approach. This would translate to "if a sample contains at least one ROI with a tumor, then the sample is cancerous". This approach is however overly sensitive to outliers. On the other hand, the average or median are not suitable either because a bimodal distribution is expected (i.e. a sample most likely contains both healthy and tumorous FOVs), in other words small tumorous areas would be missed, or cancel out good prediction. Therefore, we turn to a statistical strategy and chose the 90^{th} percentile as a good trade-off between the mean and the maximum aggregations. This would translate into the predicted probability value that 90% of the ROIs fall into.

	Accuracy	Sensitivity	Specificity
P1	91 %	91 %	**92 %**
P2	89 %	94 %	75 %
avg(P1, P2)	90 %	93 %	83 %
Algorithm	**94 %**	**97 %**	85 %

Table V.2: Per sample performance of pathologists *vs.* algorithm on entire dataset (aggregated over folds).

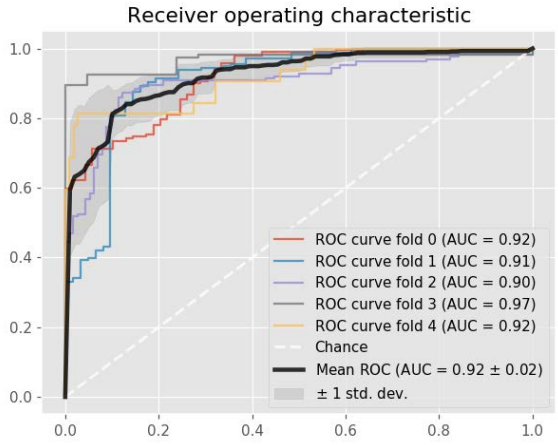


Figure V.8: ROC curves corresponding to all the 5 folds at ROI level.

Sample-level CV metrics

Accordingly, we compute the average and standard deviation over the 5 folds at sample level obtaining accuracy of $94 \pm 5\%$, sensitivity $95 \pm 10\%$, specificity $80 \pm 24\%$ and ROC AUC 0.96 ± 0.05. We notice a increase in accuracy and sensitivity and a slight decrease in specificity due to the strategy itself. The high variation of specificity over folds is due to the under representation of the negative class in the dataset which becomes more dramatic for each fold i.e. there are only a couple of negative samples per fold, i.e. one false positive would result in 50% sensitivity.

Comparison with pathologist performance

In order to remove the ambiguity related to data splits and have a clearer snapshot of the performance on the entire dataset, we compute per sample performance on the collated test sets predictions and obtain 94% accuracy, 97% sensitivity, 85% specificity and a ROC AUC of 0.96, which are actually caused by only 1 false positive and 2 false negatives.

Figure V.9: Detailed sample-wise cross-validated test results for the whole dataset: every subplot corresponds to an unique tissue sample; each point represents a FOV; on the X axis there is the FOV ID and on the Y axis there is the predicted tumor probability; the white horizontal line corresponds to the per sample aggregated prediction as the 90^{th} quantile of the FOV probabilities; colors represent the ground truth or the lack thereof: red for malignant, blue for normal, while the pale colors correspond to labels with uncertainty and white for lack of ground truth; the enclosing rectangles highlight the 6 samples missed by the pathologist(s) during the blind review, out of which 3 were correctly classified by our model and the other 3 represent the only misclassified cases.

By virtue of this computation we can now directly compare with the pathologist performance during the blind review. In Table V.3.3 we can observe that the algorithm performs better than the average of the pathologists with an increase of 2 to 4 percentage points. Moreover, in terms of accuracy and sensitivity the CNN performs much better than the respective best performance of the pathologists. For specificity, the pathologist P1 - who correctly diagnosed the most normal samples - manages to surpass the automated diagnosis, however it is worth considering that there is already an important disagreement between the two pathologists, with a performance gap of 17 points. Nonetheless, when we look at the sample count the differences seem less significant, as P1 misses 1, while P2 misses 3 and algorithm misses 2 out of the 13 negative samples.

Discussion

The detailed results containing the predictions for all ROIs in all samples (including the ones not included in the metrics due to their absent or uncertain ground truth) as well as the global predicted probability per sample are plotted in Figure V.9. There are highlighted with a square frame the 6 samples which were misclassified by (at least one of) the pathologists; 3 of those were actually correctly predicted by our models, while the other 3 are the only samples missed by the algorithm. This aspect can suggest that our trained neural network manages to encode expert level knowledge that mimics clinical reasoning. Another interesting aspect that could be noticed in the Figure V.9 is that the predictions on the ROIs having absent or uncertain labels do not change the outcome of the diagnosis, showing model robustness to difficult or ambiguous aspect of tissue under DCI imaging.

V.3.4 Qualitative Validation

Due to the black-box nature of deep neural networks and the sensitive application in medical diagnosis, we wish to establish confidence in the prediction based on visual feedback rather than just performance metrics, and to verify that the model's decision is not biased.

V.3.4.1 Class-wise Filter Bases with Linear Classifier

We recall that one of our main objectives was learning an adapted feature base, proper to tissue appearance under DCI imaging. In order fulfill this requirement, an end-to-end training was employed; however, to validate the extent of the features fidelity both to the data and the task, we need to adopt some strategies for demystifying the trained model. While the feature extracting convolutional layers are more straightforward to interpret (See Section IV.3.2), the classifiers on the other hand are much more opaque. Classifiers are most frequently (this application included) coded with fully connected layers which implies that all the inputs from one layer are connected to every activation unit of the next layer. Given the high dimensionality (e.g. 1024) of such layers and the dense information recombination, the

Table V.3: Compared metrics for MLP (i.e. proposed) *vs.* linear classifier (i.e. experiment) on features extracted from embedding GAP layer of VGG16 trained on proposed dataset *vs.* on initialization dataset ImageNet (i.e. benchmark); per ROI metrics corresponding to the test set of the first CV data split.

	Features	Classifier	Accuracy	Sensitivity	Specificity	ROC AUC
proposed	Breast DCI	MLP	96 %	91 %	100 %	0.99
experiment	Breast DCI	LogReg	95 %	97 %	92 %	0.95
benchmark	ImageNet	LogReg	91 %	80 %	88 %	0.84

logic of fully connected layers is fairly untraceable and non interpretable. Therefore, we experiment with a weaker classifier, namely the "transparent" linear model of logistic regression (LogReg).

"Weak" classifier proves "strong" feature base

Methodology-wise, we use the train/test data from the first CV split and its corresponding CNN previously trained as detailed in the previous sections. Firstly, we train a LogReg classifier on top of the existing GAP feature embedding layer, both of the trained model and the ImageNet trained VGG16 for comparison. We observe just a slight performance decrease (-1% in accuracy and -0.04 points in ROC AUC) as compared to the more complex MLP classifier, but a much bigger increase (+4% in accuracy and +0.11 points in ROC AUC) as compared to the benchmark features extracted from the ImageNet-trained CNN (See Table V.3). Based on these results, we can infer that the feature base learned is indeed representative for the diagnostic task and it holds most of the knowledge necessary for diagnosis, also showing that we trained a model that generalizes well.

Domain-specific *vs.* task-specific layers

Secondly, we reproduce this experiment for each convolutional layer in VGG16, aiming to find the diagnosis capacity of both the shallower and deeper layers. In this scope, we shall extract the data embedding from each convolutional layer by applying a GAP operation and then train a LogReg model on those features. Similarly, we compare the performance with the results obtained with the corresponding layer trained on a generic dataset. We can see in Figure V.10 the evolution of the ROC AUC metric as we go deeper in the network, as expected, the deeper the layers the most classification power they have, however the ImageNet-based features share a performance plateau between mid-level and high-level features, while for the adapted high-level features, the evolution of performance is conspicuous. Based on this experiment, we can infer that in our learned CNN model, the last 3 CONV layers encode task-specific features, while the other layers encode more generic domain-specific layers. This information could be used in the future in the case of extending this model to new tasks, as it would help to know which layers to continue training to adapt for the new task.

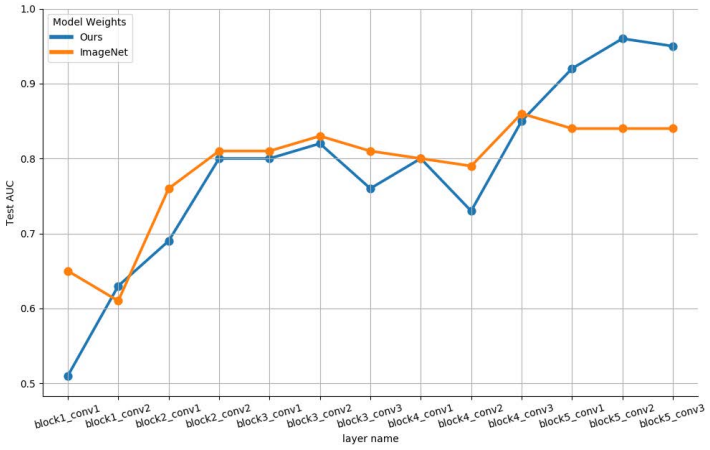
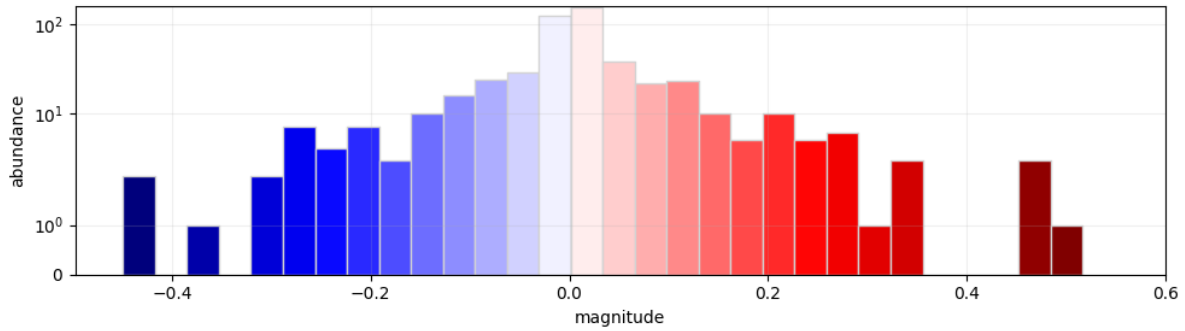


Figure V.10: Logistic regression performance metrics (AUC) on the test set data with feature vectors extracted from the convolutional layers along the VGG-16 architecture: trained on ImageNet data or fine-tuned on our data, showing an increase of 0.10 points in ROC AUC (from 0.86 on ImageNet to 0.96 on ours) and additionally deducing that we have a highly specialized last convolutional block.

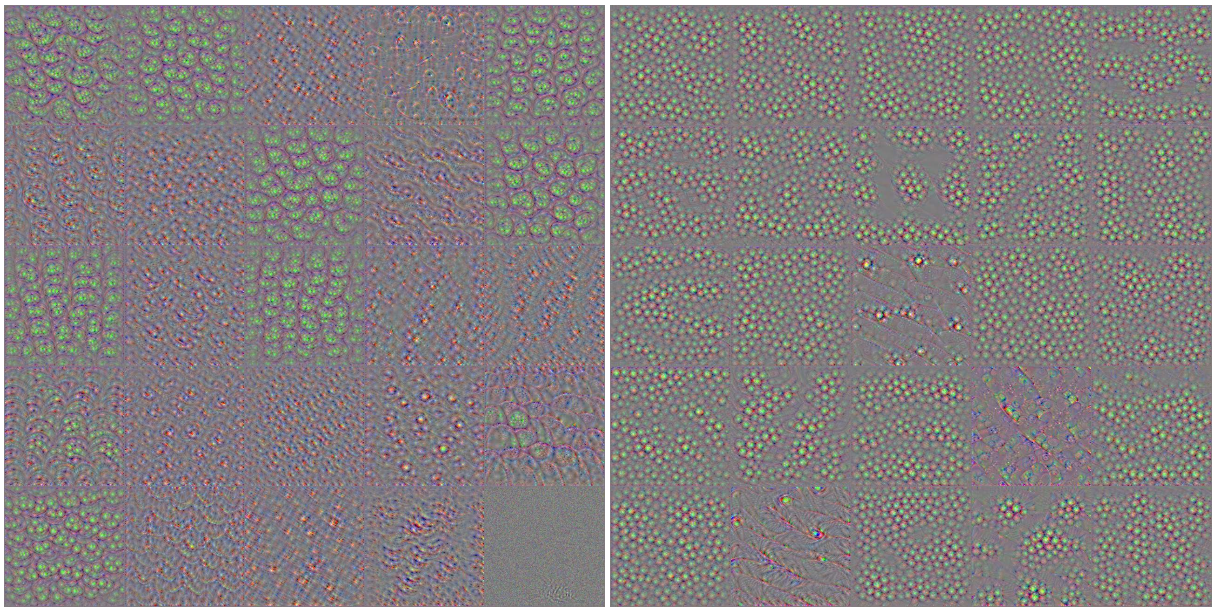
Class-wise filter bases

Given the good performance of logistic regression, we can say that the data is linearly separable in the learned feature space, therefore we can deduce that there are two quasi distinct feature bases describing cancerous and normal tissue. For the last application in this experiment, we look at the coefficients of the features in the learned decision function, or, in other words, the feature importance. In Figure V.11 a can be seen the distribution of said coefficients, it stands out that many of them evaluate around zero. This non-explicitly trained sparse model could suggest that the feature vector is too big for the given problem, but it does not represents a problem as it still generalizes well, on the contrary, it is an indication that the feature extraction model can accommodate more complexity i.e. could be extended with more data in the future. Now considering the non-zero coefficients, by taking the patterns corresponding to the coefficients with the largest absolute values we can find the most discriminating features and create two class-wise "filter bases": negative coefficient correspond to negative evidence of the cancerous diagnosis (i.e. normal class), while positive coefficients represent positive evidence (i.e. tumor). See Figure V.11 b for a visual representation of 50 of the most discriminatory visual patterns towards diagnosis: on one hand, showing small cells organized in grape-like structures resembling the main mammary tissue structures - the breast lobules - and some regular fiber patterns, typical for normal tissue; on the other hand, for the tumor class, there are depicted much bigger cells, not organized in any particular formation, together with big solitary cells intertwined between the fiber system.

We remind the reader that the patterns in Figure V.11 b are computed by sequentially maximizing via brute-force the activation of each filter in a convolutional layer and simulating the input image that would generate this maximization. The complexity and resolution of the textures are therefore dependent on the receptive field of the filter. The receptive field represents the zone of the input image that is "visible" for



a) Histogram of magnitude of logistic regression coefficients.



b1) Negative evidence filter bank showing lobule-like structures and fibers.

b2) Positive evidence filter bank showing big densely packed cells and big isolated invasive cells amid fibers.

Figure V.11: Looking under the hood of the logistic regression model and decoding the learned filterbanks: **a)** histogram of the coefficients of the logistic regression model, i.e. the distribution of weights of the data features, where each feature is the activation i.e. presence of a corresponding filter in the input image; **b)** showing the filters corresponding to the coefficient with the highest magnitudes: the top 25 negative and positive, respectively, from the total of 512 filters.

a CNN kernel is computed upon. Given the hierarchical design of CNNs, the size of the receptive fields grows proportionally to their corresponding layer depth, allowing to learn progressively more complex features. For example, here is the size of the receptive fields of the last convolutional layer in each block of VGG16: 5×5, 14×14, 40×40, 92×92 and 196×196. Accordingly, the shown features correspond to patterns at a resolution of 196×196 px.

V.3.4.2 Enlarged Nucleoli as Cancer Biomarker in DCI Imaging

The previous experiment where we discover class-specific textures belongs to our quest of finding cancer biomarkers proper to DCI images. They seem to be generic texture on cell organization, they do

not bring any new insights but come to confirm our understanding of the appearance of breast tissue in DCI. However, during our multiple experiments in search for the suitable hyperparameters and training strategies, we have noticed an interesting phenomenon, namely that a trained model with suboptimal diagnosis performance had learned some smooth information-bearing filters, which is not what is expected from an overfitted model, but rather noisy filters. This occurred when using the Adam optimizer, which is an adaptive gradient decent with momentum estimation. In other words, compared to the SGD optimizer that was finally used in the present work, Adam adapts the learning rate iteratively to the parameters performing small updates for frequently occurring features and large updates for the rarest ones. Adam is used successfully in many application from the literature and it is seen as a "silver bullet" in the DL community since it does not require as much fine-tuning on the hyperparameters as fixed rate gradient descent methods, like SGD; there are however works showing adaptive optimization methods should be taken with a grain of salt [127] as they might converge to a suboptimal local minimum.

The interesting finding observed with this Adam-trained model was first observed in in the appearance of the learned filters (see Figure V.12 c) and confirmed by the input images (example in Figure V.12 b1) as driven by the attention maps (see Figure V.12 b2). We notice round cell-like structures encompassing one or more dark dots and due to the apparent size (few microns) of those black dots, we have strong reason to conclude that they correspond in fact to *nucleoli*.

The nucleolus, first documented in the 1830s, a composing organelle of the nucleus, it is a dense structure, packed with DNA (see Figure V.12 c). We know DNA absorbs light, i.e. appearing dark, property that comes to enforce out hypothesis that those dots are indeed nucleoli. The typical nucleolus boasts 2 to 5 nucleoli, ranging in size from 0.5 to 5 μm in diameter.

However, tumor cells have larger and more numerous nucleoli, which indicates intense protein synthesis by the cell. Moreover, high nucleolar frequency and multiple nucleoli correlate with high mitotic rate in breast cancer [157], and mitotic rate is one of the main markers for grading the severity of multiple types of cancers. As a matter of fact, a strong correlation between nucleolar morphology and cancer was recognized by pathologists over 100 years ago, when it was first observed that large and abnormal nucleoli were common in cancer cells [158, 159] and allegedly, the most well documented cytological changes in cancers occur in the nucleolus [160]. Therefore, prominent nucleoli are now a widely used diagnostic marker of human cancers [161] and there are even arising automatic methods for the detection of prominent nucleoli [162].

Based on these experimental findings and literature research, we are strongly advocating for the importance of further biological investigation on the appearance of nucleoli in DCI imaging, as they bear important therapeutic potential. An interesting aspect is the differences of nucleoli analysis in classical H&E *vs.* DCI imaging; in H&E they would need to be counted and measured to be deemed as abnormal

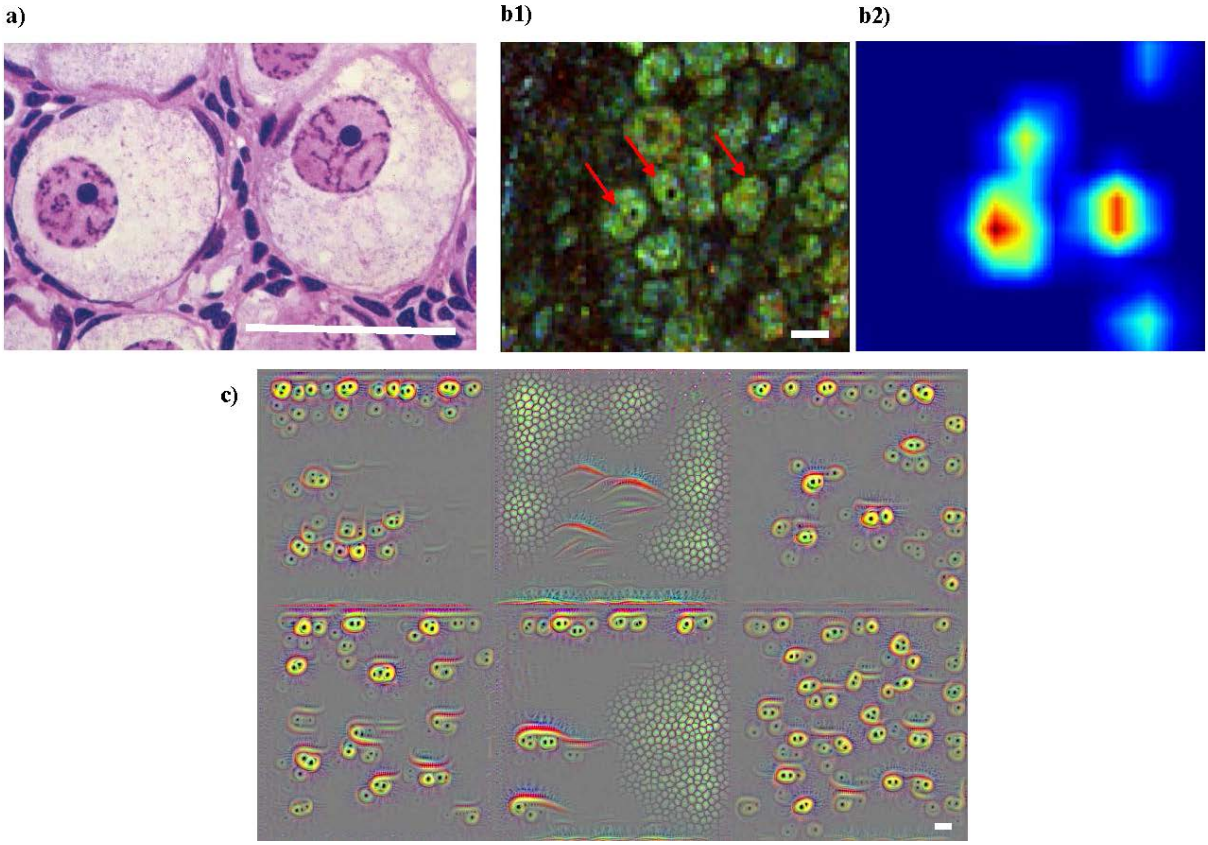


Figure V.12: a) example of cell appearance in H&E showing the cell body with cytoplasm and the nucleus containing nucleoli; **b1)** example of cells with visible nucleoli in DCI imaging and **b1)** the corresponding tumor-positive attention map highlighting the visible nucleoli; **c)** example of learned filters clearly showing cells with visible nucleoli; $10\,\mu\mathrm{m}$ scalebars.

- which is tedious task but, on the other hand, in DCI only their proliferating quality is apparent, as normal nuclei seem to fall below the diffraction limit (at least for the give breast dataset). Consequently, we suggest considering hyper-proliferating nucleoli as a viable cancer biomarker in DCI imaging.

V.3.4.3 Localizing Tumors and Normal Structures with Attention Maps

In the previous experiments we primarily looked at the learned patterns that the CONV layers respond to, but as it was presented in Section IV.3.2, this is not the only tool to look under the hood of CNNs. Another powerful method is visualizing the so called attention maps, which are heatmaps highlighting the areas in a specific input image by their levels of contribution to an output (i.e. class).

In practice, we used an extended variant of the *Gradient-weighted Class Activation Mapping* (Grad-CAM) [142] method. Originally this method displays the activation maps of all the filters from a given convolutional layer - usually the last one, as it holds the most task-specific information - modulated by the back-propagated *positive* gradient flowing from a certain output node (i.e class) to that layer. However, as we are dealing with a binary case and only one output neuron, we are looking at the

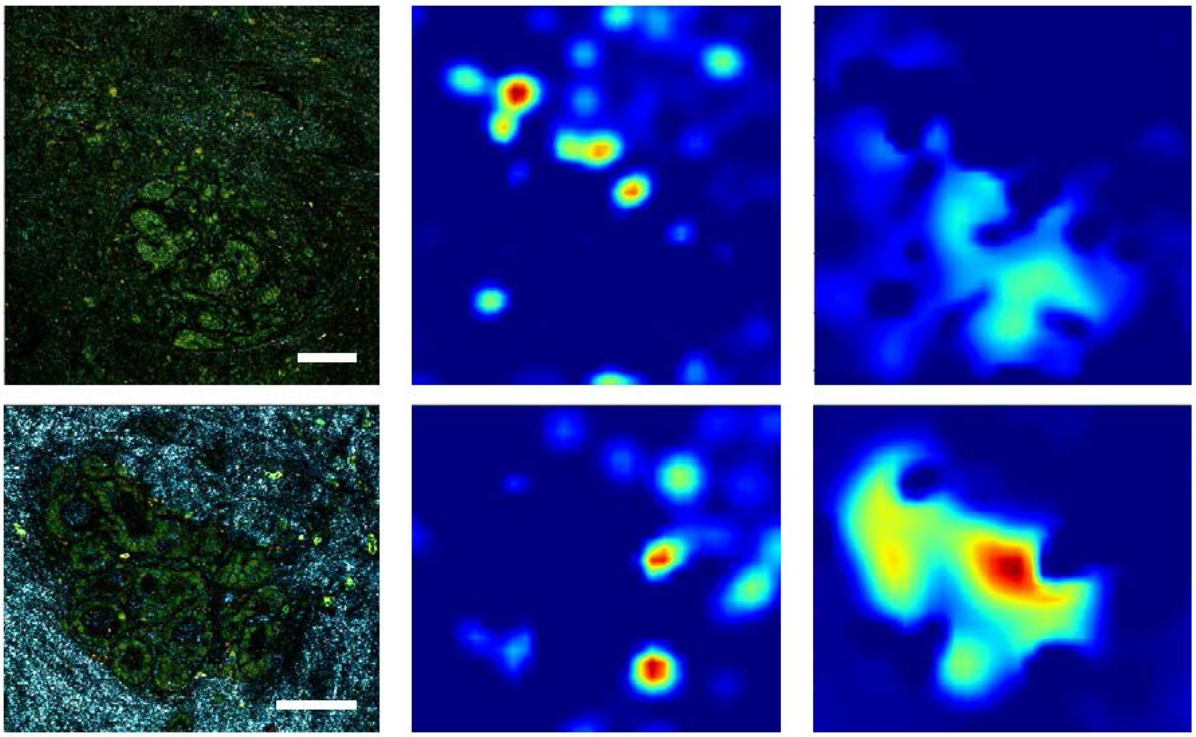


Figure V.13: Visual examples of attention maps corresponding to the last CONV layer block5_conv3 on image crops: input image crop (left), corresponding **tumor-positive** attention map (middle) and **tumor-negative** attention map (right); top example shows a normal lobule surrounded by infiltrating cancer cells (tumor sample with 79% tumor predicted probability); bottom example shows a normal lobule coming from a normal sample with 24% tumor predicted probability); $100\,\mu\mathrm{m}$ scalebar.

negative gradients separately as an indication on the absence of the tumoral class, i.e. the presence of the normal class (see Figure IV.5).

This results in a coarse localization of the evidence of the class presence and absence, respectively, in a given input ROI image. without the need for annotation at smaller scale. We say the attention map is only a coarse localization due to the fact that the corresponding layer activation maps based on which it is calculated, have a highly reduced spatial resolution as compared to the input resolution; here the image resolution is 1440×1440 px while the activation maps of the last CONV block undergoes a 16-fold downsampling, measuring 90×90 px.

Therefore, we compute tumor class positive and negative attention maps and analyse them observing that they predominantly show either cancerous cells invading the stroma, or healthy lobules and ducts, as confirmed by the pathologist's assessment (see Figure V.13). Visualizing these attention maps can serve multiple purposes, including verifying that the model is not biased or drawing attention to specific parts of the image that can assist the surgeon in rapid analysis of the sample, this is particularly useful for larger wide-field images of entire biopsies.

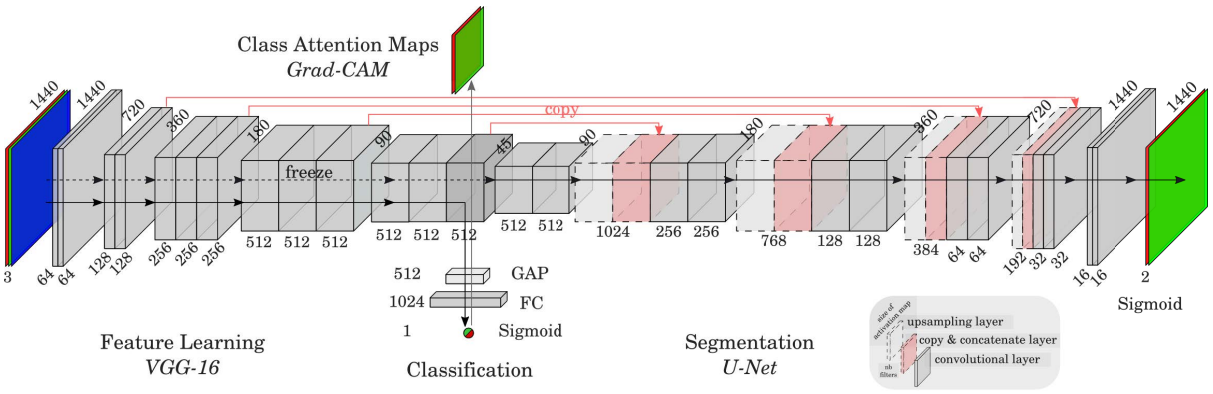


Figure V.14: Streamlined plug & play architecture containing the 3-step workflow: (i) feature extraction and classification with VGG-16, (ii) Attention Map computation (offline) with GradCAM, (iii) Segmentation by building a U-Net architecture having the pre-trained VGG-16 as backbone.

V.3.5 Streamlined Localization Architecture for Easy Deployment

In this section we shall leverage the attention maps to generate annotation masks for tumor and normal class, respectively, then train an image segmentation architecture on this auto-generated ground truth.

The obtained attention maps were confronted against and the prior interpretation of the pathologist on several interesting FOVS, especially those that contained both normal and cancerous structures. As the attention maps showed strong agreement with the pathologist's reading, we have thus decided that it is pertinent to leverage this coarse localization to guide a segmentation model. To pursue this we transformed the attention maps into segmentation mask which would serve as ground truth for training a U-Net [136] built by merging the network already trained on the classification task and adding a decoder branch. See Figure V.14 for more details on the architecture. The pre-trained branch is "frozen" meaning that we are building upon the classification features and there are only the parameters of the decoder left to train ($\sim 9M$ parameters).

Noting that there is no high-confidence ground truth available for segmentation, the processing steps of converting the attention maps into segmentation mask, as well as the choice of the loss to optimize were guided by two aspects: (i) the substantially lower resolution of the heatmaps as opposed to image resolution and (ii) GradCAM's documented weak point that it usually captures only the most discriminating part of the classified object or only one instance of the object.

The attention maps were upscaled to the input size using bilinear interpolation, followed by Otsu thresholding, morphological dilation with a circular structuring element of radius $r = 15$, and Gaussian filtering with $\sigma = 15$ to account for the uncertainty on the boundaries. We also zeroed out the tumor-positive attention maps for normal samples, knowing there are no tumor cells present in normal FOVs, but there could be several healthy structures present in cancerous FOVs. Accordingly, for each input image two

segmentation masks were generated, allegedly corresponding to normal breast tissue structures and tumoral cell clusters, respectively.

We train the decoder by minimizing **Tversky loss** [163] using Adam optimizer with a rate of $lr = 1e{-}4$. The Tversky loss is defined as:

$$L_{\text{Tversky}} = 1 - \frac{TP}{TP + \alpha FN + \beta FP} \tag{V.2}$$

The loss, which is a generalization of the more popular Dice loss that introduces unbalanced penalization of classes vs background. The penalization parameter $\alpha = 0.6$ (chosen from literature), meaning that false positives (FP) are penalized higher than false negatives (FN) i.e. modeling the fact that we have high confidence in the "pre-segmentation" already obtained through GradCAM, but we encourage an extended segmentation of the entire areas of interest; by extension $\beta = 1 - \alpha = 0.4$ relaxes the penalization on adding "new" pixels to the segmentation. Less parameters to train allow for a sightly bigger batch size of 5. Knowing that U-Net is generally fast to converge and our generated ground truth is not of 100% confidence, we stop training when the loss is stabilized, after 15 epochs.

Visually, the segmentation obtained is slightly finer and indeed including more cells, but we can not give a quantitative result at this point as the ground truth is not yet validated by a clinician. However, the main advantage of this approach resides in the quality of the model to combine the classification and segmentation tasks and to provide both diagnosis and localization. This end-to-end architecture benefits from an easy deployment capacity and it could ultimately serve for online training on the device side in the sense that clinicians could help improve the model with their immediate feedback i.e. producing annotations by correcting the model predictions.

V.4 Conclusion

To sum up, in this chapter we presented different applications based on data from different modalities - FFOCT, DCI signal and DCI processed images - and tissue types - skin and breast - but sharing the quality of being densely annotated - at pixel level or millimeter level. Therefore, we leveraged various supervised learning methodologies, all having as common denominator the same end goal: cancer detection, together with a strong focus on interpretability and validation strategies.

We have shown that building a custom CNN architecture can be beneficial in the case of well posed problems (in the sense of having a representative and adequately large training set to represent the problem) as adapting the filter kernel size (and receptive fields, respectively) or model depth and complexity can lead to increased performance. While, on the other hand, when dealing with smaller and more complex datasets using a pre-trained model is crucial.

In our exploratory work we did not take into account solely the performance metrics, but aimed to also gain insights from qualitative strategies to either validate the results or acquire new information. In this regard, we called to (i) feature importance derived from random forest based classifiers to highlight some salient oscillatory frequencies from the dynamic signal; (ii) simulated textures representing the learned features of CNN models to build class-specific feature bases and find cancer bio-markers for DCI modality; (iii) Grad-CAM algorithm to build class-specific attention maps and localize tissue structures in the input images.

To conclude, this section is a anthology of exploratory methods that can be employed to analyze FFOCT / DCI imaging under its different aspects. The insights gained from these applications will be further extended in the next section, where we tackle a scenario far more challenging, but which illustrates a real-world situation in terms of data complexity and availability.

Chapter VI

Benign *vs.* Malignant Classification from Global Diagnosis

In the previous chapter we presented a set of exploratory approaches for distinguishing tumor from normal tissue, in this regard we applied supervised learning with dense labels, i.e. all data points had an associated ground truth, however, this is an ideal setting that is difficult to reproduce in all real-world problems. As we have succinctly presented in Section III.1.2, specific medical annotation are expensive and labor intensive to obtain, nonetheless, in the case of clinical tissue assessment a histological report is always produced; it contains an overall description of the tissue architecture and cell aspect contributing to the diagnosis, hence training models directly on the information extracted from these reports would require no extra intervention from a medical expert to annotate the images.

Nonetheless, in addition to getting the global ground truth per subject, one has to remember the necessity of subsampling the images (see Section III.1.3), therefore the question that arises is how to extend the global label to the sampled image components (usually patches or ROIs). One way to tackle the problem is to assign the global label to all its sub parts and train using classical fully supervised algorithms, however the probability for the model to converge is inversely proportional with the amount of noise in the labels [164]. We know that the label is not omnipresent, but on the contrary, it can be limited to a few cancer cells (on the order of hundreds of pixels) in a whole core-needle biopsy (on the order of a billion pixels). Therefore, a more adapted solution is using the multiple instance learning (MIL) framework, detailed in the next section, which integrates the assumption that the global label might not be apparent in all sub parts of the entity.

In this chapter, it is described how we adapt the MIL framework for our dataset of breast biopsies from Section III.2.2.2 to distinguish malignant from benign specimens. The current state of the work represents a promising proof of concept as the dataset used is incomplete with respect to what is expected

from its source clinical study. However, it shows great results as to the final clinical applications, given the satisfactory performance metrics obtained and the insights inferred from post-hoc analysis of the results. We also test the employed algorithm on a classical histology benchmark dataset (i.e. breast metastasis detection in lymph node H&E stained images - CAMELYION16 [165]) and compare the results with other MIL approaches. Moreover, we take advantage of this densely annotated dataset to validate the instance-level prediction capability of our MIL approach as well.

Contents

VI.1 Multiple Instance Learning

VI.1.1 Motivation

The concept behind multiple instance learning (MIL) is not new, it was explored [166] in the 90's and the exact nomenclature was first used in 1997 by [167] to try and solve the problem of the chemists who could only determine if a molecule is qualified to make some drug or not, but they couldn't say exactly which of its states are responsible for that stability. Therefore, the paradigm came as a response to the need of formalizing incomplete knowledge or classifying heterogeneous groups. These mixed groups are generally defined in a binary manner under the assumption that negative groups only contain negative instances while positive groups can also contain negative instances and they should contain at least one positive instance, but the nature of the instances is unknown (see Figure VI.1).

This situation occurs often in the case of image classification, as there are few datasets with only one object type per image and they are carefully curated academic datasets, like MNIST [168], CIFAR-100 [169] or Caltech-256 [170]. However, in real world settings this case seldom arises, for natural scenes but even less so for medical imaging [171] where the pathological zone represents just a fraction of the (healthy) anatomical context. Taking the case of histology, the tumor to tissue proportion is heavily skewed, for example in the CAMELYON [65] dataset there is a median of 2% tumoral patches per slide (ranging from 70% down to even 0.01%) [172]. Therefore, the MIL paradigm fits the problem of gigapixel histopathology images, where the instances are represented by multiple sub-images (i.e. patches) sampled from a bigger image like a WSI which would represent a bag.

VI.1.2 Method

There are two major flavors of algorithms for MIL [173]: instance-based and embedding-based algorithms. The term instance-based denotes that the algorithm attempts to find a set of representative instances based on an MIL assumption and classify future bags from these representatives. By contrast, embedding-based algorithms make no assumptions about the relationship between instance labels and bag labels, and instead try to extract instance-independent information about the bags in order to learn the concept, in other words, each bag is defined with a single relevant embedding and thus generic single instance classification algorithms can be used.

The purpose of any MIL application is to predict the label probability $\hat{Y} \in \mathcal{P}$ (for the binary case $\mathcal{P} = [0, 1]$) for a bag of instances $X = \{x_i | x_i \in \mathcal{I}, i = \overline{1, N}\}$, where $\mathcal{I}$ is the instance space and N is the number of instances inside the bag, and the ground truth bag label Y is known, but not the instance labels $y_i \in \mathcal{L}, i = \overline{1, N}$ (for the binary case $\mathcal{L} = \{0, 1\}$). Regardless of the MIL strategy or the paradigm used to enable it (either traditional rule-based or deep learning based), implementation

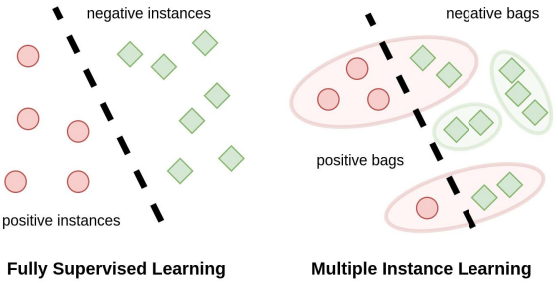


Figure VI.1: Fully Supervised Learning *vs* Multiple Instance Learning concepts illustrated for the binary classification problem.
(adapted from [174])

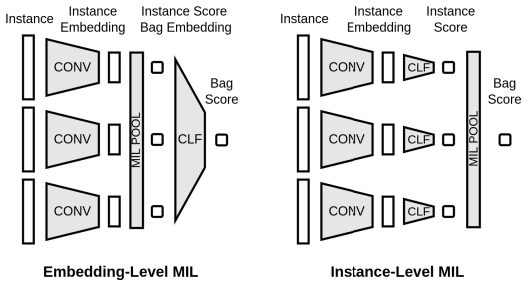


Figure VI.2: Embedding *vs* Instance-Level MIL in CNN models defined by the position of the MIL Pooling layer relative to the feature embedding convolutional layers (CONV) and the classifier layers (CLF).

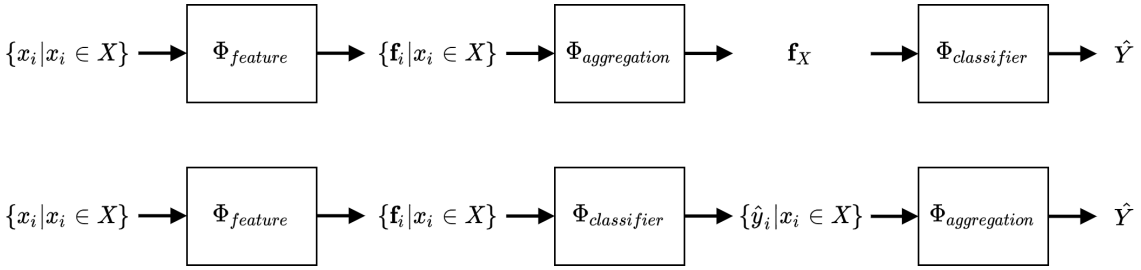


Figure VI.3: Block diagrams of **embedding-level** (top) *vs* **instance-level** (bottom) MIL frameworks with feature extractor module $\Phi_{feature}$, aggregation (or pooling) module $\Phi_{aggregation}$ and classifier $\Phi_{classifier}$, mapping input bag X to bag prediction $\hat{Y}$, passing by features $\mathbf{f}_i$ of instances x_i and either bag feature vector $\mathbf{f}_X$ (for embedding-level MIL) or instance predictions $\hat{y}_i$ (for instance-level MIL).

requires the same computational entities: feature extractor $\Phi_{feature}$, aggregator $\Phi_{aggregation}$ and classifier $\Phi_{classifier}$. Usually, in both cases, the instances of a bag X are mapped from the input domain $\mathcal{I}$ to a latent space $\mathcal{F}$ by $\Phi_{feature} : \mathcal{I} \rightarrow \mathcal{F}$ such that for each instance there is a corresponding feature vector $\mathbf{f}_i \in \mathcal{F}, \forall x_i \in X$. The main difference in implementing the two flavors resides in the relationship between the instance features and the bag prediction or in the succession of the aggregator and classifier blocks. Namely, for the embedding-based approach the instance features are combined into a global bag feature vector $\mathbf{f}_X$ by $\Phi_{aggregation} : \mathcal{F} \rightarrow \mathcal{F}$ based on which the classifier finds the bag score $\hat{Y}$ by $\Phi_{classifier} : \mathcal{F} \rightarrow \mathcal{P}$, while for the instance-based method, for each instance x_i its score $\hat{y}_i$ is predicted and the bag score is obtained by aggregating the instance scores with $\Phi_{aggregation} : \mathcal{P} \rightarrow \mathcal{P}$. See Figure VI.3 for the generic block diagrams of the two MIL approaches and Figure VI.2 for schematic of the two flavors implemented under the CNN paradigm, where $\Phi_{feature}$ is implemented by convolutional layers, $\Phi_{aggregation}$ by a pooling layer and $\Phi_{classifier}$ by more artificial neural layers.

VI.1.3 Related Work

While there exist applications where a single input image is considered the bag and the pixels are considered to be the instances [175, 176] they are used for small-scale natural images. We are interested

in the case where the instances are represented by multiple images sampled from a bigger image which would represent a bag. Even though these two perspectives revolve around the same key concept, they differ in implementation. In CHOWDER [177] they adapted the WELDON [175] method, consisting of a *min-max* aggregator, for WSI classification from image tiles. Since they use a non-specialized feature representation they obtain high variance of the model performance which is overcome with ensemble learning. This is only disclosed in a later work [178] where they employ a self supervised contrastive learning pre-training method based on [179] in order to learn data specific features, instead of using the features extracted from ImageNet. In [180] they also use contrastive learning as a helper task for pre-training the MIL feature extractor. Nonetheless, to our knowledge, there are no works that include the feature learning step in the MIL framework and thus learning a task specific embedding. In [181], while obtaining a good accuracy in breast cancer detection, they state the importance of also learning adapted features.

Attention based MIL framework using only slide labels - CLAM [182] and variant introducing a percentage of the instance labels [183], while reaching high bag-level classification performance, they also provide instance-level attention scores. However, this notion of importance or attention is quite fuzzy and should be taken with a grain of salt as it is far from being equivalent to the actual patch prediction.

VI.1.4 Objectives

On our side, we are interested in the highest level of transparency since we are looking to build a tool that brings interpretability of DCI images and builds confidence in the imaging technique for the surgeons, radiologists or any clinical personnel potentially involved with rapid extemporaneous tissue analysis. As in all previous applications presented, we are still concerned about learning adapted features. With this in mind, we define the following **requirements** for our MIL approach for classifying malignant from benign breast biopsies in DCI images:

1. obtain adequate biopsy-level predictions;
2. infer adequate tile-level predictions;
3. learn an adapted feature base;
4. build a mutable model that can be easily extended.

Henceforth, to respect the requirements, in our application we are using the instance-based flavor of the MIL framework under a set of assumptions:

Assumption 1 *All biopsies (i.e. samples) coming from the same nodule (i.e. inclusion) share the same diagnostic.*

This assumption is needed at this incipient phase of the study in order to maximize the number of labeled data entities (bags). Despite being a weak assumption, its effects could be minimized by a richer dataset which would make up for the slightly noisy labels or, on the other hand, we could expect a sample-level annotation on DCI images, hence removing the ambiguity altogether.

The standard presence-based assumption [184] on which MIL stands on is translated to our nomenclature as follows:

Assumption 2 *There is at least one malignant patch (i.e. positive instance) in each malignant sample (i.e. positive bag).*

Assumption 3 *There are no malignant patches (i.e. positive instances) in the benign samples (i.e. negative bags).*

In the next sections we will show how to put these definitions into practice, namely how is the network architecture designed and then trained on the breast biopsies dataset (see Section III.2.2.2 for details on the dataset).

VI.2 Model

VI.2.1 Multi-branch Architecture

As we have seen, the majority of MIL methods applied to WSIs are embedding-based algorithms that focus on implementing different strategies for instance interaction and training a suitable classifier, with little to no focus on incorporating the actual feature extraction in the MIL training; this is mainly due to an unavailability of the computational resources demanded by such a framework. The feature extraction part of the model is either transferred from a state-of-the-art architecture trained on another dataset (e.g. ResNet trained on ImageNet [185]) or trained on the dataset at hand on different helper tasks [186]. However, to our knowledge, there are no applications that incorporate task-level feature learning in training under the MIL assumption.

Nonetheless, we can take advantage of the slightly reduced resolution of DCI imaging (compared to WSIs) and some pre-selection of pertinent information in them, as well as a previous model V.3 trained on a similar dataset III.2.2.1 and design an instance-level MIL model with task-specific features and instance-level predictions.

VI.2.2 Information Fusion with Global MIL Pooling

To come back to the two previously mentioned forms of MIL: the instance-based and embedding-based approaches, they translate into the multi-branch network architecture by the level at which the fusion takes place, in this regard, for the first case the bag output is obtained by aggregating the instance-level outputs and for the latter the fusion happens at the instance feature vectors (See Figure VI.2). In other words, the same feature extraction network is applied on all instances in a bag, then the instance information could be somehow fusioned to a have a unified vector to represent a bag on top of which the bag-level classifier is added. However, by virtue of the usual black-box nature of the relationship between the intermediate fusion mechanism and the classifier, this method would not reveal interpretable insights about the instances themselves. Conversely, the whole network (feature extractor + classifier) could be applied on all instances, then a specific MIL pooling on the instance-level outputs is defined as to obtain the bag-level output; with this approach the predictions at instance-level can be directly inferred. As one of our top concerns when building the models is interpretability, we opt for the latter approach.

As for the MIL pooling, there are multiple strategies that could be adopted, as long as they follow two important requirements dictated by set theory (as each bag X respects the set properties). Firstly, the model output needs to be permutation-invariant meaning that its output needs to be independent from the ordering of the instances in a bag. Secondly, the model needs to accept any input size (as bags X could be of variable size, with a different number of instances) and be independent from the set abundance (in the sense that the number of instances should not affect the result either). Therefore $\Phi_{aggregation}$ needs to be permutation-invariant and independent from input size itself. The possible function to employ are maximum, average, distribution-based (i.e. quantile [187]) or attention-based. From a computational point of view, they can be equally applied under the two MIL paradigms, but conceptually, it would make less sense averaging the instance scores as compared to averaging the instance features, for example. Therefore, in our case, for aggregating over the instance scores in order to obtain the bag score $\hat{Y} = \Phi_{aggregation}(\hat{y_i})$, we turn to the maximum function, as it also respects perfectly the fundamental Assumptions 2 & 3.

We build a multi-branch architecture (Fig. VI.4) whose constituting branch is transferred entirely: convolutional layers (CONV) + fully-connected classifier (FC), from the CNN model trained with full supervision on breast excision dataset (see III.2.2.1) and duplicated k times, where k is the number of instances per bag. All the weights are shared across branches (this is also known in the literature as *siamese network* [135]), on top of that it is added a *Max* pooling layer which aggregates the instance-level predictions under the standard MIL assumption to give the bag-level prediction. For each sample $X = \{x_i | i = \overline{1, N}\}$ containing N tiles x_i, $i = \overline{1, N}$ whose global ground truth Y is known (and

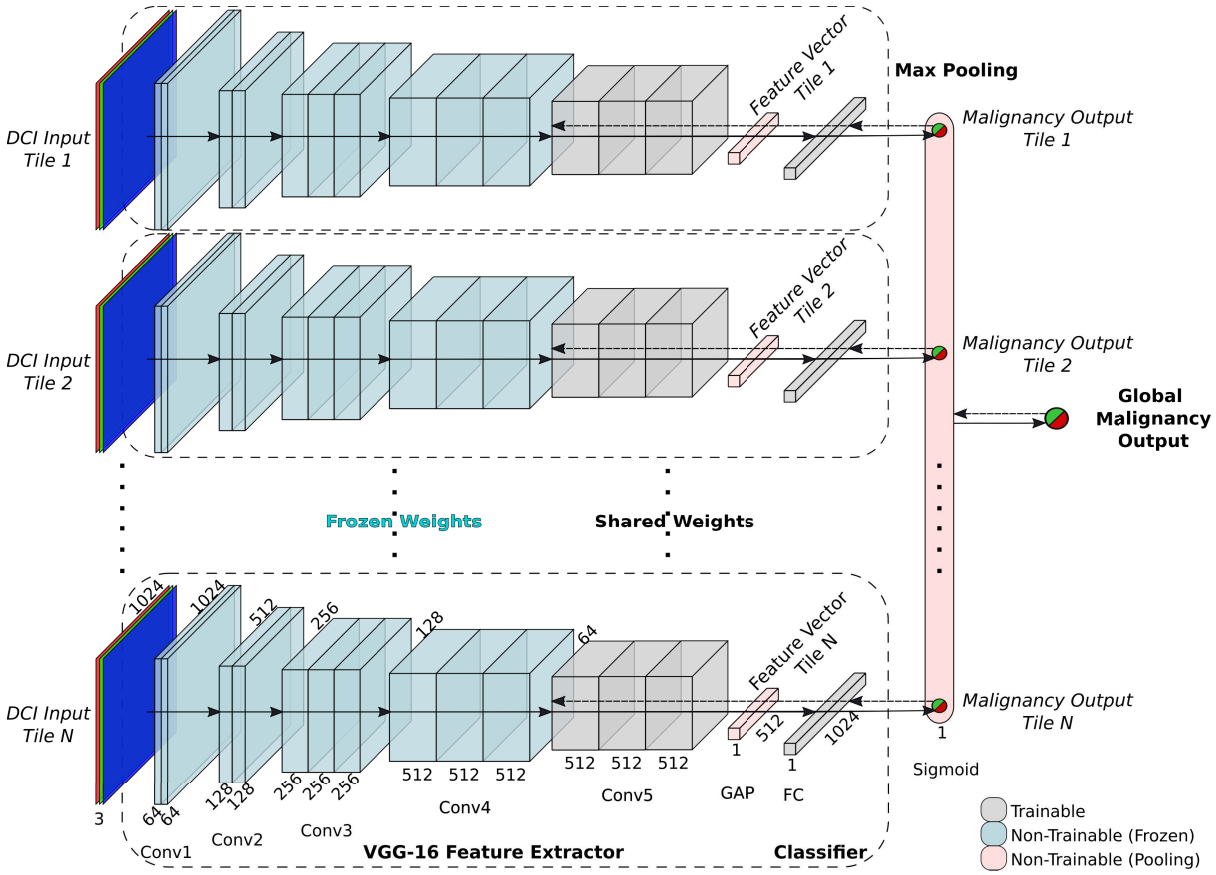


Figure VI.4: Architecture for MIL

represents the sample diagnosis), we build the corresponding bag by selecting a subset $X_b \subseteq X$. The main branch encodes a function $\Phi = \Phi_{feature} \circ \Phi_{classifier}$, $\Phi : X \mapsto [0, 1]$, mapping each tile x_i to its predicted probability $\hat{y}_i = \Phi(x_i)$, which are then aggregated by the MIL pooling layer to obtain the sample prediction $\hat{Y} = \max_{i \in \{1,k\}}(\hat{y}_i)$. Let us emphasize that the tile labels y_i are unknown, however we can still obtain tile predictions $\hat{y}_i$.

The forward pass involves all the instances of the bags in one batch, which results in a probability score $\hat{y}_i$ for each instance x_i and then, by virtue of the MIL pooling, the top ranked instance of each bag is used for training through backpropagating the gradient of the loss function computed on the bag ground truth and the highest scoring instance probability. Accordingly, only one instance contributes at each step in the training, but as the weights are shared, the highest scoring instance is bound to change during training, aiding generalization.

VI.2.3 Weight Transfer, Sharing & Freezing

In our case, we can take advantage of the already trained model on a very similar dataset (see Section III.2.2.1) of the same modality, same organ and same pathology, but a different clinical setting, i.e. surgery vs biopsy, which also leads to a difference in the negative class, i.e. healthy vs benign, while the

positive class has the same definition, i.e. malignant. For simplicity, we shall refer to the pre-training dataset in Section III.2.2.1 as D1 and the fine-tuning dataset in Section III.2.2.2 as D2.

The initial state Φ_0 of the main branch model Φ is based on a VGG-16 [131] backbone which takes inputs of any size by virtue of its *global average pooling* (GAP) bottleneck between CONV and FC. This model was trained in a FS manner on D1 by minimizing the weighted binary cross entropy loss with *stochastic gradient descent* (SGD) optimizer and a learning rate $lr = 10^{-4}$ (See Section V.3).

Since Φ_0 was trained on the same modality and tissue type we shall leverage the low-level features and fine-tune the last layers, which intuitively learn more specialized features. However, to confirm this hypothesis and find the exact layers to train, we compare the logistic regression classification results of VGG-16 trained on ImageNet and D1. In this regard, we extract the features (with GAP) from each layer of the network and compare their performance evolution between the two models, until we find a dramatic performance improvement of the D1-trained model at the first layer of the last convolutional block (refer to previous experiment in Section V.3.4.1 and Figure V.10 for more details).

Consequently, the first 4 convolutional blocks are frozen during training and only the 3 convolutional layers comprising the last convolutional block are further trained on the current task (see Fig. VI.4), as we deduced that those are highly specialized feature extractors for the final task, while the others encode generic textures proper to the modality and organ.

VI.3 Training

When training neural networks, the limiting bottleneck is the available memory of the GPU, which is in continuous expansion thanks to the technical advancements, for our available setup, the VRAM can go up to $48\,\mathrm{GB}$). Accordingly, when designing the network architecture and choosing the training hyperparameters (e.g. the batch size), one has to take into consideration the available memory at hand. For example, with the presented VGG-16 architecture and input size, while its 15M weights take up only $60\,\mathrm{MB}$ of space, the activations occupy around $1.2\,\mathrm{GB}$ of memory for one branch only. Intuitively, this quantity becomes directly proportional with the number of network branches i.e. instances in a bag or patches in an image and with the batch size. For example, with a bag size of 5 and a mini-batch of 2, the necessary memory increases 10X totaling to $12\,\mathrm{GB}$ of VRAM; not to mention that this quantity corresponds to the forward pass alone, which according to [188] takes less than a quarter of the memory necessary for the whole training process, as computing the corresponding gradient maps for the backward loss propagation are more resource consuming. Reducing the memory required for storing the gradient maps at back-propagation is the main reason why MIL approaches use pre-trained feature extraction networks; since end-to-end training is indeed unrealistic for this paradigm, we opt for a feature

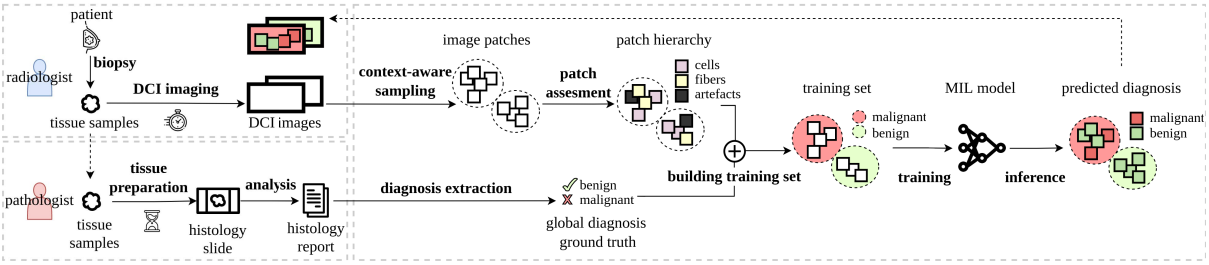


Figure VI.5: Workflow for MIL training on breast biopsies: from image acquisition to inference.

extraction model trained on a similar dataset and still choose to re-train the last more specialized layers in the MIL setting.

VI.3.1 Tackling Computational Constraints with Content-aware Bag Generation

Just as in the case of gigapixel histopathology images, end-to-end training is hindered by the computational resources, since the entire image together with the computational graph cannot fit in the GPU RAM. Therefore, further partitioning and filtering of the image is needed; the primary goal is to remove the pixels bearing no information, but encumbering the resources nonetheless, like the background pixels, secondly, image artifacts could be removed as they introduce noise that would burden model convergence.

Even if at training time the maximum bag size is limited by the memory of the GPUs at hand which is already generous in this particular case it is not a blocking constraint since we have a way to select the most representative patches for each sample, and therefore minimize the computational burden without losing useful information. Therefore, when choosing the maximum allowed bag size one needs to consider some relevant *a priori* data statistics, which in this case is is the median number of patches per sample containing cells (i.e. 6). Therefore, the constraints on the value for k_{max}^{train} are lower bounded by the average number of patches containing cells in a sample and upper bounded by the memory resources (depending also on mini-batch size): $6 \leq k_{max}^{train} \leq 8$, we choose $\mathbf{k_{max}^{train}} = 8$. For training, for each sample X_j, for its corresponding bag X_j^b the bag size is $k_j^{train} = minimum(k_{max}^{train}, N_j)$, and the batches are constructed taking into account equally populated bags. However, for testing, there is no limit on the bag size $\mathbf{k_{max}^{test}} = \infty$, thus all patches of a sample are considered in the diagnosis, therefore $k_j^{test} = N_j$.

When sampling the images we tried to keep to a minimum their fragmentation, to respect the constraint imposed by the maximum bag size parameter. In this regard, we choose the patch size in concordance with the width of the biopsying needle i.e. 1024×1024 px and we try to capture the entire surface of the imaged tissue with minimally overlapping patches. What is more, we opted for texture-aware patchification to avoid slicing homogeneous morphological features, for that we used a method based

on the SLIC superpixel algorithm *SoSLeek*, detailed in Section III.1.3, in this manner we obtain $\simeq 2000$ tiles.

VI.3.2 Tackling Difficult Samples with Focal Loss

For dealing with the imbalanced data, we propose to utilize the focal loss (FL) [189] instead of the common used binary cross-entropy (BCE). Focal loss is proposed in the object detection community to solve the problem of extreme foreground-background class imbalance, however it performs very well on other problems like semantic segmentation or image classification, to the point where it has been shown that it also helps train more reliable and confident models [190]. Focal loss is based on a simple yet powerful idea, during model training, it reduces the contribution of the already well-classified candidates but to focus on the hard, misclassified ones.

It differs from the classic binary cross-entropy by virtue of the modulating factor γ called "focusing" parameter, which helps scale the loss for misclassified or "hard" training examples. However, given the general definition of FL, with $\gamma = 0$ FL is equivalent with BCE. See Figure VI.6 for an example of the influence of different values of the γ parameter.

After multiple experiments of hyper-parameter tuning, BCE was proven to not generalize well, even in the weighted case. The focal loss brought a 8% gain in test accuracy as compared to the common cross-entropy loss, this is due to its quality of "hard mining". Therefore, to quantify the error between the predicted sample diagnosis $\hat{Y}$ and the true bag label Y, we use the class-weighted focal loss:

$$L(Y, \hat{Y}) = \begin{cases} -w_p \alpha \left(1 - \hat{Y}\right)^{\gamma} \log(\hat{Y}) & \text{if } Y = 1 \\ -w_n (1 - \alpha) \hat{Y}^{\gamma} \log(1 - \hat{Y}) & \text{if } Y = 0 \end{cases} \tag{VI.1}$$

where γ controls the influence of higher-confidence correct predictions (or "easy" examples), α weights errors for the positive class (therefore if $\alpha < 0.5$ false positives are penalized more than false negatives and vice-versa, while for $\alpha = 0.5$ the misclassified samples are considered regardless of their true class). We also add fixed weights (w_n, w_p) computed according to the nodule diagnosis ratio, i.e. $w_p = 1.29$, $w_n = 0.82$. We used the suggested values for the focal loss parameters, namely $\gamma = 2$ and $\alpha = 0.25$ (or $\alpha_p = 0.25$ and $\alpha_n = 0.75$). Therefore we choose to penalize more the false positives than the false negatives, this is due to the slight uncertainty of the positive diagnosis (recall Assumption 1), however, in defiance of this, in medical applications it is desirable to minimize the presence of type II errors (i.e. false negatives) or errors of omissions, as they are more dangerous than type I error since they imply dismissing a sick patient, instead of performing more tests which could eventually lead to the true diagnosis. The loss is minimized via SGD with momentum and $lr = 10^{-4}$, on a mini-batch size of 3 bags, thanks to an Nvidia A40 GPU with 48 GB VRAM.

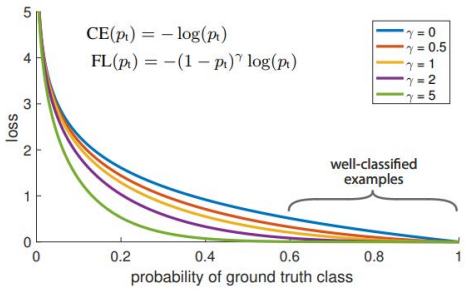


Figure VI.6: Focal Loss vs. Cross Entropy ($\gamma = 0$) and influence of parameter γ for $y = 1$ (Figure from [189]).

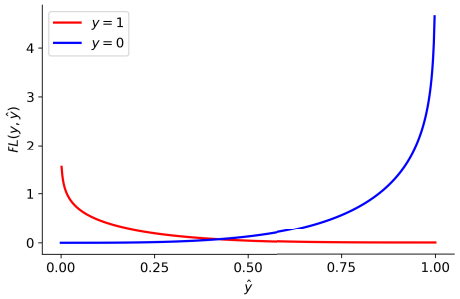


Figure VI.7: Binary Focal Loss for $\alpha = 0.25$ and $\gamma = 2$ when $y = 1$ or $y = 0$.

VI.4 Results

When defining the clinical study, the main criteria of the clinicians in favor of adopting the LightCT™ Scanner in the current practice for rapid diagnosis was reaching a Positive Predictive Value (PPV) > 70%, below this performance it was deemed useless to inform the patient about their diagnosis minutes after the biopsy procedure. Other expected metrics are Sensitivity > 80%, Specificity > 90% and Negative Predictive Value (NPV) > 80%.

In this section we present and analyze the results obtained with the proposed approach for automatic diagnosis of breast cancer in DCI images.

VI.4.1 Training Strategies Comparison

These results are strongly supported by the previous work detailed in the previous chapter in Section V.3, namely the CNN model (referred to as Φ_0 in this chapter) trained in a fully-supervised manner on ROIs coming from the breast excisions dataset presented in section III.2.2.1 (referred to as **D1** in Table VI.1) which achieved a cross-validated sample-level accuracy of 94% (*Scenario #1* in Table VI.1). Hereinafter, we shall go over the logical course of action in training strategies leading to the presented method allowing to train a malignancy classifier on the dataset with global labels presented in Section III.2.2.2 (referred to as **D2** in Table VI.1).

First of all, we test the malignancy detection capacity of the model trained on surgical excisions on the biopsy dataset (*Scenario #2*) and we unsurprisingly observe high sensitivity and low specificity, or a redundant detection with a high number of false positives; this is due to the nature of the datasets themselves, the model trained on malignant vs normal tissue has no power of discrimination between malignant and benign appearance i.e. the model having *normal* as negative class finds *benign* samples as pathological.

Table VI.1: Per nodule classification results on different training scenarios, paradigms and datasets. **D1** dataset: surgical excision breast samples healthy vs. tumoral with dense local labels annotations. **D2** dataset: breast biopsy samples benign vs. malignant with weak global labels annotations. Note that test metrics are computed per sample by aggregating the scores of their constituting instances (ROIs/FOVs for D1 and patches for D2). In bold best performance obtained on D1 and D2, respectively, as test sets

Training Paradigm	**Scenario**	**Dataset**		**Test Metrics**			
		Train	**Test**	**ACC**	**SNS**	**SPE**	**AUC**
Fully Supervised Learning	#1	D1	D1 ROI	**88**	**89**	87	**0.93**
			D1 Sample	**94**	**97**	84	**1**
	#2	D1	D2 Biopsy	57	**88**	37	0.78
			D2 Nodule	50	**100**	18	0.82
	#3	D2	D2	N.A. (label noise)			
Multiple Instance Learning	#4	D2	D2	N.A. (memory constraints)			
	#5	ImageNet	D2 Biopsy	74	53	87	0.77
		D2	D2 Nodule	72	57	82	0.75
	#6	D1	D2 Biopsy	**85**	75	**90**	**0.85**
		D2	D2 Nodule	**86**	89	**84**	**0.86**
	#7	D1	D1 ROI	57	34	**100**	**0.93**
		D2	D1 Sample	72	63	**100**	**1**

One might think of training on D2 in a similar fully supervised manner (*Scenario #3*), but this can only be possible if one assumes that all instances (i.e. tiles) of a bag (i.e. image) share the global bag label, albeit it being feasible in some applications, this extension would introduce a substantial level of noise in the labels and it would impede convergence.

On these grounds, we turn to Multiple Instance Learning (MIL), but based on the aspects mentioned at the beginning of this chapter, we cannot train end-to-end on D2 alone (*Scenario #4*) due to memory limitations. Fine-tuning a network pre-trained on ImageNet (*Scenario #5*) still gives unsatisfactory results with a sensitivity little above 50% which is most certainly caused by the low-level feature extractor blocks which are not adapted as we can train only the last convolutional block of VGG-16 and the fully-connected classifier layer. When applying the same strategy, but this time the transferred weights are adapted to our modality and tissue type (*Scenario #6*) we achieve a dramatic increase of 32 points in sensitivity (89%) and manage to match the specificity achieved in the fully supervised application (84%). Finally, one might ask what happens if we go back (*Scenario #7*) and test this mode on D1 (by

unplugging the main branch). While all normal cases are correctly ruled out of malignancy (so actually improving the specificity of the initial model), the sensitivity is low.

All in all, we observe that intra-domain pre-training improves performance. Most importantly, we deduce that, as malignancy is defined in relation with the negative class (normal or benign), having the same model perform well on the two tasks would need explicit simultaneous training, which would probably also improve the performance on the individual tasks.

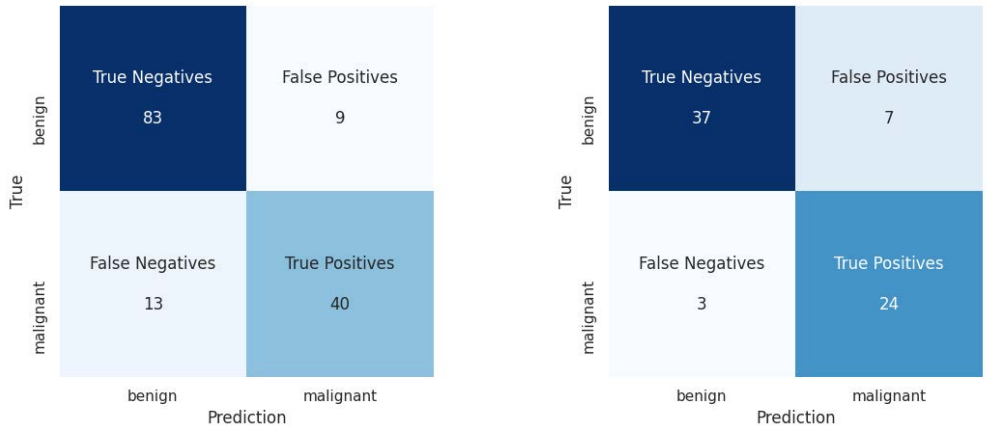
VI.4.2 Cross-Validation Results

Given the reduced dataset, it is in our interest to leverage it in its entirety, therefore having a separate holdout dataset on which to test the performance of the model is not feasible. Accordingly, we opt for analyzing the cross-validated results in order to have a clear picture about the method efficiency. For accurate test results we decide on 3-fold cross-validation which leads to a broader test set hopefully reflecting better the data distribution. The split is done with respect to the instances i.e. samples from the same nodule will be present together in either the train or test sets. Moreover, please note that for the test split we consider all the patches coming from those samples, while for the train set we comply with the method constraint imposed by the maximum bag size parameter, therefore we obtain for each fold:

- a train-set comprising: 100 samples coming from 48 inclusions, having $\sim$800 patches (given the maximal bag size of 8);
- a test-set comprising: 50 samples coming from 24 inclusions, having $\sim$700 patches (given all image patches are used for computing the metrics).

In Table VI.2 you can inspect the test binary metrics (accuracy, sensitivity, specificity, positive predictive vale, negative predictive value) obtained for a fixed probability threshold $T = 50\%$ as well as the area under the ROC curve, computed both at biopsy and nodule level, for each fold, with their average and standard deviation over fold, along with the results aggregated on the 3 test sets. For the aggregated results we also show the confusion matrices both at sample and nodule level in Figure VI.8. Note that the aggregated results were used to compare the different training strategies (previous section) in order to have a global picture for the entire dataset.

Looking at the CV results, we notice high variance between folds, which is a mere reflection of the high variance of the data between folds and is a clear indicator for the need of incorporating more data. Nonetheless, in the next section we seek to understand what is the nature of the relationship between the data and its prediction.



(a) Sample-wise **(b)** Nodule-wise

Figure VI.8: Confusion matrices for aggregated 3-fold prediction results.

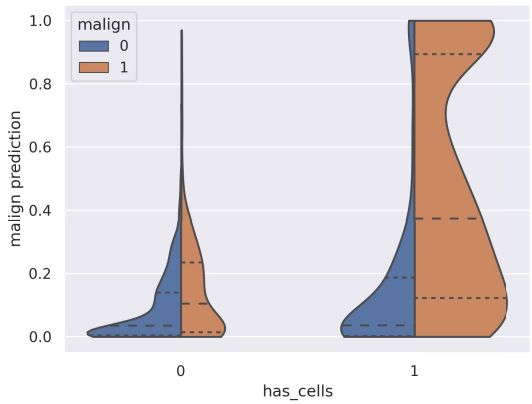


Figure VI.9: Patch malignancy prediction according to patch cellularity and sample ground truth.

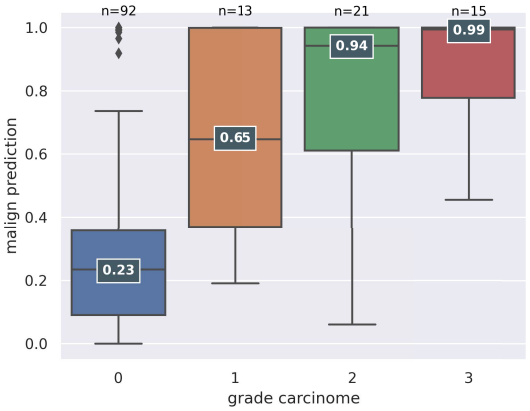


Figure VI.10: Sample malignancy predicted probability according to malignancy grade (*NB*: grade = 0 concerns benign samples).

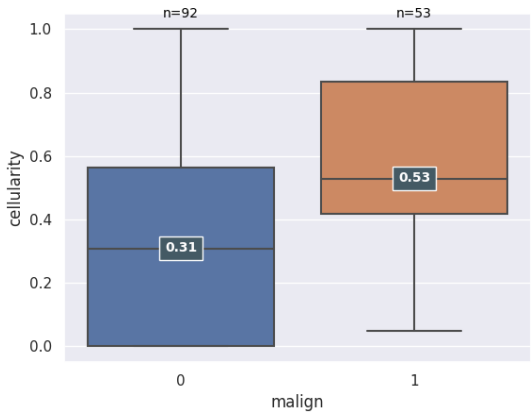


Figure VI.11: Sample cellularity according to malignancy ground truth.

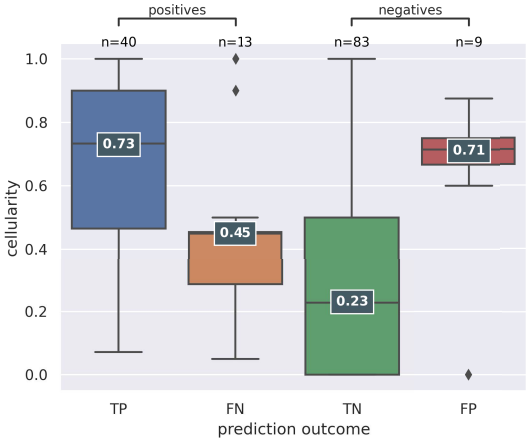


Figure VI.12: Sample cellularity according to prediction outcome.

Table VI.2: Breast biopsies dataset MIL 3-fold cross-validated test metrics computed per biopsy and per nodule as per the maximum prediction obtained on the constituting tiles: accuracy(ACC), sensitivity (SNS), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV) all computed via a 50% thresholding value on the predicted probabilities, expressed in percentage and the area under the ROC curve (AUC); reported test metrics per fold with their average and standard deviation (STD) over fold and the aggregated metrics over all folds i.e. on the whole dataset.

	Metric	Fold 1	Fold 2	Fold 3	Mean $\pm$ STD	Aggregated
Biopsy	**ACC**	87.23	84	83.33	84.85 $\pm$ 1.7	84.83
	SNS	93.33	71.43	64.71	76.49 $\pm$ 12.22	75.47
	SPE	84.38	93.10	93.55	90.34 $\pm$ 4.22	90.22
	PPV	73.68	88.24	84.62	82.18 $\pm$ 6.19	81.63
	NPV	96.43	81.82	82.86	87.04 $\pm$ 6.66	86.46
	AUC	0.92	0.90	0.83	0.88 $\pm$ 0.04	0.85
Nodule	**ACC**	83.33	91.67	82.61	85.87 $\pm$ 4.11	85.92
	SNS	88.89	100	77.78	88.89 $\pm$ 9.07	88.89
	SPE	80	86.67	85.71	84.13 $\pm$ 2.94	84.09
	PPV	72.73	81.82	77.78	77.44 $\pm$ 3.72	77.42
	NPV	92.31	100	85.71	92.67 $\pm$ 5.84	92.50
	AUC	0.86	0.96	0.87	0.9 $\pm$ 0.04	0.86

VI.4.3 Prediction Analysis

VI.4.3.1 Patch-Level Predictions

As we obtain actual instance predictions but have no instance-level ground truth, we shall look at different aspects of the available data as to infer their correctness. Only a quarter (24.4%) of the total patches (205 / 839) coming from malignant samples were predicted as malignant, with an average of $23.53\% \pm 27.36\%$ malignant patches predicted per malignant sample. While we have no way of assessing the correctness of the patch-level prediction with the current information available, we relate to a study [191] conducted on 300 breast cancers which revealed that the average tumor to tissue ratio in core-needle biopsies (under the standard histology protocol) is $42\% \pm 16.16\%$. Therefore, we can deduce that there might be an under-detection, either caused by the detection algorithm or the imaging itself because not all cells might be apparent in DCI as opposed to standard histology, moreover, the physical slice thickness is bigger than that of the optical slice, $\sim 5\,\mu\mathrm{m}$ as opposed to $1\,\mu\mathrm{m}$. On the other hand, looking at the false detection, only 2.9% (33 / 1151) of the patches coming from benign samples were predicted as cancerous, with an average of $4\% \pm 12.52\%$ malignant patches predicted per benign

sample. Still, these 33 misclassified instances, which we can be sure they correspond to false positives, gave rise to 9 misclassified samples coming from 7 biopsied nodules (see Figure VI.8).

VI.4.3.2 Degree of Cellularity

We remind the reader that for each patch has been visually inspected and labeled for presence or absence of cells. We extend this insight at the sample-level by defining the *degree of cellularity* of a sample as the percentage of patches containing cells from the total number of patches containing tissue (as opposed to background or imaging artifacts). Having this information, we will investigate if cell presence actually has an overall influence on the malignancy prediction.

Looking at the ground truth sample diagnosis we note that the malignant samples have a median cellularity of $53\% \pm 29\%$, while in the benign samples there are $31\% \pm 31\%$ patches containing cells per sample (See Figure VI.11). Also, according to Spearman's rank correlation the two variables, malignancy *vs.* degree of cellularity, are significantly correlated with a *p-value* $\ll 10^{-6}$ and $r = 0.4$.

Now when it comes to the malignancy prediction, cell absence and malignancy prediction are dependent in the sense that only a negligible number (15 / 1148 or 1.3%) of patches without cells were predicted as cancerous. This was expected and enforces our belief that stromal disorganization is not a diagnosis factor for breast cancer in DCI and that cell morphology and organization are the discriminating aspects. In Figure VI.9 it can be observed the distribution of predicted probabilities as per patches with or without cells coming from benign samples or malignant samples. We note that cell absence is correlated with malignancy absence, while for patches containing cells the prediction distribution is bi-modal in the malignant case (there are both tumor positive and negative cell-containing patches) and uni-modal in the benign case. Although we expect to have both normal and malignant cellular groups in a malignant sample, there is no way at this point of assessing their correctness.

However, at sample level, the correct prediction is overall independent from cellularity level, as the two variables are not statistically correlated with *p-value* $= 0.14$ and $r = -0.12$. Nevertheless, when we look at the different prediction outcomes (See Figure VI.12) we notice that the average cellularity of the falsely classified samples is closer to that of the predicted class than its true class. More specifically, false positives (FP) and true positives (TP) have higher degrees of cellularity (71%, 73%) than true negatives (TN) and false negatives (FN) (23%, 45%). For the case of FN, they could actually correspond to TN and could possibly reflect the noisy labels introduced by *Assumption 1*. While for the FPs, they could reveal a bias of the model to associate malignancy with hyperplasia. However, this bias was also concurred by the radiologist who themselves misclassified highly cellular benign samples as malignant, still, we had no feedback from the pathologist at that point in the study.

A possible fix to reduce false positives is by employing hard negative mining strategies, one such approach could be prioritizing high scoring but negative patches at bag creation. On the other hand, to reduce false negatives, we should refine *Assumption 1* by introducing some kind of uncertainty related to the shared diagnostic among same-nodule biopsies. Intuitively, this uncertainty should be correlated with the size of the biopsied nodule, as smaller nodules are more difficult to accurately sample. Luckily the information on the nodule size is available, thus extractable from the pathology reports and injectable into the model.

VI.4.3.3 Carcinoma Grade

Seeing the previous analysis results and knowing that one main characteristic of high grade cancers is extreme cell proliferation, we expect that more advanced cancers (with a higher carcinoma grade) are easier to detect. Their high tissue heterogeneity, cell abundance and deviation from the normal appearance, would potentially minimize all diagnosis ambiguity. In this regard, we analyze the statistics of the malignancy predicted probability per sample with respect to the ground truth carcinoma grade (1 to 3 on the Elston-Ellis grading system), where available. In Figure VI.10 we can see the corresponding box plots for each grade (1, 2, 3) and for the non-malignant samples (indicated as grade = 0) for completion, showing that the malignancy grade is directly proportional to the predicted malignancy probability. Hence, the average prediction probability for grade 1 is 65%, while for grade 2 raises to 94%, as for grade 3 it reaches 99%. Therefore, we can firmly state that high grade cancers have a lower chance to be missed. More precisely, by fixing the classification threshold $T = 50\%$ probability, for grade 1 the true positive rate is 70%, for grade 2 it increases slightly to 76%, while for grade 3, the most advanced cancers, the true detection reaches 87%. By comparison, for all cancer grades as a whole the true predictive rate is 75%, while for the benign class the true prediction reaches 90%. The result of this analysis could influence a shift of strategy in the use of the DCI technique towards the field of emergency intervention.

VI.5 Validation on Benchmark Dataset CAMELYON16

In our approach we focused on one directing axis - build a highly fluid MIL architecture which could be assembled from other model which benefited from fully-supervised training and then it could also be deconstructed in order to provide instance-level predictions. Therefore, we want to study the importance of intra-domain pre-training by applying the method introduced in this chapter also on a benchmark dataset. What is more, given that we do not dispose of dense annotations in our working dataset, we want to to validate the veracity of the instance-level predictions by taking advantage of the finer annotated CAMELYON16 data.

VI.5.1 Data

The 2016 ISBI Camelyon Challenge [165] was the first ever data challenge using histopathology images, namely 400 WSI provided by 2 centers in the Netherlands. The objective of the challenge was to asses the performance of deep learning algorithms for automated detection of metastases in hematoxylin and eosin (H&E) stained whole-slide images of lymph node sections adjacent to the breast. The dataset is split into 2 parts: one for training containing 270 WSIs (110 with metastasis and 170 without), together with their pixel-level ground truth as segmented metastatic zone and the held-out test set of 129 WSI on which the AUC would be computed in order to rank the different proposed submissions. The ground truth was verified by immunohistochemical staining and the pathologist performance was also reported. The best-performing pathologist did not have any time constraints when analyzing the images, while the others, i.e. a panel of 11 pathologists, had a 2h limit in order to simulate standard clinical circumstances.

The challenge implied solving 2 distinct problems: identification of individual metastases in WSI (i.e binary classification task at zone-level) and classification of every WSI as either containing or lacking metastases (i.e. binary classification task at image-level).

For the lesion-level detection task, 27.6% of individual metastases were not identified by the pathologist, corresponding to a lesion level sensitivity of 72.4%. For the slide-level classification task, the best pathologist achieved a sensitivity of 93.8%, a specificity of 98.7%, and an AUC of 0.966, While the panel of 11 pathologists achieved a mean sensitivity of 62.8% with a mean specificity of 98.5%, and the mean AUC of 0.810.

We are going to take a closer look at the WSI classification task, here the best algorithm [192] achieved an AUC of 0.994 by training an ensemble of 2 networks based on the InceptionV3 architecture and laborious data handling with stain standardization, extensive data augmentation and hard negative mining. This model was trained in a fully-supervised manner by taking advantage of the pixel-level annotation, however, we are interested in the MIL approach using only global labels.

In this regard, we turn to another data challenge aimed at learning from global labels which also provides the pre-sampled CAMELYON16 slides together with the tile features extracted with ResNet50 architecture (trained on ImageNet), therefore we bypass the strenuous data management task. In the provided dataset there are 279 slides in the train-set (112 with metastasis and 167 without) and 120 slides in the test-set (chose ground truth is unknown); from each slide there are sampled a maximum of 1000 tiles/slide of size 224×224 at $10\times$ mgnification, together with their 2048-dimensional ResNet50 features per tile. What is more, 11 slides (or approx 4%) are fully annotated, summing to 10K annotated tiles (out of which only $\sim$700 pathological).

VI.5.2 Method

The inspiration for this approach came from the need of rather having a specialized feature extractor instead of using transfer learning, however, for this proof-of-concept we only train an MLP on top of the ImageNet features for each tile by leveraging the dense labels. After training the MLP in a fully supervised fashion on the 10K tiles, we use it in a MIL context as the building block of a siamese multi-branch network, where each branch takes a image tile as input and the per-tile predictions are aggregated by MaxPooling to give the image-wise prediction. Compared to the previous approach, here the features of each instance are pre-computed, therefore having as input directly a vector with 2048 values instead of an RGB image (which would sum up to 150K values for a tile of size 224×224) and also having a considerable part of the network removed, allowing to create a network with more branches, namely a maximum of 1000 branches given that in the dataset this is the maximum number of instances per sample.

The MLP has two fully connected layers of 100 and 10 neurons with sigmoid activations. Since its input is a 2048-dimensional vector corresponding to the tile descriptor extracted from ResNet50, the MLP has $\sim$205K parameters to train which is disproportionate with regard to the number of data points i.e. 10K, therefore heavy dropout of 50% is applied. It was trained using the Adam optimizer with learning rate $lr = 0.001$ to minimize the binary cross entropy (BCE) loss weighted according to the class proportions (which are even more skewed than the image-level ones, unsurprisingly) with $w_0 = 0.53$ and $w_1 = 7.15$ for 30 epochs on batches of 100. The MIL training is performed using similar parameters i.e. optimizing weighted BCE ($w_0 = 0.84$ and $w_1 = 1.25$) for 30 epochs using the Adam optimizer with 0.001 learning rate and a batch size of 10. The 50% dropout is kept at this stage too.

Table VI.3: Metrics computed on the CAMELYON16 dataset at the sample-level: comparison between the proposed MIL method (and influence of pre-training the classifier); at tile-level: comparison between supervised training with tile labels (metrics aggregated over the 5 folds) and MIL training with sample annotations.

(a) Sample-level metrics

Training Paradigm	Train CV AUC	Test AUC
MIL	0.853 ± 0.099	0.833
Pretrained MIL +4% labels	0.962 ± 0.011	0.858
CHOWDER R=5 [177]	0.903	0.858
CLAM [193]	-	0.895
CLAM +10% labels [183]	-	0.924

(b) Tile-level metrics

Training Paradigm	AUC	ACC	SNS	SPE
FS	0.956	96	81	97
MIL	0.913	97	62	99

VI.5.3 Results & Discussion

With this experiment we tested two hypotheses:

1. the improvement of a MIL approach brought on by an intra-domain pre-trained backbone (See Table VI.3a);
2. the veracity of the patch-level predictions of a MIL model trained only on slide-level labels (See Table VI.3b).

For the first question, we trained a 2 layer perceptron on top of the ResNet50 extracted features in a multi-branch MIL model first randomly initialized via the Glorot [144] uniform method, then pre-trained on the tiles of 4% of the WSIs if a fully-supervised manner. The metrics in Table VI.3a represent the average $\pm$ standard deviation AUC obtained by 5-fold cross validation training on the challenge train-set, together with the competition AUC on the held-out test-set (model trained on the entire train-set). We observe an average improvement of 0.109 points on the train-set and 0.025 on the final set.

However, for this data sampling with small patches, the classic MIL assumption of the highest scoring instance should be less efficient than in the case of our context-aware sampling at millimeter-scale as an instance does not contain enough information to indicate the subtype, not to mention that a WSI is comprised of 100 times more tiles. Nonetheless, it seems like intra-domain pre-training compensates for this weak assumption, as the proposed method obtains the same performance on the test set (AUC 0.858) as CHOWDER which represents an image by $2 \times R$ tiles i.e. 10.

For the second question, we trained a MIL model on all non-densely annotated slides i.e. 268 slides (containing $\sim$ 250K patches), then extracted its constituting branch and used it to predict the labels of the $\sim$ 10K tiles coming from the 11 fully-annotated slides. Note that all 11 slides were correctly classified by the MIL model as containing metastases. In Table VI.3(b) we present the tile-level metrics and compare the fully-supervised test metrics aggregated over the 5 folds with the test metrics of the MIL model. Unsurprisingly, the fully supervised approach has superior prediction capabilities, the most stringent improvement being the 19 points gain in sensitivity compared to the MIL approach which still reaches a respectable 62%. On the other hand, correctly classifying non-malignant areas seems a less challenging task as both approaches reach very high specificity (97-99%), which together with the class imbalance causes very high accuracy (96-97%) too.

In [192] patch-level accuracy is 98% on the whole fully-annotated dataset, but it should be taken with a grain of salt since the class specific metrics are not disclosed. Unfortunately, if we were to compare with other MIL approaches, to our knowledge, there are no methods trained on the slide-level labels of the CAMELYON16 dataset that also allow computing reliable patch-level prediction. In [194] they infer a patch-level AUC of 0.64 obtained based on CHOWDER [177] by analyzing the scores and inferring

some thresholds related to the class prediction, however, the authors of CLAM [183] warn about the direct inference of the class from the patch attention scores. From another perspective, it might be a stretch to compare with the pathologist performance on the lesion detection task where they achieved 72% sensitivity [165].

VI.6 Conclusion

To conclude, the study in this chapter is showcasing the challenges of a real-world CAD problem: limited data and annotations, unconventional contrast and constant evolution of its applications.

We are advocating for plug-and-play model design that can be extended under different circumstances. More specifically, we present how a model trained in a fully supervised manner on a similar domain is adapted for the MIL framework and how it improves the results compared to an extra-domain pre-training. We also take full advantage of the limited dataset by smart sampling, adding pertinent complementary information etc.

We manage to achieve 89% sensitivity and 84% specificity by training on a breast core-needle biopsies dataset having only the global histopathology diagnosis as ground truth. We believe that the convergence of the method despite our small dataset is greatly due to the pre-trained architecture on similar data.

Given the transparent formulation of the problem, the method can be easily extended to include instance-level ground truth in the learning process. Nonetheless, minimizing the auxiliary effort of expert annotators greatly reduces the bottleneck of training DL models on DCI images. This is an important step towards building safe and robust tools that would support the clinical adoption of DCI. Upon gathering more data, we would obtain improved prediction results, in [195] they demonstrate that the MIL performance is significantly increased (in terms of reducing both the variance and the average validation error) and their order of grandeur was of thousands of WSIs. This method powered by a continuous cycle of data collection and algorithm improvement has potential to ultimately reform the extemporaneous diagnosis field, improving patient outcome.

Chapter VII

FFOCT *vs.* DCI Cross-Modal Representation Learning

In the previous chapters we have approached the two imaging modalities proper to the LightCT™ scanner separately. The choice of imaging was done according to the particularities of the pathologies studied, with a deeper focus on DCI as it bears paramount cellular information and it raises more questions by its nature and novelty, calling for more research. However, with the continuous technological improvements of the setup, acquiring both modalities concurrently in one sitting is perfectly feasible, therefore we are considering merging the information extracted from the two modalities in order to take advantage of their complementarity. Therefore, in this chapter, we introduce a proof of concept for learning a common latent space for DCI and FFOCT images via siamese convolutional neural networks and contrastive learning.

Contents

VII.1 Motivation

Our main motivation is to exploit the multimodal nature of our setup, in a quest to "overcome ones weaknesses with the other ones strengths". To be more clear, we are expecting to surmount the high inter-acquisition variations of DCI with the long-established robustness of FFOCT. In this regard we shall explicitly train on registered image pairs of both modalities which would lead to defining a common inter-modal embedding, endeavor which falls under the umbrella or representation learning. It would allow extracting mutual information from DCI and FFOCT that would most likely encode fiber characteristics. This is particularly of interest as fibers in FFOCT have a robust and reproducible rendering, to the point that their biological properties could be inferred from the FFOCT image. On the other hand, in DCI imaging fibrous tissue suffers great fluctuations in appearance, therefore it is not reproducible between acquisitions while also being the main source of imagining artifacts.

According to the theoretical formulation of DCI imaging, fibers should not be apparent at all in the image, as they do not contain actively moving scatterers like cells do. However, in the real world setting, where we face with both external (environmental) and internal (setup related) movements, the perfect theoretical conditions do not apply. These nuisances induce some random oscillations at fiber level, making them apparent mostly in the color channels corresponding to low and mid-range oscillatory frequencies (i.e. blue and green). In practice, we observed that there can be variations in fibers intensity both intra- and inter-acquisitions. The source of these permanent perturbations are unknown and not quantified yet. On top of that, there are also more severe imaging artifacts induced by singular events, like a door being slammed in the imaging room or liquid flowing inside the tissular creases present amid fibrous material. These phenomena cause saturated pixels in the affected zones hindering the image interpretation because the highly contrasted areas attract viewer's attention unnecessarily.

All these insights were gathered after experimenting with DCI longtime, important intuition about this behavior of DCI was also gained by comparing the images with their FFOCT counterparts of the same tissue sample, getting a sort of "ground truth" for the tissue architecture. However, these observations are purely subjective and not formally defined yet. Therefore, by the present work, we are aiming to build a

model able to "see" beyond these imaging inconsistencies of DCI and we are planning to implement it by having FFOCT as a training guide, akin to the human agent.

The work in this chapter is conducted on the breast excision dataset (in Section III.2.2.1) as it is the best curated one available to date. However, once the method is validated, it has broad potential to be expanded to multiple tissue types in order to incorporate general knowledge on fiber characteristics in DCI. This is namely the main reason in favor of this approach - generalization. It could play a powerful role in better understanding and future iterations of DCI technique. Another aspect in support of this statement, which actually also influences its general quality, is the "infinite" dataset generation: already registered FFOCT / DCI image pairs are seamless to acquire and there is no need of extra human intervention, like annotation or validation. Therefore, the proposed method directly exploits the intrinsic nature of data in an "unsupervised" fashion.

Even if we noticed in the qualitative results on the classification task on the same dataset (Section V.3) little importance attributed to fibrous structures in DCI, extra-cellular matrix remains an important marker in cancer diagnosis and overall tissue characterization. Normal stroma is significantly different from tumor associated stroma. In breast cancer, tumor-associated stroma contains an increased number of fibroblasts and immune cell infiltrates, enhanced capillary density, increased collagen and fibrin deposition, all which alter the structure and stiffness of the extra-cellular matrix [196]. Therefore, we expect this approach could also serve in other downstream tasks, like diagnosis or biomarker discovery.

VII.2 Context

VII.2.1 Unsupervised Feature Learning

Unsupervised learning is a type of machine learning where the observation itself is prioritized over the labels. Unsupervised learning is not used for classification or regression; instead, it is used to uncover underlying patterns, cluster, denoise, decompose data or detect outliers, to name a few. When the data quality does not correspond entirely with the prediction task to solve, one can opt for using networks pre-trained on huge datasets like ImageNet as a starting point. However, the knowledge encoded there, even if powerful, might not be sufficient for the task at hand as the two domains might be too divergent. In that case, pre-training directly on the target problem domain, but on an auxiliary task, might be the way to go. This is where self-supervised pre-training [197, 198] comes into play. The idea has been put into practice in various manners, but their common denominator is training a network on a pretext task that is more accessible, without needing manual label annotation. These methods are supposed to produce good features for other downstream tasks that learn with supervised learning objectives, e.g., classification. As the labels of these pretext tasks are generated automatically, they have the advantage

of "infinite" training set generation (depending on the task). Usually the performance on the pretext task is ignored as long as the thereby learned feature encoders perform well on the downstream target task.

Historically, encoder-decoder networks were proposed for projecting the input into a lower dimension and then reconstruct it; then the encoder would be reused for downstream tasks like classification. This type of networks that map the input to itself are called auto-encoders (AE) [199], there are variants learning a slightly modified input, like denoising AE [200] or masked AE [201]. Auto-encoders were indeed used for unsupervised feature learning [202], but they were short-lived as they have been proven not to be able to capture fundamental concepts about the data; in [203] they experimentally dismiss this statement *"Features good for reconstruction are inherently good for classification tasks"*. Nonetheless, in 2022, pre-training with auto-encoders is showing signs of revival [204–206] as they seem to perform better when coupled with transformers [207] - the newest type of neural network architecture that promises to disrupt the classic AI paradigm.

If learning the input itself was controversial, the community has come up with more complex toy tasks to train their AI "babies", passing from a "learn by seeing" to a "learn by doing" philosophy. Some methods rely on linear transforms, in [208] they predict rotation angle, which stands on the hypothesis that only if a model has the visual commonsense of what the object should look like in its natural state could it recognize the correct rotation of an object, therefore this is not suitable for orientation invariant problems like tissue analysis. Other approaches are based on patch shuffling, like predicting relative patch locations [209] or learning patch permutations by solving jigsaw puzzles [210], this was also successfully applied to medical imaging [211]. There is also a generative family of proxy tasks like predicting missing pixels, a.k.a. inpainting [186] or coloring images, trained by decoupling grayscale and color channels in Lab colorspace, inspired by the visual cortex, thus called split-brain autoencoder [212]. What all these methods have in common is how seamlessly they could be applied on a variety of datasets, as they address elemental properties of images in they generality, however, this comes with the drawback of not injecting specific knowledge into the model. After all, the choice of the method resides in the capacity to encode useful information for the downstream task from the dataset at hand, whose success is not guaranteed.

VII.2.2 Contrastive Learning

A "richer" problem formulation for tackling self-supervised pre-training is arbitrated by contrastive learning (CL). As compared to the previous examples which deal with one data sample at a time, contrastive learning considers groups of samples, thus learning through input relationships. This paradigm exploits the idea of comparing data points against each other so there can be identified characteristics that are shared by different data classes and characteristics that distinguish one class from another. In

this context we define the concepts of positive and negative samples, where positive samples are pairs of data points belonging to the same class, while negative samples are "contrasting". Please note that in this context the term "class" has a broader meaning than in the classification framework.

Given the level of a priori knowledge available on the dataset, we can distinguish two families of CL methods:

1. self-supervised contrastive learning (SSCL) [213, 214], where there is no class information available, so as in the case of auto-encoders, the original sample and a transformed version of it are used to form the positive pairs, while any two different samples shall construct the negative ones;

2. supervised contrastive learning (SCL) [215, 216], provided that some additional information is available to distinguish between samples, positive pairs are built from different samples sharing a known characteristic i.e. class, while negative pairs are constructed from samples of different classes, with the result that samples from the same class are brought together in the learned latent space, while samples from different classes are distanced in that same latent space; which makes this concept interpretable as a metric learning [217] problem.

Moreover, a CL solution can be implemented under two similar settings: either pairs are considered independently and iteratively or they are considered in groups of three, where one sample is the anchor and the others are their positive and negative counterparts, respectively; therefore, in this latter case, both relationships are evaluated concomitantly. These formulations are translated in the model architecture and loss function, respectively.

The leading methods in self-supervised contrastive learning, SimCLR [218] and MoCo [179], are closely competing against each other in a race against time, suggesting the high stakes at play in finding a generalized task-agnostic pre-training framework. MoCov2 [219] combines the key concepts from both methods: momentum contrast for effective batch generation and training from MoCo with more extensive data augmentation and adding an MLP projection head as in SimCLR. While these methods use multiple negative pairs for each positive pair, BYOL [220], on the other hand, uses only positive pairs; idea which incited some controversies further analyzed in [221] to confirm on why it manages to generalize well regardless. In NNCLR [222] they use the nearest neighbor in the latent space in order to construct positive pairs. In [216] class labels are leveraged for generating positive pairs and SimCLRv2 [223] uses self-supervised contrastive learning with supervised fine-tuning and knowledge distillation to improve classification accuracy with as little as 1% of the class labels.

We have overviewed the leading methodological works, however, the strength of contrastive learning lies in the myriad of diverse applications it can power. In [193] CL pre-training is used for histopathology images, however the main improvement from baseline is given by a transformer network based on multi-head self attention [207, 224] rather than CL, while in [225] CL pre-training is used to improve segmentation on medical images with limited annotations and in [178] MoCo pre-training on the target domain is used to improve the performance and robustness a previous MIL approach [177] on histology image classification. One recent breakthrough in computer vision is DALL·E [226], an OpenAI model that creates hyper-realistic images from text captions for a wide range of concepts expressible in natural language. And this is not the only application where the fields of computer visions and natural language processing (NLP) come together under a multimodal [227] contrastive learning framework. Other examples include mainly cross-modal image and sentence retrieval [228–230] which can be seen under a representation learning lens as they implicitly learn the individual concepts present. As for the medical field, in [231] they make use of the textual interpretation of x-ray images, while in [232] they combine histology images with genomics data. However, we are most interested in the case were multimodality is understood as a combination of different imaging techniques, which are plentiful in the case of medical imaging modalities, some examples include: multimodal medical image fusion [233], registration of multi-stain histopathology / immunohistochemistry images in the common latent space [234], co-training on separate stain channels [235] from H&E. CL can be also used in longitudinal studies to encode time invariant properties, in [236] there is an example of training on the same modality (aerial photos) at different time steps.

VII.3 Method

For our objective, we shall train a model in contrastive fashion in order to learn a common latent space for FFOCT and DCI imaged samples. The said model is embodied by a siamese network which has the quality of being applicable on multiple inputs from which it extracts corresponding feature vectors, based on which a metric is learned, for the present case, the cosine distance is used. All the implementation details are presented in this section, from the architecture design to its training process which focuses on fully exploiting the data via online batch generation.

VII.3.1 Architecture

VII.3.1.1 Siamese Network

Due to their multiple inputs, all the applications enunciated in the previous section are possible thanks to multi-stream network architectures. The composing individual sub-networks can differ greatly among

them or not at all, depending on the nature of the relationships between the inputs themselves and the input-output relation. For example, in the case of an image + text contrastive learning model, the individual streams are often constructed via CNN and LSTM architectures (which represent the state-of-the-art in CV and NLP, respectively). To generalize, regardless of the input type, if the information to extract from the individual inputs is different, then the networks would be disjoint too, with a late information fusion of the obtained features vectors. On the other hand, when dealing with similar modalities (e.g. images only) and looking to extract similar information from all inputs, like in our case, the intuition is to use similar encoders for the inputs. The concept of an identical multi-stream network is represented by Siamese Neural Networks [135], also called twin networks, which have been around for as much time as the author of this manuscript 😄. The idea was introduced in 1993 by Baldi & Chauvin for fingerprint recognition [237] and shortly after it was "baptized" by LeCun and used to solve signature forgery verification [238] as an image matching problem. As the name suggests, a siamese network is most often constructed by a pair of identical sub-networks which work in tandem to extract common features from the inputs and are joined by an energy function at the top.

The goal of a siamese network is to learn a common encoding function Φ that maps two different representations $x_i^a \in \mathcal{I}^a$ and $x_i^b \in \mathcal{I}^b$ of the same instance $s_i \in \mathcal{S}$, but with $x_i^a \neq x_i^b$ onto a common domain $\mathcal{F}$ such that $f_i^a = \Phi(x_i^a),\ f_i^a \in \mathcal{F}$ and $f_i^b = \Phi(x_i^b),\ f_i^b \in \mathcal{F}$ and their distance Δ in this domain is near-zero, i.e. $\Delta(f_i^a, f_i^b) \to 0$. By extension, for two different samples s_i and s_j the distance should be greater $\Delta(f_i^a, f_j^b) \gg 0$.

In most of the applications present in the literature and recalled above, x^b is a transformed version of x^a s.t. $x^b = T(x^a)$ where T is a linear operator from the data augmentation family, which implies that the input domains are quasi-equivalent $\mathcal{I}^a \sim \mathcal{I}^b$. However, in our case, on each tissue sample s_i there are applied two very different imaging projections, one corresponding to FFOCT modality: $\Omega^a : \mathcal{S} \to \mathcal{I}^a$ and one to DCI: $\Omega^b : \mathcal{S} \to \mathcal{I}^b$, making the two input domains $\mathcal{I}^a$, $\mathcal{I}^b$ more challenging to bring closer into the common embedding domain $\mathcal{F}$. Intuitively, this discrepancy between the two domains, pushes the network to learn and encode finer, more meaningful and elemental qualities of the two domains, in case of convergence.

We have adopted VGG16 as the backbone architecture to keep consistency with prior work (c.f. Chapters V and VI). However, we find through experiments that, unlike the previous problems solved, the present problem is less sensitive to the network architecture chosen, other SOTA networks (i.e. ResNet50, InceptionV3) performing comparably (at least judging by the train / test losses obtained, without the thorough analysis of Section VII.4). Given the nature of the problem, the two feature extraction sub-networks can be completely disjoint (i.e. learning independent features from each input) or completely coupled (i.e. learn common features between the inputs) or somewhere in between, with

early or late weight sharing or fusion. The most used approach that also fits out problem is the case of identical networks that share their weights and fusion at the very end. Therefore, the function Φ would be encoded by a VGG16 model applied unequivocally on both inputs x^a, x^b. Due to this design the model is symmetrical, meaning that the inputs could be inversed without a change in the output. Please note that as DCI images are 3-channel RGB images and FFOCT images grayscale, we shall clone the FFOCT channel to match the dimensionality of its counterpart.

Each of the two fully convolutional branches are followed by an operation meant to construct the feature vectors f^a and f^b, operation computed through a flattening layer which collapses the spatial dimensions of the activations of last convolutional layer into a 1D vector. To make a comparison with the previous classification approaches, there we used a GAP pooling method, which gatherers the spatial dimensions and reduces the feature vector to the average activation of each kernel in the respective layer. This approach is useful for detecting patterns in data, agnostic to their coordinates, however, in the present case, localization is of uttermost importance as we are looking to match corresponding features from the two images, hence the choice of the flattening layer.

We can take advantage of the fact that no class information is needed to split the $1\,440 \times 1\,440$ registered DCI and FFOCT images further into tiles to produce a richer training set. The choice of the patch size, and consequently the size of the inputs, is made mostly empirically as the minimum size needed to contain sufficient fibers to make a decision. The chosen size for the inputs x^a, x^b is 480×480, which given the architecture (see Figure VII.1) leads to a size for the resulting feature vectors f^a, f^b of $15 \times 15 \times 512 = 115\,200$ each. But most importantly, the presented architecture can handle any input size, as there are no fully connected layers succeeding the feature embedding, with the only condition that the two inputs have the same size.

VII.3.1.2 Cosine Similarity Computing Embedded Node

Having defined the independent vectors $f^a = \Phi(x^a)$ obtained on inputs $x^a \in \mathcal{I}^a$ from one domain and the feature vector $f^b = \Phi(x^b)$ resulting from the input in the second domain $x^b \in \mathcal{I}^b$, we now have to define the distance function Δ.

The Euclidean distance is the classical example of the distance in a metric space. It is calculated as the square root of the sum of the squared differences between the two vectors, however, in computation intensive settings like machine learning it is common to remove the square root operation in an effort to speed up the calculation. Given its formulation $\left\| f^a - f^b \right\|$ we can see that if the vectors are not normalized, the metric is unbounded and vectors with large values will dominate the distance measure.

However, for our problem, we are looking not at the amount of activation response, but rather at the fact that the same corresponding pixel region is activated in both images. Moreover, we intuitively expect

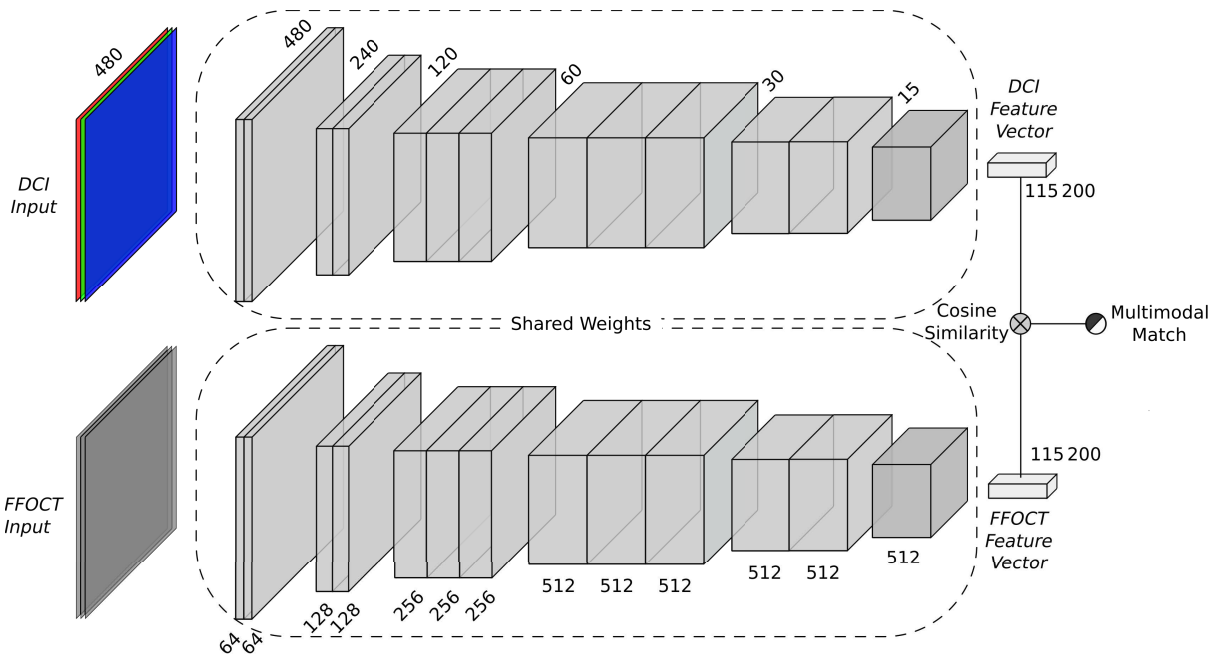


Figure VII.1: Siamese network architecture with VGG-16 convolutional block and integrated node computing cosine similarity serving as extractor of common features shared between DCI and FFOCT images

the important intensity variations present in DCI images to be propagated in the level of activation, therefore, the metric chosen should be robust to this aspect.

Cosine distance is a measure[1] based on the cosine of the angle θ between two vectors, which at its turn is a dimension of similarity i.e. the cosine similarity. The cosine of two vectors is computed as their inner product divided by the product of their lengths:

$$\text{Cosine Similarity:} \quad cos(\theta(f^a, f^b)) = \frac{f^a \cdot f^b}{\|f^a\| \cdot \|f^b\|} = \frac{\sum_{k=1}^{n} f_k^a f_k^b}{\sqrt{\sum_{k=1}^{n} f_k^{a2}} \sqrt{\sum_{k=1}^{n} f_k^{b2}}} \tag{VII.1}$$

$$\text{Cosine Distance:} \quad \Delta(f^a, f^b) = 1 - cos(\theta(f^a, f^b)) \tag{VII.2}$$

For two proportional vectors $f^a = c \cdot f^b$ i.e. their angle null $\theta(f^a, f^b) = 0$ have a cosine similarity of 1, while two orthogonal vectors ($\theta = \frac{\pi}{2} = 90°$) have a similarity of 0. Cosine similarity and, by extension, cosine distance are criteria of orientation agnostic to the magnitude of vectors. Another major perk of using this metric in the present scope is the fact it is neatly bounded by [0, 1] in the positive domain.

In the feature embedding space we are comparing two very sparse vectors in very high dimensions and cosine similarity is affected only by the terms the two vectors have in common, whereas Euclidean has a term for every dimension which is non-zero in either vector. In relation with our problem, the fibrous

[1] N.B. Cosine distance is not technically a metric as it does not satisfy the triangle inequality.

structures we are aiming to match in the two image modalities are always present in the FFOCT image and by extension in the embedding vector, while in DCI imaging not all fibers are always detectable, moreover, they might appear with highly unstable intensities independent of biological factors. This motivates the choice of cosine similarity as a metric, accordingly, we shall add a neuron computing the normalized dot product of the feature vectors f^a and f^b at the output of the network. Expected output for positive input pairs (x_i^a, x_i^b) will be closer to 1 and for negative pairs (x_i^a, x_j^b) nearing to 0.

VII.3.2 Training

In this section we detail the training of the network described above and depicted in Figure VII.1 via online batch generation and using a classification inspired loss. Present experiment has been conducted on the 47 breast excision dataset (detailed in Section III.2.2.1) containing carefully curated ROIs imaged with both FFOCT and DCI. The train / test split follows the same 80 / 20 rule resulting in 37 samples for training and 10 for testing. When it comes to ROIs, we are training on 403 image pairs and testing on 124 pairs, all sharing the same size i.e. $1\,440 \times 1\,440$ pixels.

VII.3.2.1 Online Batch Generation

As we have mentioned in the introduction, one motivation towards this contrastive representation learning approach is leveraging the already registered images, which serve as an inexhaustible training set. Firstly, there are $N^+ = 403$ unique positive image pairs and $N^- = C\binom{N^+}{2} - N^+ > 80K$ possible negative pairs. To this we add the image augmentations which consist in 3 levels of image contrast, for DCI images only, and flips (vertical and/or horizontal), synchronized for DCI and FFOCT positive pairs. Furthermore, we use extracted sub-regions of 480×480 that artificially grows the dataset. During each training epoch we are extracting k patches from each image, and since it is online generation[2], the patches are slightly different between each epoch, acting like data augmentation, helping convergence and generalization. In theory, from each $1\,440 \times 1\,440$ image we could extract $k = C\binom{960}{2} > 450K$ distinct patches of size 480×480, however, the insignificant variation introduced by displacements of only one pixel would not justify the great training overhead, therefore, we choose $k = \frac{\text{image width} \times \text{image height}}{\text{patch size}^2} = 9$. We can estimate that the number of positive input pairs seen by the network during training time is equal to $N \times k \times card(\mathcal{T}) \times n_{\text{epochs}} \approx 10\,800 \times n_{\text{epochs}}$, where N is the number of positive image pairs, k the number of patches randomly extracted from each image at each training step, $\mathcal{T}$ is the set of image transformations applied for augmentation and n_{epochs} is the total number of training epochs. For

[2] In this context, the term *generation* is employed in relation with data generators used in programming, which are functions that can be called multiple times and yield an iterator at each call, hence, they are used to encode the logic for feeding batches to the training loop.

validation generators we always sample the central patch of the images to ensure reproducibility among epochs in order to monitor the test metrics confidently.

For positive pairs of images we randomly choose a position and extract corresponding patches from the image pairs DCI and FFOCT (x_i^a, x_i^b), while for negative pairs of patches (x_i^a, x_j^b) we randomly choose two non-matching images and select a random positions from each. To respect the class balance principle we create symmetrical batches with the same number of negative and positive pairs. Again, the strategy for choosing the batch size is in accordance with the biggest size that fits into the available memory. Accordingly, the batch size equals to 12 with 6 positive pairs of patches and 6 negative pairs of patches. Note that at the point of this experiment the GPU used in the memory-hungry MIL approach was not available, so batch size can be potentially increased in future experiments, given the rapid evolution of GPU technology. Also, regarding hardware aspects, given the online batch generation that requires multiple I/O calls[3], it is introduced a high computational overhead, but this approach ensures convergence due to the diversity injected into the network, especially the high variation of negative samples seen.

VII.3.2.2 Loss Function

In the scope of deep metric learning two families of loss functions are generally used to train for the positive and negative evidence: pairwise losses, which encode the negative evidence implicitly, and triplet losses for explicit negative association. Regardless, both categories compare distances between representations of data samples.

One could possibly apprehend training directly on the distance of the positive samples only, by directly minimizing the distance as a loss function i.e. forcing the distance towards zero. However, by simply doing so would lead to a collapsed solution, since both the distance and the loss could be made zero by setting the feature vectors to a constant [239]. Therefore, negative samples indeed add fundamental value.

Triplet loss is a loss function where a reference input (called anchor) is compared to a matching input (positive) and a non-matching input (negative). The distance from the anchor to the positive $\Delta(f, f_+)$ is minimized, and the distance from the anchor to the negative input $\Delta(f, f_-)$ is maximized such as their difference is greater than a margin m.

$$L_{\text{triplet}}(f, f_+, f_-, m) = \max(0, \Delta(f, f_+) - \Delta(f, f_-) + m) \tag{VII.3}$$

[3]Input/output communication. Here, repeated transfers of data from storage (hard drive) to memory (RAM, VRAM).

However, one important drawback is the need for a three-branch architecture as opposed to classic siamese networks which requires more computational power, inevitably sacrificing the batch size. Moreover, in Google's paper for face recognition using triplet loss [240] it is mentioned the difficulty to train this kind of networks as well as the necessity of hard triplet mining for convergence (i.e., searching for negatives similar to the anchor and positives more dissimilar) which introduces more overhead and increases the development burden.

A common pairwise ranking loss that comes under many names is the contrastive loss a.k.a margin or hinge loss, which, much as the earlier example, ensures that the distance between positive pairs is small while the distance between negative pairs is at least m. However, as opposed to the triplet loss these relationships are established independently and sequentially.

$$L_{\text{margin}}(f_i, f_j, m) = \begin{cases} \Delta(f_i, f_j)^2 & \text{if } i = j \\ \max(0, m - \Delta(f_i, f_j))^2 & \text{if } i \neq j \end{cases} \tag{VII.4}$$

With its two distinct case, contrastive loss bears a resemblance to the traditional cross entropy loss. As a matter of fact, an insightful work [241] proves theoretically and experimentally that classification-based losses are analogous in performance and even definition with ranking losses especially designed for metric learning. By the same token, more works [242–249] find classification-based losses to work better in their applications.

Based on these finding and given that our distance function Δ is defined on $[0, 1]$, we found pertinent to start by testing a binary cross-entropy loss first and, in light of the obtained results (see Section VII.4), we finally deem it extremely effective. As opposed to the previous loss definitions given in a general form in relation to any distance function Δ, we formulate the loss function used in relation with the cosine distance and the predicted angle $\hat{\theta}$ between the cross-modal feature vectors. Given we cannot actually estimate the exact true value of θ, we apply a hard margin constraint on the distance definition, so we can express the true relation between feature vectors, as: $\Delta(f_i^a, f_i^b) \to 0 \implies f_i^a \simeq f_i^b$ and $\Delta(f_i^a, f_j^b) \gg 0 \implies f_i^a \neq f_j^b$ which formulates the problem as a binary classification problem. Accordingly, the loss function is defined as follows and plotted in Figure VII.2, given the angle between vectors θ and the angle predicted by the siamese network $\hat{\theta}$:

$$L(\theta, \hat{\theta}) = \begin{cases} -\log(\cos(\hat{\theta})) & \text{if } \theta = 0 \\ -\log(1 - \cos(\hat{\theta})) & \text{if } \theta > 0 \end{cases} \tag{VII.5}$$

From the conducted experiments, we observe that the present approach is not sensitive to training hyper-parameters either, therefore suggesting that the problem is sufficiently well-posed. The results reported

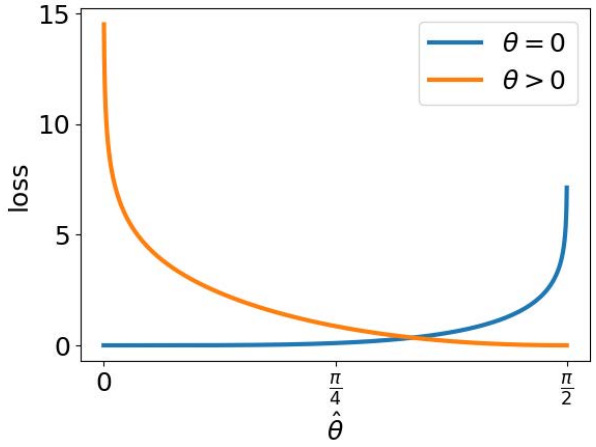


Figure VII.2: Training loss for matching image pairs (x_i^a, x_i^b) having colinear feature vectors therefore $\theta(f_i^a, f_i^b) = 0$ and for the case of non-matching image pairs $\theta(f_i^a, f_j^b) > 0$.

in the next section are obtained by training on the described loss via a SGD optimizer with learning rate of $lr = 0.01$ and weight decay l_2 regularization of $1e - 5$, on a batch size of 12 (6 negative pairs and 6 positive pairs). The method was fast to converge and training was stopped after 35 epochs when losses stopped significantly improving.

VII.4 Results

In the contrastive learning literature, the performance on the pretext tasks is usually deemed not important as the focus is on the main task. Nevertheless, in our case, the objective is learning an accurate and reproducible representation of DCI images in analogy with the FFOCT modality which we shall verify through both quantitative and qualitative methods. It is worth mentioning that in our criteria of choosing the best performing model, qualitative results outweighed quantitative results, as we had obtained similar performance metrics (i.e. low validation losses) across multiple experiments, regardless, some models produced noisy filters or uninterpretable attention maps.

For the sake of completeness, we have tested the classification power of the features extracted with this approach from DCI images and we obtained an accuracy identical to when training directly on FFOCT images, i.e. 75% binary classification accuracy baseline. This result can actually be seen as a successful sanity check; it represents a good indication that the learned latent space offers a complete picture of tissue architecture in DCI images.

In the following sections we report the obtained cosine distance between all combinations of cross-modal image pairs coming from the training data set and testing set, respectively, both positive and negative pairs. We also define two measures meant to asses for the robustness of the method, the identity error and

the symmetry error, inspired by the metric learning flavor of this contrastive learning method. Finally, we shall look at the learned filters and obtained attention maps through some examples.

VII.4.1 Quantitative Results

VII.4.1.1 Learned Distance

We show the metrics obtained with the presented approach in Table VII.1 by reporting the statistics of the cosine distance obtained on the positive data pairs ($N^{+}_{\text{train}} = 403$, $N^{+}_{\text{test}} = 124$), as well as on *all* the possible combinations of negative pairs ($N^{-}_{\text{train}} > 80K$, $N^{-}_{\text{test}} > 7.5K$). Please note that while training was conducted on multiple 480×480 px extracted tiles, the metrics are computed based on the features extracted on the entire $1\,440 \times 1\,440$ px FOVs, therefore the feature vectors f are 9 times bigger compared to those of the tiles used for training. This proves that the method is scalable and effective. Another argument in favor of this statement is the closeness of the train and test metrics: the 5$^{\text{th}}$ to 95$^{\text{th}}$ quantile interval is identical to up to 3 decimal places for positive samples, while for negative samples the distances spread out more toward lower values for the test set (see Figure VII.3 for the distribution of the pair distances on the test images). What is more, we have a well-delimited margin between the maximum distance of positive pairs ($\max_i(\Delta(f^a_i, f^b_i)) = 0.283$) and the minimum distance of the negative ones ($\min_{i \neq j}(\Delta(f^a_i, f^b_j)) = 0.526$) - a $\min - \max$ hard margin of 0.24 and an interquartile soft margin of 0.58. By making a parallel with the other losses explicitly integrating a margin m in the training process (i.e. L_{triplet} and L_{margin}), we notice that a margin is nonetheless implicitly learned in the case of our cross-entropy loss, in accordance with the findings in [241]. Consequently, there are no confusions between positive and negative pairs on the test data, however, interestingly, this is not the case for the train test, where we notice some false "pairings". Namely, five positive pairs have an abnormally high distance $\Delta(f^a_i, f^b_i) > 0.5$ in the embedding space, but upon further analysis we observe that the majority were due to data curation errors (three mismatched and one flipped) and one difficult example due to architectural homogeneity and lack of salient fibers to correlate on. There are also two non-matching pairs with $\Delta(f^a_i, f^b_j) < 0.5$ which upon visual inspection seem to be indeed quite similar. If anything this is an indication of a good generalization and lack of overfitting on training data. Regardless, looking at the inter-quantile distance of the training set it also presents a wide soft margin of ~ 0.6.

VII.4.1.2 Identity Error

If we are viewing this problem from the metric learning optic, then we shall look at the identity property, which is one of the key properties that define a function as a metric. It states that if the distance between two entities is zero then the entities are identical. In order to test this hypothesis for our case, we shall

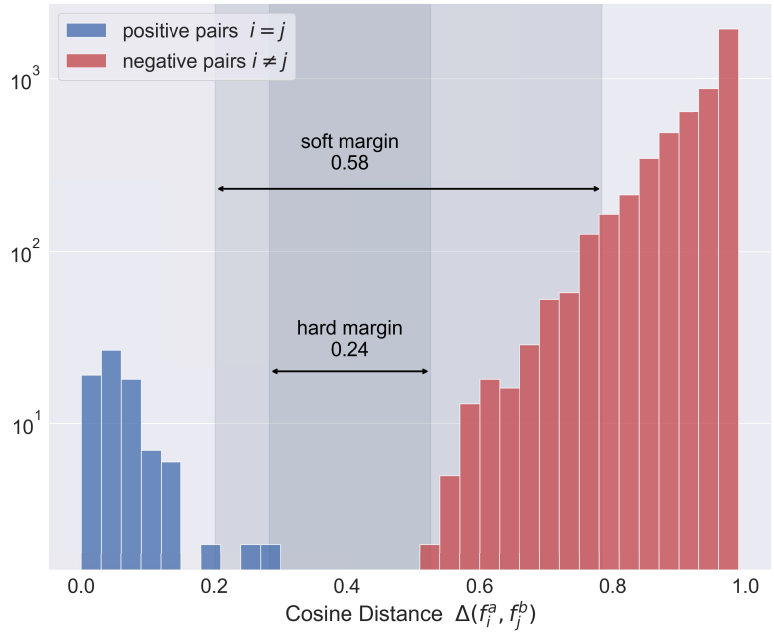


Figure VII.3: Distribution of the cosine distance of all possible combinations of image pairs over the test set, both positive and negative pairs.

compute the distances between multiple transformed versions of a DCI image $x_{i'}^a = T'(x_i^a)$ and its original FFOCT counterpart x_i^b, under the assumption that the transformation $T' \in \mathcal{T}$ should not alter the distance between the images. Note that $\mathcal{T}$ represents the set of image transforms that were also used for training, they correspond to different LUTs for coding the DCI signal in the RGB space, resulting in different contrast and saturation values which lead to different saliency levels of the present structures. We extract all the corresponding feature vectors $f_{i'}^a$ and we are looking to define the identity error $\varepsilon_{ii'}$ quantifying the deviation among all the computed distances $\Delta(f_{i'}^a, f_i^b)$ between the transformed DCI images and the FFOCT image as follows:

$$\varepsilon_{ii'} = \frac{1}{\text{card}(\mathcal{T})} \sum_{i'} \left| \Delta(f_{i'}^a, f_i^b) - \overline{\Delta(f_{i'}^a, f_i^b)} \right| \text{ where } x_{i'}^a = T'(x_i^a) \text{ and } T' \in \mathcal{T} \tag{VII.6}$$

We compute the statistics of $\varepsilon_{ii'}$ for all positive pairs, the mean identity error $\bar{\varepsilon}_{ii'}$ could be interpreted as a percentage given it is bounded by 0 and 1; therefore, the mean identity error for the train set is $\bar{\varepsilon}_{ii'} = 0.4\% \pm 0.8\%$, while for the train set is $\bar{\varepsilon}_{ii'} = 0.4\% \pm 1.1\%$.

VII.4.1.3 Symmetry Error

We shall also look at the symmetry property, which is another key property of a metric function. It states that the distances between two entities be equal regardless of the direction. In our case, if we consider the entity as being the real-life tissue sample, then the distance between two tissue samples s_i and s_j should be equal in the cross-modal space, regardless of the imaging modality chosen for either. By computing

Table VII.1: Metrics obtained on the cross-modal pairs on the train and test sets: **cosine distance** on **positive pairs** $\Delta(f_i^a, f_i^b)$, cosine distance on **negative pairs** $\Delta(f_i^a, f_j^b)$, as well as the **identity error** of transformed positive pairs $\varepsilon_{ii'}$, as well as the **symmetry error** between negative pairs ε_{ij}.

		mean ± std	**min - max**	**Q5 - Q95**
Test	$\Delta(f_i^a, f_i^b)$	0.071 ± 0.059	0.007 - 0.283	0.012 - 0.202
	$\varepsilon_{ii'}$	0.004 ± 0.008	1e-5 - 0.048	1e-4 - 0.019
	$\Delta(f_i^a, f_j^b)$	0.939 ± 0.073	0.526 - 1	0.784 - 0.998
	ε_{ij}	0.009 ± 0.012	1e-6 - 0.145	1e-4 - 0.03
Train	$\Delta(f_i^a, f_i^b)$	0.078 ± 0.121	0.006 - 0.986	0.013 - 0.207
	$\varepsilon_{ii'}$	0.004 ± 0.011	1e-5 - 0.147	1e-4 - 0.011
	$\Delta(f_i^a, f_j^b)$	0.966 ± 0.044	0.173 - 1	0.873 - 0.998
	ε_{ij}	0.008 ± 0.015	1e-7 - 0.810	1e-4 - 0.03

the (cosine) distance matrix between the feature vectors of all FFOCT and DCI images we observe that it is asymmetric, therefore the distance between the DCI rendering of one sample and the FFOCT of another is not equal to the distance between the DCI image of the former and FFOCT of the latter, i.e. $\Delta(f_i^a, f_j^b) \neq \Delta(f_j^a, f_i^b)$ when $i \neq j$. However, this is not surprising, given the little robustness of DCI imaging. In order to put a number on this discrepancy, we define the pairwise symmetry error:

$$\varepsilon_{ij} = \left|\Delta(f_i^a, f_j^b) - \Delta(f_j^a, f_i^b)\right| \text{ where } i \neq j \tag{VII.7}$$

We compute the statistics for all negative pairs, the mean symmetry error $\bar{\varepsilon}_{ij}$ could be interpreted as a percentage given it is bounded by 0 and 1; therefore, the mean symmetry error for the test set is $\bar{\varepsilon}_{ij} = 0.9\% \pm 1.2\%$ and for the train set, respectively, $\bar{\varepsilon}_{ij} = 0.8\% \pm 1.5\%$.

Based on these metrics detailed in Table VII.1 we can conclude that the proposed method is a robust solution for characterizing DCI images in relation with the more reproducible FFOCT and the learned representation could serve for future developments.

VII.4.2 Qualitative Results

Since this is a representation learning method employed in an effort to obtain a robust extraction of fiber characteristics from DCI imaging, we also apply qualitative analysis methods in order to get a grasp of the nature of the features learned. As in previous experiments, we are visualizing attention maps via Grad-CAM and learned patterns via maximization of filter activation with gradient ascent.

From visualizing the activation maps we observe two validating phenomena:

- *Low contrast fibers are captured by the network*, so what is otherwise inconspicuous to the naked eye can be revealed using the present method. See example 1 of Figure VII.4 where the fibers in the lower half of the DCI image (a1) are less contrasted than those in the upper half, to such extent that their orientation can not be estimated; regardless, the activation map trained (d1) on the modality matching task captures the correct orientation, which can be verified by looking at the corresponding FFOCT image (b1). Note that the obtained map is agnostic to the FFOCT image, as it was computed on the DCI image, moreover, the sample belongs to the test set.

- *Imaging artifacts are understood by the network, not confused with fibers and discounted*, which is usually not the case for a human user. See the second example from Figure VII.4 where the DCI image (a2) displays artifact as highly saturated green-yellow traces, which are present both at the fiber level (top left corner) and caused by the liquid medium flowing in the interstitial space (bottom half and top right). We deduce this by comparing with the FFOCT acquisition (b2) and observe that in the attention map (d2) this difference was also captured.

In both attention maps (Figure VII.4 d1 & d2) we also notice a natural side-effect of the training objective; namely, besides the fiber features, the holes in the tissue serve as matching features for associating the two modalities. While this seems to not overshadow the fiber detection, it might be undesirable in other downstream tasks and should be acknowledged.

Another aspect worth highlighting is the feature complementarity among the features learned via the tumor classification task (detailed in Section V.3) and the modality matching task. With these two methods combined we achieve a *complete representation* of DCI images, learning features related to cell appearance as well as fiber structures, respectively. This can be best noticed by looking at the two examples of pairs of activation maps in Figure VII.4 (c1 & d1, c2 d2) which offer a "reading" of the corresponding DCI image, acting as an aids to interpretation. This complementarity of image characteristics is also nicely illustrated in Figure VII.5 showing some examples of learned patterns encoded by the deeper layers of the two networks.

VII.5 Conclusion

To sum up, we presented a method of representation learning that leverages the intrinsic duality quality of the imaging setup producing registered DCI and FFOCT images of the same tissue. In an effort to overcome the high undesired variability of DCI with the robustness of FFOCT, we designed a siamese network and trained it in a contrastive fashion to allow for reproducible fiber representation in DCI

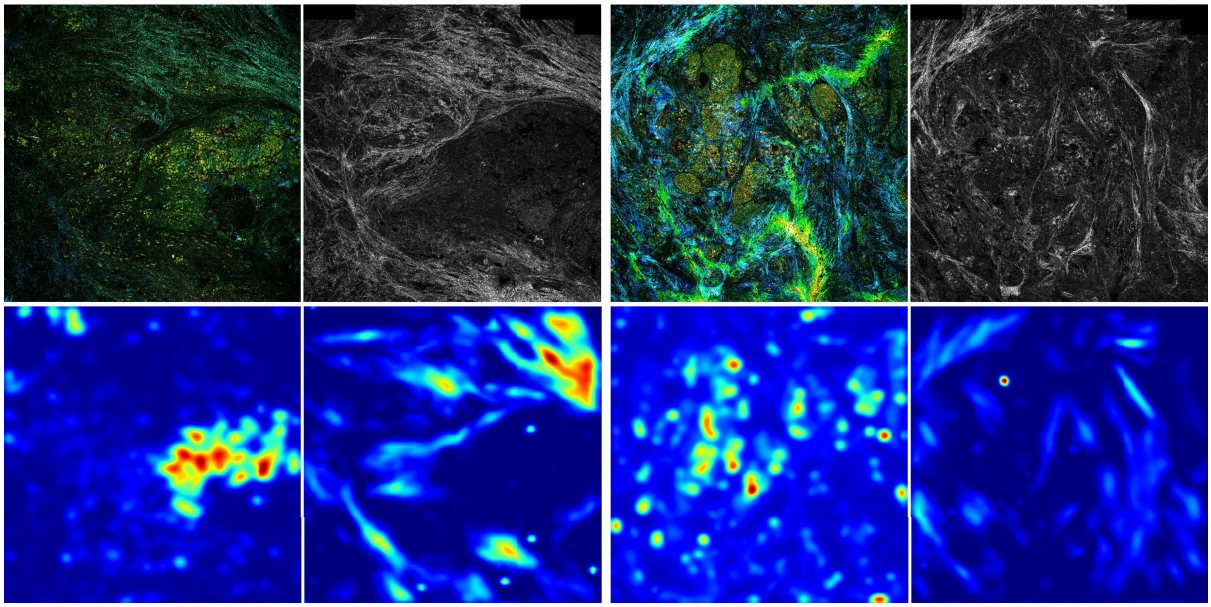


Figure VII.4: Two comparing examples of **activation maps** of the VGG-16 model trained on the **tumor classification task** (c1, c2) and the **modality matching task** (d1, d2); note that the activation maps are based only on the **DCI images** (a1, a2), while the corresponding **FFOCT images** (b1, b2) are showed for a better apprehension of the fiber architecture of the respective FOV.

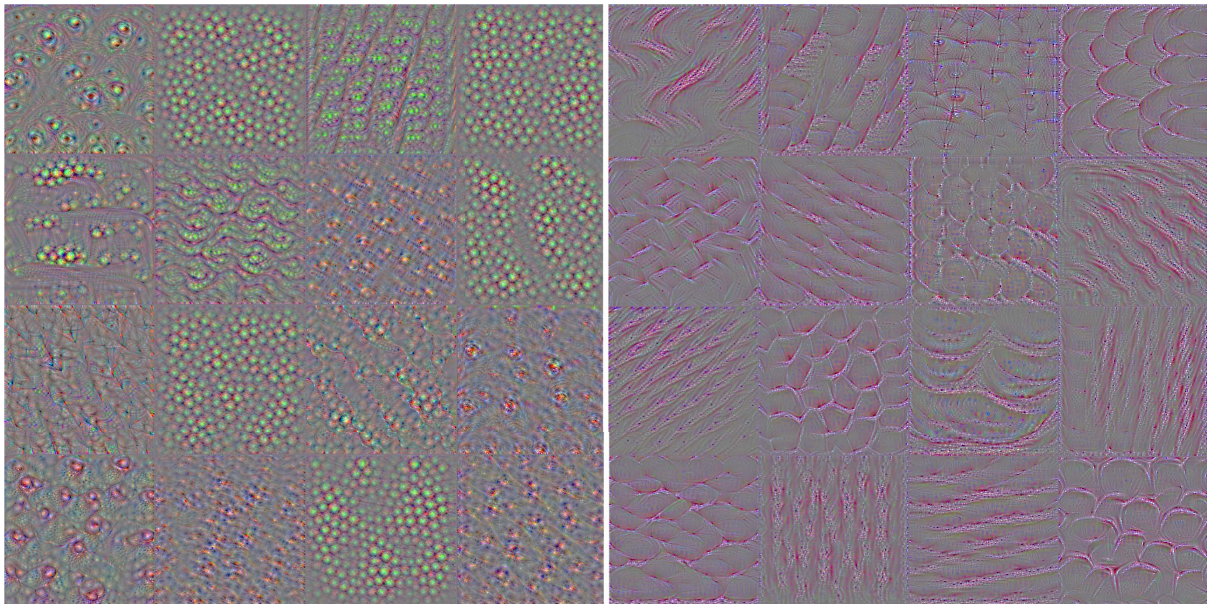


Figure VII.5: Comparison of **learned filters** from the last convolutional layers of VGG-16 model trained on the **tumor classification task** (left) and the **modality matching task** (right) showing **cell** (left) and **fiber** (right) characteristics.

images. We confirm the strength of the approach with multiple experiments both qualitative and quantitative and, together with previous work (e.g. Section V.3), we can firmly state that we managed to develop two complementary models that offer a *complete characterization* of DCI images of breast tissue.

Moreover, this cross-modal matching approach has the advantage of being easily applied to any similar dataset, and an unstructured combination of different datasets thereof, ultimately leading to an exhaustive definition of fiber appearance in DCI which could answer some theoretical questions via a data-driven approach. This effort could lead to finally finding a fiber signature in DCI which could potentially be achieved jointly with the source separation approach in Section V.2. One possible application of having a fiber signature would allow DCI raw signal filtering and therefore better image contrast. Or, a simpler, more brute-force method for improving image readability could be *image multiplexing* by combining relevant structures from each imaging (cells from DCI and fibers from FFOCT) guided by the localization maps illustrated in Figure VII.4, for example.

Another great extension potential lies in the area of improving classification tasks. The features obtained from cross-modal matching are not enough to be used for diagnosis on their own, since we obtained similar performance as when training directly on FFOCT images. We have also experimented with the same architecture prototype but trained with a multitask loss, with the idea to leverage the multimodal pairing as a regularizer, a helper intermediate task along with the main classification task. However, this multitask setting did not surpass the classical supervised task alone (presented in Chapter V) but we believe this could be improved by a different information fusion strategy. In the same philosophy, we get inspiration from [250] where they use a complementary contrastive loss in a MIL setting.

Ultimately, a similar approach of contrastive learning could be used for DCI and classical histology mutual information mining. We remind the reader about the similarities between DCI and standard histology depicted in Figure III.4, however, this development is currently hindered by the lack of an an appropriate dataset i.e. adequately registered intra-modality images. As H&E is the current state-of-the-art in tissue diagnosis, transferable knowledge between FFOCT / DCI and classic H&E histology could greatly speed up the disambiguation and adoption of our imaging modalities.

Chapter VIII

Conclusions

Contents

VIII.1 Summary of Contributions

In the present work we laid the groundwork for FFOCT / DCI image analysis towards cancer diagnosis with focus on a better understanding of the signal. This manuscripts gives a variety of methodological "keys" towards unlocking this objective. We gave a snapshot of all aspects involved in the problematic: from the medical concerns, to the challenges related to the optics and particularities of the data, together with the strengths and limitations of the technical tools available, the modeling choices being driven by the intersection of those aspects. In this regard, we have employed adapted data mining strategies and built 3 datasets, based on which we developed multiple algorithms by leveraging domain knowledge and machine learning methods:

- a proof-of-concept for testing the feasibility of using DL methods on our imaging applied on a *skin cancer dataset*, a collection of widefield FFOCT images coming from 40 skin samples, with pixel-level annotations; for this well posed problem, we design a custom CNN architecture with improved performance from SOTA architectures with 95% sensitivity and 97% specificity in classifying normal dermis *vs.* basal cell carcinoma.

- multiple exploratory methods based on the dully curated *breast surgical margins dataset*, comprised of ~ 400 individually annotated ROIs imaged with both DCI and FFOCT, coming from 47 surgical excisions:

  - a signal decomposition method based on source separation with non-negative matrix factorization revealing oscillatory signatures and their spatial localization;
  - a fully supervised cancer classification model trained on DCI processed images by fine-tuning a narrow bottleneck adaptation of VGG16 architecture offering 97% sensitivity and 85% specificity at the sample level, surpassing the pathologist performance;
  - instance localization of tumors and healthy structures via positive and negative attention maps supported by the straightforward design of the previous model; based on the obtained localization maps we extend the classification model to also accommodate segmentation, this streamlined architecture should serve for easy deployment onto the product side;
  - class-wise filter bases that encode salient textures for healthy lobules and proliferating tumor cells, by replacing the MLP classifier with a linear classifier, which confirms the strong discriminative power of the learned feature extractor alone;
  - compelling evidence towards considering enlarged nucleoli as a (breast) cancer biomarker in DCI imaging, achieved by "looking under the hood" of trained CNNs;
  - robust fiber characterization in DCI images achieved via a multi-modal contrastive learning method by minimizing the cosine distance between corresponding DCI and FFOCT images in the joint latent space learned, which bypasses an important drawback of DCI - the low repeatability, by relying on the robustness of FFOCT.

- a diagnosis method suitable for real-world data acquisition scenarios based on a *breast biopsies dataset* acquired by a radiologist in the clinical setting, without any tailored expert annotations; the dataset is comprised of widefield DCI images of 150 biopsies coming from 72 breast nodules and weak annotations at nodule-level are extracted directly from the pathology reports; the trainset is built via an extensive data engineering pipeline: texture-aware sampling with the *SoSleek* method (which has wide applicability and therefore we made available the open-source code), followed by non-expert agent labeling and ranking based on information content (i.e. entropy) of image patches; we trained a versatile instance-level multiple instance learning model obtaining 89% sensitivity and 84% specificity, followed by thorough post-hoc analysis revealing interesting insights; the framework is easily adoptable at large scale as it minimizes the need of a human annotator, moreover, its design allows obtaining instance-level predictions and, by extension, incorporating dense labels in the training.

To sum up, we propose solutions to accommodate multiple training settings of increased difficulty, from well cured data aiming to answer a precise question to more unstructured data issued from real-world use of the LightCT™ scanner. There is not a one size fits all solution, but we show how we can, and

should, transfer knowledge between tasks. Moreover, given the unknowns regarding both the imaging and the DL framework, we confer special attention to method validation and results analysis in order to deliver explainable models that can be trusted by the clinical users, but also reveal new insights like finding specific biomarkers.

VIII.2 Discussion & Perspectives

Tuning the Model Performance

In the present work we did not focus on improving model performance at the decimal point precision, as a high metric, e.g. accuracy, does not guarantee generalization, instead, we focused more on building unbiased models by learning appropriate features. However, once the training data is sufficiently representative for a final application, which in turn is well-defined in terms of use cases, and the respective model is ready to be moved to production, then the focus can be shifted toward tuning the performance. For classification objectives, the standard strategy (which was also employed here) is thresholding the output probabilities at 50% to determine the final class, however, the decision threshold can be moved to tune the balance between sensitivity and specificity to better fit the problem. In relation with the cost of errors, a higher specificity might be prioritized for screening purposes, while a high sensitivity is crucial for diagnosis. Following the same philosophy, different aggregation strategies can be adopted, to pass from patch-level to sample-level diagnosis.

Another way for improving model performance, at the cost of losing interpretability, used especially in industrial applications and data competitions is ensemble learning. It consists of training multiple models on the same task and obtain the final prediction as a combinations of each their predictions, i.e. voting. As guidance, in [177] they use a set of 50 similarly trained models, with the only difference being the initialization of the weights.

Anyhow, we believe that the aspect of tuning the performance should be dictated by the business plan and is out of the scope of this work.

Deployment

Another obstacle caused by the lack of clearly defined use cases is model deployment. There are two main questions regarding deployment, *what* and *how*. In term of whats, the main challenge of data-driven approaches is that there is rarely a "one size fits all" solution. In our case, integrating a universal diagnosis model into the scanner that could cater to all possible use cases is far fetched, as the scanner can be used with various organs, pathologies and biopsying procedures which differ greatly among themselves to be treated collectively. On the other hand, such embedded solutions may be integrated for answering more general questions, like assessing image quality, detecting cell presence, etc.

In term of hows, the most elegant solution and ultimate objective is embedding the model into the system, most probable at the level of the associated image viewer. The prediction of the model could be presented in textual form as a report and/or as a layer superposed on the acquired image for localization, e.g. class attention heatmaps. Another ambitious option is decoupling the data and algorithms from the actual physical system via a dedicated cloud storage and computing platform.

Nevertheless, we believe that in order to deploy robust trustworthy clinical-grade models, they should be first trained of more data, ideally multi-centric. In order to cater to this need of data collection as well as make the algorithms available to the clinicians, a viable solution is using already existing tools, like Cytomine [30], a collaborative platform destined for storage, annotation and, more recently, algorithm deployment. It enables communication between the clinician - the one who performs the acquisition, the pathologist - the one who establishes the final diagnostic and the developer / data scientist - the one who builds the image analysis pipelines. This way, medical experts can annotate the images they acquire, but also interact with the algorithm's predictions in order to correct or refine them, information which would help improve the aid-to-diagnosis model, either offline or even online.

Here we skimmed some ideas, but a thorough analysis of MLops solutions should be conducted to make deployment decisions.

Ethics and Regulations

When releasing a product to the end user it comes up against regulations enforced by dedicated organism, even more for medical related products. For the European Union, in April 2021, the European Commission publishes the first-ever legal framework on AI and a new Coordinated Plan with Member States. For the USA, in September 2021, FDA issued the "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan". Both policymakers express their concerns on the opacity of AI algorithms in general and advocate for the promotion of the transparency of these devices. Moreover, human-in-the-loop strategies are encouraged, in the sense that no AI device should make decisions, but rather guide the human agent through predictions accompanied by a confidence grade.

It is well known that AI models will encode the biases present in their training data, in the medical cohorts those biases are most often related to ethnicity and gender [251, 252] under-representation. To take a practical example, ethnic groups can be differentiated by the melanin content which is a light absorbing compound. Intuitively, the quantity of melanin might influence OCT imaging as it captures the optical properties of the samples. Indeed, in [253] they find significant difference in quantitative OCT image quality between light and dark skin types at both epidermal and dermal levels, while in [254] they find differences in OCT retinal images in patients with albinism.

The clear trend towards passing to Web3.0 and Blockchain technologies need not be ignored, as it shall impact data ownership status in clinical trials. Medical data, especially images, could be coined as NFTs to be tracked, owned and even monetized. It would be possible to see exactly where the data is used and the patient could consent on the use of their data. This would also ensure greater clinical trial transparency and also ease the enforcement of regulations.

Therefore, data collection should include a diverse cohort and model design should handle known biases as much as possible. When it comes to the methods in this work, we have forestalled the ethical concerns of AI by going the extra mile to design interpretable models.

Multimodality

Moving away from pragmatic aspects towards new methodological directions to extend the present work, we first propose the development of an NLP method to automatically extract clinical information from the pathology reports [255]. To recall the work in Chapter VI, extensive effort has been employed to extract the ground truth from textual data and it shall be repeated on new data, so an automated text parser would reduce the workload on the human agent but also reduce the error, making the extraction reproducible. Ideally, this NLP model should be developed with an expert supervision. We believe that the multiple instance learning approach developed in Chapter VI has already potential to make automated diagnosis in FFOCT / DCI imaging a reality and this moment will only be expedited by an automated ground truth extractor.

In the same endeavor, we think it might be pertinent to achieve an automated correlation with gold standard histology, in order to transfer knowledge from this extensively studied technique and leverage existing multi-centric annotated histology datasets. We have already presented a method to define a common latent space from different imaging techniques in Chapter VII. There we were dealing with already registered DCI and FFOCT images, but in the case of histology and DCI correlation, the main difficulty is obtaining an (even partially) matching dataset. However, we believe that as long as the same structures are visible is both modalities, a similar approach can be applied to learn a common DCI / H&E image representation, therefore it might be worthwhile obtaining an appropriate dataset.

Metabolic Analysis

There are still a lot of questions about the underlying biological mechanism producing the DCI signal, the first step towards uncovering this was reported in [25] where they observed that blocking the aerobic respiration made no change in the acquired signal, however, blocking glycolysis (i.e. the anaerobic respiration) caused the DCI signal to drop to the level of dead-tissue, therefore, they conclude that glycolysis related micro-movements must be at the source of the DCI signal.

Cellular respiration is a metabolic pathway that breaks down glucose (sugar) in order to produce energy in the form of adenosine triphosphate (ATP). The steps of cellular respiration are: first 1) *glycolysis*: breaking down 1 mole of glucose into pyruvate and 2 ATP; followed by 2a) *oxidative phosphorylation* (OXPHOS): in the presence of oxygen, pyruvate is oxidized inducing an electro-chemical potential at the membrane of mitochondria that produces 34 ATP, this is the aerobic respiration; or 2b) in the absence of oxygen the pyruvate is fermented into lactate which can be transformed by the liver back into glucose to fuel glycolysis. However, in our case, i.e. ex-vivo, there is no way to continuously synthesize glucose, therefore, the signal is diminishing with the consumption of the glucose storage, which finally leads to apoptosis, i.e. cell death.

Normal cells utilize glucose to derive 70% of their required ATP through OXPHOS. However, despite the fact that the aerobic metabolic pathway is up to 15 times more efficient than anaerobic metabolism (34 ATP *vs* 2 ATP), it has been found that cancer cells prefer glycolysis even in the presence of sufficient oxygen supply. This phenomenon, discovered more than 50 years ago, is known as the *Warburg effect* or aerobic glycolysis and it is commonly recognized as a hallmark of cancer. What is more, since cancer cells are significantly energy-consuming in order to fuel their malignant properties, like excessive proliferation they perform glycolysis at a 10-fold rate as compared to the norm. Cancer metabolism is being extensively studied for potential anti-cancer therapeutics development [256].

In the light of these literature findings and motivated by the framework presented in Section V.2 meant for fine signal characterization by means of isolating salient oscillatory frequencies proper to cells only, we believe that further experimentation in the cancer cell biology and metabolomics field could shine a light on the quantification of the metabolic activity from DCI signal it presumably captures and, ultimately, it could represent a new quantifiable cancer biomarker.

To conclude, the present work is pioneering for FFOCT / DCI image analysis by offering educated and thoroughly validated solutions, but also by opening up some research directions in the field. We are looking forward to the adoption of FFOCT / DCI imaging as a standard of care for fast tissue assessment and we believe this work significantly contributes to this endeavor.

Résumé francophone détaillé

I Introduction

Motivation

Le cancer est un tueur silencieux, qu'il est crucial de détecter de façon précoce pour augmenter les chances de guérison et de survie. L'imagerie bio-médicale est particulièrement utile dans la prise en charge du cancer, pour le dépistage, le diagnostic et le traitement. Malgré les progrès apportés en imagerie, la recherche reste très active afin d'aider à améliorer le suivi, la qualité de vie et la survie du patient.

On propose l'utilisation d'une nouvelle famille de techniques d'imagerie, la tomographie par cohérence optique plein champ statique (FFOCT) et dynamique (DCI), qui permettent une analyse nettement plus rapide du tissu par rapport à l'étalon-or qui est l'histopathologie. De cette manière, elles peuvent améliorer la prise en charge du patient à la fois pour la biopsie (en réduisant le temps d'attente d'un résultat négatif et ainsi l'angoisse du patient) aussi bien que pour la chirurgie (en optimisant la quantité du tissu excisé car pour réduire la récidive il faut enlever toutes les cellules cancéreuses, tout en préservant les cellules saines). Reposant sur le principe de cohérence lumineuse, ils ont l'avantage d'obtenir un bon contraste sans aucun préparation du tissu ainsi que la possibilité d'imager l'échantillon du tissu en profondeur. Cependant, leur nouveauté et leur contraste différent des autres techniques les rendent difficiles à adopter en milieu clinique.

Afin de faciliter l'interprétation par les médecins des images FFOCT et DCI obtenues, on emploie des approches exploratoires sur plusieurs fronts: essayer d'abord de caractériser le signal interférométrique dynamique brut, fournir des méthodes d'aide au diagnostic basées sur l'apprentissage profond sur plusieurs jeux de données et essayer finalement de décoder ces modèles de boîte noire pour obtenir une interprétation du diagnostic, et finalement trouver des biomarqueurs spécifiques dans nos images.

La structure de la thèse

Dans le chapitre introductif on établit le contexte du présent travail, notamment en rappelant l'évolution des techniques d'imagerie impliquées dans le traitement du cancer. On donne également un bref aperçu de l'évolution de l'informatique et de l'intelligence artificielle menant à des méthodes d'aide au diagnostic.

Dans le Chapitre II on approfondit les techniques d'imagerie étudiées, la tomographie par cohérence optique plein champ (FFOCT) et l'imagerie cellulaire dynamique (DCI). On présente les éléments théoriques nécessaires à la compréhension des mécanismes qui caractérisent les deux imageries, en partant d'une brève introduction à l'optique et la façon dont la lumière interagit avec la matière biologique, jusqu'à la description des spécifications techniques, comme les composants du scanner ou la formation des images.

Le Chapitre III est consacré aux données. On décrit ici les étapes de traitement pour obtenir les 3 jeux de données conçus et exploités dans cette thèse. On présente une contribution originale, représentée par une méthode d'échantillonnage d'images tenant compte de la texture, qui est applicable à un large spectre de problèmes d'imagerie. Ici on identifie aussi les défis du présent travail liés à l'imagerie et à la nature des données, qui déterminent les choix méthodologiques dans les chapitres suivants.

Le Chapitre IV présente quelques notions indispensables à la compréhension de l'apprentissage profond, avec un accent particulier sur la conception, l'entraînement et la validation des réseaux de neurones convolutifs (CNN), qui reposent au cœur de la méthodologie exploitée dans cette thèse.

Le Chapitre V est un riche chapitre exploratoire couvrant de multiples aspects de l'imagerie FFOCT / DCI, réunis par l'objectif commun de la classification entièrement supervisée des échantillons de tissus sains *vs.* cancéreux. On utilise de multiples techniques d'extraction de caractéristiques sur les images FFOCT, le signal brut DCI et les images traitées DCI, ainsi qu'une variété de stratégies de classification telles que: la construction d'une architecture CNN, l'entraînement de classifieurs arborescents sur des caractéristiques à séparation de sources, le peaufinage d'une architecture de pointe, ainsi que le décodage des caractéristiques apprises, etc.

Le Chapitre VI aborde un problème du monde réel, celui de l'apprentissage à partir de données acquises dans un contexte clinique sans traitement particulier ni annotations d'experts, la vérité terrain étant extraite de rapports de pathologie déjà disponibles. À cet égard, nous élaborons un pipeline de classifications de biopsies malignes ou bénignes en utilisant une approche d'apprentissage multi-instances; le modèle bénéficie d'une définition transparente qui permet également d'accéder au diagnostic inféré des sous-parties de l'image.

Dans le Chapitre VII nous exploitons désormais la dualité de l'imagerie FFOCT / DCI, ici nous adoptons une approche d'apprentissage contrastif pour maîtriser les artefacts de la DCI avec la robustesse de la FFOCT. Nous développons une méthode fiable de caractérisation des fibres à partir de l'imagerie DCI en utilisant les images FFOCT correspondantes comme point de repère, en minimisant la distance cosinus entre les paires d'images correspondantes dans un espace latent commun appris via un réseau de neurones siamois. De plus, nous accordons une attention particulière à la validation, en définissant deux métriques, les erreurs d'identité et de symétrie.

Dans le dernier chapitre, nous résumons les contributions du présent travail et proposons quelques idées pour les développements futurs.

II Tomographie par cohérence optique plein champ statique et dynamique

La tomographie par cohérence optique plein champ [23] (FFOCT de *Full-Field Optical Coherence Tomogrphy*), développée [1] et perfectionnée [23] par l'équipe du Pr. Claude Boccara de l'ESPCI et commercialisée par LLTech depuis 2011, est utilisée dans le domaine médical et de la recherche pour analyser la morphologie et la fonction des tissus biologiques. Elle est particulièrement utile car elle ne nécessite aucune préparation du tissu (par exemple, une coloration) et offre une résolution[1] de $\approx 1\,\mu\mathrm{m}$ dans les 3 dimensions. Dans une étude clinique récente [24], deux chirurgiens ont obtenu une sensibilité et une spécificité moyennes de 87% pour le diagnostic de la malignité du sein sur des images FFOCT. De plus, la tomographie par cohérence optique plein champ dynamique [25] également connue sous le nom de **imagerie cellulaire dynamique** (DCI de *Dynamic Cell Imaging*) fait progresser la technique en révélant des informations complémentaires liées aux structures cellulaires vivantes grâce au contraste endogène dérivé de l'activité cellulaire, ce qui permet d'améliorer la sensibilité de 90% et la spécificité de 96%, selon la même étude [24].

III Des données cliniques aux données computationnelles

Les jeux de données

Dans le présent travail, nous traitons trois ensembles de données différents. Il est intéressant de noter que l'ordre de présentation des jeux de données correspond à l'ordre chronologique d'acquisition, ce qui donne également un aperçu de l'évolution de la technique d'imagerie: de la FFOCT plus "historique", au passage à la DCI quand il n'était pas possible d'imager que certains champs de vue (FOV de *Field of*

[1] Pour donner une idée de l'ordre de résolution, la taille des organites (par exemple les mitochondries) varie de 1 à $10\,\mu\mathrm{m}$, alors que les noyaux mesurent environ $10\,\mu\mathrm{m}$, les cellules animales mesurent de l'ordre de dizaines de microns jusqu'à 100 μm.

View) au début, puis finalement à l'acquisition de biopsies entières en DCI. De plus, la granularité des labels associés à chaque ensemble de données, depuis les annotations au niveau du pixel jusqu'au niveau du patient, indique une tendance vers la diminution du niveau de supervision lorsque la complexité et l'ampleur des ensembles de données augmentent:

- **cancer de la peau - excisions de la chirurgie de Mohs**: 40 images FFOCT haute résolution (10 contenant du carcinome basocellulaire), annotées au niveau des pixels, sous-échantillonnées en utilisant le suréchantillonnage de la classe minoritaire pour garantir un équilibre entre les classes, ce qui donne environ $50\,000$ patchs pour chaque catégorie.

- **cancer du sein - excisions des mastectomies**: 400 FOVs unitaires DCI distribués aléatoirement dans 47 échantillons (11 normaux et 36 cancéreux) provenant de 33 patientes ayant subi des mastectomies, annotés individuellement.

- **cancer du sein - biopsies**: un ensemble de biopsies imagées en DCI par le radiologue qui les a aussi réalisées, parmi lesquelles ont été conservées pour l'entraînement 150 biopsies provenant de 72 nodules mammaires, ne comportant pas d'annotations mais avec une vérité terrain extraite des rapports de pathologie (issus des lames H&E), sous-échantillonnées en tuiles respectant la texture à l'aide de la méthode SoSleek (github.com/dmandache/sleek-patch).

Ces ensembles de données, résumés dans le tableau suivant, serviront à construire des pipelines d'analyse de données dans le but d'extraire des informations précieuses sur cette technique d'imagerie révolutionnaire en vue d'améliorer la pratique clinique.

Tableau: Aperçu des jeux de données.

étude clinique	**modalité d'imagerie**	**échelle des images**	**niveau d'annotation**	**taille du patch**	**no. patchs**	**no. annotations**
cancer de la peau	FFOCT	gigapixel	pixel	256	100K	100K
cancer du sein marges chirurgicales	FFOCT DCI	megapixel (patch)	patch	1440	400	400
cancer du sein biopsies	FFOCT DCI	gigapixel	groupe d'images	1024	2K	150

Défis liés aux données

Même si nous sommes toujours dans le contexte du big data, nous sommes confrontés au défi des **données limitées**, omniprésentes dans le domaine médical, en outre, il existe de multiples inconnues liées à l'aspect novateur de la technique et à la nature même de l'imagerie:

- **données provenant d'un seul centre**: pour chaque application les données sont collectées dans le cadre d'études cliniques ciblées; les données proviennent d'un seul centre et par conséquent, par rapport à l'ensemble des données médicales qui alimentent les méthodes de pointe, nous sommes d'autant plus confrontés à la rareté des données;

- **bruit d'étiquetage**: à l'heure actuelle, aucun expert médical n'est en mesure de poser un diagnostic fiable sur la seule base des images FFOCT / DCI; l'annotation des images résulte soit de la collaboration entre un expert en imagerie et un expert médical (un pathologiste) par une mise en correspondance entre nos images et la lame d'histologie, soit d'une extraction directe du rapport de pathologie basé sur le même échantillon. Par conséquent, la procédure d'annotation risque d'introduire un certain bruit dans les étiquettes en raison de l'éventuelle différence de composition des tissus visible sur la lame d'histologie découpée mécaniquement, par opposition au découpage optique FFOCT / DCI sans préparation;

- **artéfacts d'image**: la nature dynamique de la DCI donne lieu à des aberrations d'image liées à des instabilités physiques (vibrations externes) qui peuvent entraver l'interprétation correcte des images, de plus, ce bruit est stochastique par nature et on ne sait pas à ce stade comment le modéliser et par conséquent le filtrer;

- **phénomènes biologiques captés non définies**: la source du signal DCI n'est pas encore totalement caractérisée, ce qui implique que les critères d'apparence des cellules ne sont pas validés biologiquement; par exemple, on ne sait toujours pas si ce que nous percevons comme "cellule" correspond au noyau ou à la cellule entière, c'est-à-dire le cytoplasme, ou comment l'intensité mesurée à l'intérieur du volume cellulaire, censée être représentative de l'activité cellulaire, pourrait être corrélée aux processus biologiques.

Compte tenu des points énumérés, il est clair que les approches axées sur les données (plutôt que sur les modèles) sont adaptées à toutes les inconnues sous-jacentes. Nous allons donc nous tourner vers la puissante famille d'algorithmes d'apprentissage automatique dédiée à la vision par ordinateur pour construire des solutions efficaces d'aide au diagnostic.

IV Principes théoriques sur les réseaux neuronaux convolutifs

L'apprentissage profond (DL de *Deep Learning*) [115, 116], qui est considéré comme la solution idéale pour les problèmes difficiles à définir, modélise les relations entre les entrées et les sorties d'un système sans beaucoup d'informations sur ce système lui-même. L'apprentissage profond trouve ses racines en 1957 avec l'invention du **perceptron** [45], qui servait de classifieur linéaire, mais qui est en fait le prototype du neurone artificiel moderne. L'apprentissage profond est synonyme de réseaux neuronaux profonds, qui représentent un modèle de calcul d'inspiration biologique constitué de couches interconnectées de neurones artificiels. La sortie d'un neurone représente l'entrée d'un ou plusieurs neurones de la couche suivante sans connexion entre les neurones d'une même couche, un réseau neuronal est donc un graphe acyclique. Un neurone peut servir de classifieur faible en soi, et un réseau peut donc être considéré comme une combinaison de classifieurs faibles qui forment un classifieur plus fort. L'intelligence d'un réseau réside dans les pondérations de ses neurones, qui représentent également ses paramètres de réglage. Pour parvenir à résoudre correctement une tâche, le réseau ajuste ses paramètres par le mécanisme d'apprentissage basé sur le principe du *essai et erreur*. Cet algorithme puissant est la **rétropropagation** [46], elle consiste à répartir l'erreur de prédiction entre tous les neurones interconnectés d'un circuit, proportionnellement à leur contribution à l'erreur. L'erreur entre la vraie solution attendue y et la solution prédite $\hat{y}$ peut être calculée à l'aide d'une variété de fonctions, appelées **fonctions de perte**. Le choix dépend fortement de l'objectif du problème. Pour la classification, la fonction d'entropie croisée, aussi appelée perte logarithmique, est la plus utilisée. Elle mesure la différence entre la distribution de probabilité prédite par un modèle de classification par apprentissage automatique et la vraie distribution: $L(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$.

Un modèle parfait aurait une perte nulle $L = 0$, en conséquence la rétropropagation est utilisée pour ajuster les poids afin de minimiser la perte. Mathématiquement, cela implique de calculer le gradient (ou la dérivée) de la fonction de perte par rapport aux poids d'une multicouche de neurones $\nabla_W L(W)$ en appliquant *la loi de la dérivation en chaîne*: pour $y = f(h(x))$, $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial h} \frac{\partial h}{\partial x}$. Ensuite, avec la dérivée de la perte calculée par rapport à chaque poids du réseau, les poids sont ajustés dans la direction négative du gradient modulé par η *la taille du pas*: $w_i \leftarrow w_i - \eta \frac{\partial L}{\partial w_i}$, soit **descente de gradient**. Il existe plusieurs **algorithmes d'optimisation** [124] utilisés pour mettre en œuvre différentes stratégies de descente de gradient. Dans le cas de la descente de gradient stochastique (SGD de *Stochastic Gradient Descent*) nous remplaçons le vecteur de gradient réel par une estimation stochastique du vecteur de gradient. Pour un réseau de neurones, l'estimation stochastique signifie le gradient de la perte pour une seule instance ou, plus souvent, pour un sous-ensemble. La SGD momentum [125], similaire au principe d'inertie en physique, permet d'amortir les oscillations causées par les solutions partielles calculées avec SGD. La SGD avec momentum est actuellement la méthode d'optimisation de pointe pour de

nombreux problèmes d'apprentissage profond. Il existe egalement d'autres méthodes, généralement appelées méthodes adaptatives qui sont particulièrement utiles pour les problèmes mal conditionnés. L'estimation adaptative du moment (Adam) [126] repose sur l'idée d'adapter η pour chaque paramètre pendant l'apprentissage, c'est-à-dire l'augmenter si la descente a une direction constante et le diminuer lorsque la direction change. Adam conserve une moyenne mobile exponentielle du gradient passés similaire au momentum et une moyenne mobile exponentielle du gradient au carré.

Les réseaux de neurones convolutifs (CNN de *Convolutional Neural Networks*) sont une classe d'architectures conçues pour les entrées matricielles, notamment les images. Leur configuration de neurones s'inspire de l'organisation du cortex visuel, chaque neurone répondant à une petite zone du champ de vision appelée champ récepteur. Dans le contexte des réseaux neuronaux, cette organisation est mise en œuvre par le biais de connexions parcimonieuses, ce qui signifie que chaque neurone est connecté à un petit nombre de neurones (par exemple une région carrée de 3×3) de la couche précédente. En outre, les poids sont partagés entre les nœuds adjacents, une propriété qui rend le CNN invariant à la translation. Les CNN sont un exemple réussi d'intégration d'informations spécifiques à un domaine (dans ce cas, pour des tâches de vision) dans la conception de l'architecture d'un réseau neuronal.

Un aspect essentiel de la conception des CNN consiste à établir les connexions entre les couches qui définissent le flux d'informations dans le réseau. Les anciens réseaux *feed-forward* à chemin unique partagent les mêmes principes de conception, c'est-à-dire qu'ils alternent les couches convolutionnelles avec les couches de regroupements, suivi des couches entièrement connectées, mais les nouveaux modèles gagnent en complexité pour s'adapter aux ensembles de données plus riches, comme ImageNet [48] - un ensemble de 1,3 million d'images naturelles haute résolution appartenant à 1000 classes. Pour énumérer quelques architectures importantes: *LeNet* [47] - la première, *AlexNet* [50] et *VGG* [131] - les pionniers des problèmes complexes, *Inception* [132] introduit la croissance en largeur, plutôt qu'en profondeur, *ResNet* [133] introduit des connexions de saut, etc. D'autres approches sont: les architectures multi-flux prenant plusieurs entrées (par exemple, les réseaux siamois [135]), les architectures codeur/décodeur pour la segmentation d'images (par exemple, *Unet* [136], *MaskRCNN* [130]), ou les réseaux adversatifs génératifs (GAN) [137] pour la synthèse d'images, etc.

Étant donné leur nature de boîte noire, il est difficile de comprendre les décisions qu'une CNN prend pour obtenir un certain résultat. Par conséquent, la seule façon d'avoir une idée du raisonnement est d'employer des méthodes de validation qualitative, comme la visualisation et l'évaluation des filtres appris ou des cartes d'attention. *Erhan et al.* [141] propose une méthode qui permet de visualiser les motifs appris qui sont encodés par les noyaux convolutifs, en effectuant une ascension de gradient dans l'espace d'entrée en maximisant l'activation du filtre examiné. Grad-CAM [142] est une méthode qui

permet de mettre en évidence dans une certaine image d'entrée la zone d'intérêt qui contribue le plus à la prédiction d'une certaine classe, sous la forme d'une carte de chaleur.

V Classification sain *vs.* malin par apprentissage entièrement supervisé

V.1 Classification à partir des images FFOCT

Nous présentons d'abord une application qui représente une preuve de concept sur la faisabilité de l'utilisation d'une aide au diagnostic automatisée pour notre imagerie unique. Elle correspond à la détection du carcinome basocellulaire, un sous-type de cancer de la peau, qui est homogène et visible pour être discriminé du tissu normal, notamment en FFOCT, alors c'est un problème plutôt bien posé.

Nous avons commencé par expérimenter avec certaines architectures populaires comme VGG-16 [131] ou InceptionV3 [132] qui alimentent déjà de nombreuses applications d'imagerie dans divers domaines. Ces architectures pré-entraînées ont offert une précision de 89.30% et 90.79%, respectivement. Résultats que nous jugeons insatisfaisant étant donné la formulation réductrice du problème (annotations très denses et à haut degré de confiance, input de petite taille) et améliorons les résultats en construisant une architecture de réseau sur mesure. Nous avons construit un CNN moins profond avec 10 couches comprenant: la partie extraction de caractéristiques, composée de 4 paires de couches convolutives suivis d'un max-pooling avec 25% de dropout et un classifieur composé de deux couches entièrement connectées de 512 et 64 neurones, respectivement, avec 50% de dropout. Dans la conception, nous avons pris en considération la taille des champs réceptifs les mieux adaptés à notre problème. Par exemple, le premier a 7×7 (par opposition à 3×3 pour les autres architectures) et le dernier 90×92 (par opposition à 212×212 pour le VGG16). On obtient une précision de classification de 95,93%, correspondant à une sensibilité de 95,2% et à une spécificité de 96,54% au niveau du patch, cela prouve que lorsque les données le permettent, la construction d'architectures adaptées peut se révéler bénéfique.

V.2 Classification à partir du signal DCI

Pour rester en phase avec l'évolution de la technique, nous portons notre attention sur la DCI. Afin de sonder l'importance du signal métabolique révélé par l'imagerie, nous allons dissocier de la structure du tissu en prenant en compte seulement les profils dynamiques trouvés dans chaque FOV. Nous les utiliserons ensuite comme caractéristiques à partir desquelles classer les tissus cancéreux et normaux en explorant plusieurs modèles basés sur des arbres de décision. La motivation est d'isoler différentes structures dans le domaine dynamique car de multiples comportements devraient être présents: signaux combinés de l'échantillon et des perturbations, multiples types de diffuseurs dans le tissu, multiples sources de signal dans un pixel (par exemple, superposition de fibres et de cellules). Par conséquent, nous

avons utilisé une approche de séparation des sources à l'aveugle, la méthode de factorisation par matrices non négatives (NMF de *Non-negative Matrix Factorization*) [146, 147] pour ses résultats hautement interprétables en vertu de sa contrainte de positivité conduisant à une décomposition par parties.

Le but de la NMF est de factoriser une matrice de données $X \in \mathbb{R}^{n \times d}$ en deux matrices positives de rang bas $H \in \mathbb{R}^{k \times d}$ et $W \in \mathbb{R}^{n \times k}$ représentant la base de caractéristiques extraites et son activation correspondante: $X \approx W \cdot H$, où n est le nombre de points de données, d la dimension de chaque point de données et k le nombre de composantes choisies dans lesquelles diviser. La recherche des deux matrices de composition est réalisée en minimisant l'erreur entre la matrice de données originale et le résultat de la factorisation: $\min_{W \geq 0, H \geq 0} \|X - W \cdot H\|_F^2$. Un des inconvénients de la NMF est le choix empirique du du rang k, sur la base d'expériences nous avons choisi $k = 5$ et les composantes révélées semble correspondre à:

- **signal de ligne de base**: niveau de bruit du signal;
- **fibres**: composantes spectrales de haute magnitude avec des structures fibreuses dans la composante spatiale correspondante;
- **erreur** d'échantillonnage: un pic au dernier bin de fréquence correspond à l'énergie à la fréquence de Nyquist;
- **cellules**: formes de cellules révélées dans la localisation spatiale;
- **artefacts** de mouvement: la composante de fréquence bruyante avec des pics dans la partie supérieure du spectre et l'activation spatiale correspond aux fibres hautement réfléchissantes sont des indicateurs d'une modulation de phase du signal DCI induite par un mouvement externe.

Nous avons utilisé la décomposition NMF, plus précisément les composantes dynamiques, comme moyen d'extraction de caractéristiques pour classifier les FOVs cancéreux et normaux avec une précision de 71% et nous avons révélé certaines fréquences plus importantes. Pourtant, sur la base de ces résultats, nous concluons que le signal dynamique seul (traitant donc le scanner plus comme une sonde que comme un dispositif d'imagerie) ne suffit pas pour le diagnostic. Néanmoins, cette méthode pourrait être utilisée pour améliorer le contraste de l'image ou analyser le signal de manière plus quantitative.

V.3 Classification à partir des images DCI

Sur la base des idées précédentes, on se tourne vers une solution DL pour classer les images DCI traitées provenant du tissu mammaire. Mais cette fois-ci, par rapport à l'application de la peau, nous sommes confrontés à un organe, une pathologie et une modalité d'imagerie plus difficiles, ainsi qu'à des annotations à plus petite échelle.

Pour surmonter ces inconvénients, nous nous appuyons sur un modèle pré-entraîné. En [153] ils ont révélé que l'apprentissage par transfert a connu une montée en force, montrant son efficacité sur de

petits ensembles de données. Par conséquent, nous avons décidé d'utiliser ImageNet comme base de données de pré-entraînement car elle s'est avérée plus efficace à environ 15% par rapport à d'autres jeux de données spécifiques dans le cas des images histologique du cancer du sein [154]. En ce qui concerne l'architecture, c'est une fois de plus la simplicité qui gagne puisque la VGG16 [131] obtient les meilleures performances parmi les plusieurs architectures testées. Notre CNN adopte les couches convolutionnelles du VGG16 mais emploie une stratégie différente de regroupement des informations pour la construction du vecteur de caractéristiques, le pooling à moyenne globale (GAP de *Global Average Pooling*). Le GAP permet au réseau d'agir comme un modèle d'apprentissage par dictionnaire, d'accepter n'importe quelle taille d'entrée d'image et de réduire le nombre de paramètres du classifieur à seulement 4% du nombre total.

Cette approche encourage l'apprentissage d'une base de caractéristiques fortement discriminantes. Cela, associée à des méthodes de validation quantitative et de désambiguïsation de CNNs, conduit à la construction d'un dictionnaire linéairement séparable de textures spécifiques à l'organisation des cellules normales ou malignes, et aussi à découvrir les nucléoles hypertrophiés comme biomarqueurs du cancer dans le DCI.

En ce qui concerne les résultats quantitatifs, on obtient une précision par FOV de $89 \pm 4\%$, ce qui correspond à une sensibilité de $88 \pm 4\%$ et une spécificité de $86 \pm 6\%$ pour une validation croisée à 5 plis. De plus, au niveau de l'échantillon, l'algorithme surpasse la performance moyenne de deux pathologistes lors d'un diagnostic en aveugle.

VI Classification bénin *vs.* malin par apprentissage faiblement supervisé

Les données médicales sont chères à obtenir, les annotations encore plus, c'est pourquoi nous avons développé une méthode qui apprend à partir des rapports de pathologie déjà disponibles. Dans le présent travail, nous construisons un outil de diagnostic en utilisant uniquement les étiquettes globales grâce au paradigme d'apprentissage par instances multiples (MIL de *Multiple Instance Learning*).

Dans le contexte du MIL, les instances sont représentées par plusieurs sous-images (patchs) échantillonnées à partir d'une image plus grande qui représenterait un groupe d'instances. Le concept a été introduit par [167] en réponse au besoin de formaliser des connaissances incomplètes ou de classifier des groupes hétérogènes. Ces groupes mixtes (ou sacs) sont généralement définis de manière binaire en partant du principe que les sacs négatifs ne contiennent que des instances négatives tandis que les sacs positifs peuvent également contenir des instances négatives et qu'ils doivent contenir au moins une instance positive, mais que la nature des instances est inconnue.

Étant donné notre objectif de rendre les images DCI interprétables par les utilisateurs cliniques, nous choisissons une variante de MIL qui permet d'obtenir également des prédictions locales. De plus, nous sommes préoccupés par l'apprentissage de caractéristiques adaptées à l'imagerie dans la perspective de découvrir des biomarqueurs propres à notre contraste unique, raison pour laquelle nous choisissons d'entraîner l'extracteur de caractéristiques à cette tâche également.

Ces aspects se distinguent des méthodes existantes, car la majorité des méthodes MIL appliquées aux images de pathologie [177,178,181,182] se concentrent sur différentes stratégies de mises en relation entre les instances dans l'espace latent, puis sur l'entraînement d'un classifieur. L'intégration de l'extraction de caractéristiques dans le formalisme du MIL est rarement prise en compte. Ceci est principalement dû à la limitation des ressources computationnelles demandées par de telles architectures multi-branches. Cependant, nous contournons ce problème en effectuant une présélection des instances en définissant une certaine hiérarchie et aussi en transférant les poids du modèle entraînés sur un domaine similaire et en affinant uniquement les couches spécifiques à la tâche. En outre, on intègre le classifieur au niveau de l'instance afin d'obtenir également une prédiction pertinente de la malignité au niveau local.

Nous construisons une architecture multi-branches dont la branche constitutive est entièrement transférée (couches convolutionnelles et classifieur entièrement connecté) du modèle CNN précédemment entraîné avec supervision complète sur le jeu de données d'excision du sein et dupliqué k fois, où k est le nombre d'instances par sac. Tous les poids sont partagés entre les branches, en plus de cela on ajoute une couche de regroupement *Max* qui agrège les prédictions au niveau de l'instance sous l'hypothèse standard MIL pour donner la prédiction au niveau du sac. Pour chaque échantillon $X = \{x_i | i = \overline{1, N}\}$ contenant N tuiles x_i, $i = \overline{1, N}$ dont la vérité terrain globale y est connue, nous construisons le sac correspondant en sélectionnant un sous-ensemble $X_b \subseteq X$. La branche principale encode une fonction $\phi : X \mapsto [0, 1]$ faisant correspondre chaque tuile x_i à sa probabilité prédite $\hat{y}_i = \phi(x_i)$, qui sont ensuite agrégées par la couche de regroupement MIL pour obtenir la prédiction de l'échantillon $\hat{y} = \max_{i \in \{1,k\}}(\hat{y}_i)$. Nous insistons sur le fait que les étiquettes de tuiles y_i sont inconnues, mais que nous pouvons tout de même obtenir des prédictions de tuiles $\hat{y}_i$.

Pour traiter les données déséquilibrées, nous utilisons la fonction de perte focale [189] lors de l'apprentissage. Une considération importante est de définir une capacité maximale des sacs k_{max}, ce qui nous amène à créer une hiérarchie de patchs et à choisir en priorité les patchs contenant des cellules et ayant un contenu d'information plus élevé. Nous choisissons $k_{max} = 8$ en fonction du nombre moyen de patchs contenant des cellules par échantillon (soit 6) et des ressources hardware disponibles. Cependant, au moment de l'inférence et lors du calcul des métriques, nous prenons en compte tous les patchs.

Nous obtenons une augmentation remarquable de 32 points en sensibilité (89%) par rapport au modèle utilisant la branche principale pré-entraînée sur ImageNet et nous arrivons à égaler la spécificité obtenue dans l'application entièrement supervisée (84%).

Sur ces bases, nous énonçons les contributions suivantes de notre approche MIL pour la classification des biopsies mammaires malignes *vs* bénignes dans DCI:

1. on obtient des prédictions adéquates au niveau de la biopsie;
2. on déduit des prédictions adéquates au niveau des patchs;
3. on apprend une base de caractéristiques adaptée à la tâche;
4. on construit un modèle mutable qui peut facilement être étendu à d'autres problèmes.

VII Apprentissage de la représentation intermodale FFOCT *vs.* DCI

Dans ce chapitre, notre principale motivation est d'exploiter la nature multimodale du système optique. À cet égard, nous entraînons explicitement un réseau siamois sur des paires d'images recalées des deux modalités, ce qui conduit à définir un encodage inter-modal commun, effort qui entre dans le cadre de la représentation des connaissances. Cela permettrait d'extraire des informations mutuelles de la DCI et de la FFOCT qui encoderaient très probablement les caractéristiques des fibres. Ceci est particulièrement intéressant car les fibres en FFOCT ont un rendu reproductible entre les acquisitions, alors qu'en DCI le tissu fibreux souffre de grandes fluctuations au niveau de son apparence, tout en étant la principale source d'artefacts d'imagerie. Nous prévoyons de surmonter les inconvénients de la DCI avec la robustesse de la FFOCT.

Le but d'un réseau siamois est d'apprendre une fonction d'encodage commune Φ qui met en correspondance deux représentations différentes $x_i^a \in \mathcal{I}^a$ et $x_i^b \in \mathcal{I}^b$ de la même instance $s_i \in \mathcal{S}$, mais avec $x_i^a \neq x_i^b$ sur un domaine commun $\mathcal{F}$ tel que $f_i^a = \Phi(x_i^a),\ f_i^a \in \mathcal{F}$ et $f_i^b = \Phi(x_i^b),\ f_i^b \in \mathcal{F}$ et leur distance Δ dans ce domaine est proche de zéro, $\Delta(f_i^a, f_i^b) \to 0$. Par extension, pour deux échantillons différents s_i et s_j, la distance devrait être supérieure $\Delta(f_i^a, f_j^b) \gg 0$. Dans la plupart des applications présentes dans la littérature, x^b est une version transformée de x^a tel que $x^b = T(x^a)$ où T est un opérateur linéaire de la famille de l'augmentation des données, ce qui implique que les domaines d'entrée sont très similaire $\mathcal{I}^a \sim \mathcal{I}^b$. Cependant, dans notre cas, sur chaque échantillon de tissu s_i sont appliquées deux projections d'imagerie très différentes, l'une correspondant à la modalité FFOCT $\Omega^a : \mathcal{S} \to \mathcal{I}^a$ et l'autre à la DCI $\Omega^b : \mathcal{S} \to \mathcal{I}^b$, ce qui rend les deux domaines d'entrée $\mathcal{I}^a$, $\mathcal{I}^b$ plus difficiles à rapprocher dans le domaine d'intégration commun $\mathcal{F}$. Intuitivement, cette divergence entre les deux domaines pousse le réseau à apprendre et à encoder des caractéristiques plus significatives des deux domaines, en cas de convergence.

La fonction Φ est codée par un modèle VGG16 appliqué sur les deux entrées x^a, x^b. L'architecture est choisie pour garder une cohérence avec les travaux antérieurs, mais nous trouvons que, contrairement aux problèmes résolus précédemment, le présent problème est moins sensible à l'architecture de réseau choisie, des autres modèles (ResNet50 et InceptionV3) ayant des performances comparables.

La distance Δ est définie comme la distance cosinus, une mesure[2] basée sur le cosinus de l'angle θ entre deux vecteurs, qui est à son tour une dimension de la similarité: $\Delta(f^a, f^b) = 1 - cos(\theta(f^a, f^b) = 1 - \frac{f^a \cdot f^b}{\|f^a\| \cdot \|f^b\|}$

Le processus d'apprentissage est facilité par la génération en ligne de batchs. Pour les paires d'images positives, on choisit au hasard une position et on extrait les patchs correspondants de taille 480×480 des images DCI et FFOCT (x_i^a, x_i^b), tandis que pour les paires de patchs négatifs (x_i^a, x_j^b), on choisit au hasard deux images non correspondantes et on sélectionne une position aléatoire dans chacune. Pour respecter le principe d'équilibre des classes, nous formons des batchs symétriques avec le même nombre de paires négatives et positives.

Les résultats montrent qu'il n'y a pas de confusion entre les paires d'images positives et négatives sur les données de test, de plus, nous obtenons une marge dure entre la distance maximale des paires positives $\max_i(\Delta(f_i^a, f_i^b))$ et la distance minimale des paires négatives $\min_{i \neq j}(\Delta(f_i^a, f_j^b))$. Étant donné que nous considérons ce problème sous l'angle de l'apprentissage d'une métrique, nous étudions les propriétés d'identité et de symétrie en définissant des fonctions d'erreur adaptées.

Grâce à la validation qualitative, nous en déduisons que même les fibres à faible contraste sont capturées par le réseau et que les artefacts d'imagerie sont bien compris par le réseau (et non pas confondus avec les fibres) et ignorés, deux épreuves habituellement difficiles pour un agent humain non formé.

Cette approche a l'avantage de se prêter à l'application à tout ensemble de données similaires, ainsi qu'à une combinaison de différents ensembles de données, pour arriver finalement à une définition exhaustive de l'apparence des fibres en DCI.

VIII Conclusions

Synthèse des contributions

Dans ce travail, nous avons posé les bases de l'analyse des images FFOCT / DCI vers le diagnostic du cancer en nous concentrant sur une meilleure compréhension du signal. À cet égard, nous avons employé des stratégies de préparation de données adaptées et nous avons construit 3 ensembles de données, grâce auxquels nous avons développé de multiples algorithmes par apprentissage automatique:

[2] N.B. La distance cosinus n'est pas vraiment une métrique car elle ne satisfait pas l'inégalité triangulaire.

- une preuve de concept pour tester la faisabilité du DL sur notre imagerie appliquée sur un ensemble de données du *cancer de la peau* - une collection d'images FFOCT à large champ provenant de 40 échantillons, avec des annotations au niveau du pixel. Pour ce problème bien défini, nous avons conçu une architecture CNN sur mesure avec des performances améliorées par rapport aux architectures SOTA avec 95% de sensibilité et 97% de spécificité pour la classification du derme normal *vs.* carcinome basocellulaire.

- plusieurs approches exploratoires basées sur un ensemble de données des *marges chirurgicales du sein*, composé de 400 champs imagés à la fois par DCI et FFOCT et annotés individuellement, provenant de 47 excisions chirurgicales:

  - une méthode de décomposition du signal basée sur la séparation des sources - factorisation en matrices non négatives - révélant les signatures oscillatoires et leur localisation spatiale;
  - un modèle de classification du cancer entraîné sur des images DCI par l'apprentissage fin d'une adaptation de l'architecture VGG16 avec un goulot d'étranglement (*bottleneck*) réduit, offrant une sensibilité de 97% et une spécificité de 85% au niveau de l'échantillon, surpassant les performances des pathologistes;
  - la localisation des tumeurs et des structures saines via des cartes d'attention positive et négative, sur lesquelles nous étendons le modèle précèdent pour qu'il prenne également en compte la segmentation, cette architecture intégrée devrait permettre un déploiement facile du côté du produit;
  - l'extraction des bases de filtre par catégorie qui codent les textures saillantes des lobules sains et des cellules tumorales en prolifération, en remplaçant le classifieur MLP par un simple classifieur linéaire, ce qui confirme le fort pouvoir discriminant de l'extracteur de caractéristiques à lui seul;
  - des preuves convaincantes en faveur de la prise en compte de l'hypertrophie des nucléoles comme biomarqueur du cancer (du sein) dans l'imagerie DCI, obtenues par des techniques de désambiguïsation appliquées aux réseaux neuronaux entraînés;
  - une caractérisation fiable des fibres dans les images DCI, obtenue par une méthode d'apprentissage contrastive multimodale qui minimise la distance cosinus dans l'espace latent conjoint entre les images DCI et FFOCT correspondantes, ce qui permet de contourner un inconvénient important de la DCI - la faible répétabilité - en s'appuyant sur la robustesse de la FFOCT.

- une méthode de diagnostic adaptée aux scénarios d'acquisition de données du milieu clinique basée sur un jeu de données de *biopsies mammaires* composé d'images DCI grand champ de

150 biopsies provenant de 72 nodules mammaires, les diagnostics à l'échelle du nodule sont extraits directement des rapports de pathologie. Le jeu d'entraînement est construit via un pipeline adapté comprenant une nouvelle approche d'échantillonnage tenant compte de la texture - *SoSleek*, suivie d'un étiquetage par un agent non expert et d'un classement des patchs basé sur le contenu en information. Nous avons développé un modèle d'apprentissage à instances multiples ayant une sensibilité de 89% et une spécificité de 84%, puis nous avons effectué une analyse post-hoc des résultats révélant des informations intéressantes. Ce cadre est facilement adoptable à grande échelle car il minimise le besoin d'un annotateur humain. De plus, sa conception permet d'obtenir des prédictions au niveau localisé et, par extension, d'incorporer des étiquettes denses lors de l'apprentissage.

Perspectives

Dans ce travail, nous nous sommes davantage concentrés sur la construction de modèles fiables et non biaisés en apprenant des caractéristiques appropriées plutôt que d'améliorer les performances à la virgule près. Cependant, une fois que le modèle est prêt à être mis en production, l'accent pourrait se mettre sur le réglage des performances: en ajustant le seuil de décision afin de trouver l'équilibre entre sensibilité et spécificité pour mieux s'adapter au coût des erreurs. L'apprentissage ensembliste, consistant à faire voter plusieurs modèles pour les prédictions finales, est un autre moyen d'améliorer les métriques au prix d'une perte d'interprétabilité, utilisé notamment dans les applications industrielles et les compétitions de données. Un autre aspect à prendre en compte pour les prochaines étapes est la solution de déploiement qui doit respecter les directives éthiques et réglementaires imposées par les organismes de régulation.

Pour les recherches futures, nous pensons qu'il est utile de collecter des données adaptées pour effectuer la correspondance modale avec l'histologie classique afin de transférer les connaissances de cette technique largement étudiée et de tirer parti des ensembles de données histologiques annotées multicentriques existants. De plus, nous pensons que des expériences plus poussées dans le domaine de la biologie cellulaire et de la métabolomique du cancer pourraient mettre en lumière la quantification de l'activité métabolique à partir du signal DCI, ce qui permettrait de proposer un nouveau biomarqueur quantifiable du cancer.

En conclusion, le présent travail est pionnier pour l'analyse d'images FFOCT / DCI en offrant des solutions adaptées et soigneusement validées, mais aussi en ouvrant certaines directions de recherche dans le domaine.

Bibliography

[1] E. Beaurepaire, A. C. Boccara, M. Lebec, et al., "Full-field optical coherence microscopy," *Optics Letters*, vol. 23, no. 4, pp. 244, feb 1998.

[2] G. Sines and Y. A. Sakellarakis, "Lenses in Antiquity," *American Journal of Archaeology*, vol. 91, no. 2, pp. 191, apr 1987.

[3] L. Schermelleh, A. Ferrand, T. Huser, et al., "Super-resolution microscopy demystified," *Nature Cell Biology*, vol. 21, no. 1, pp. 72–84, jan 2019.

[4] L. Fass, "Imaging and cancer: A review," *Molecular Oncology*, vol. 2, no. 2, pp. 115–152, 2008.

[5] G. B. Faguet, "A brief history of cancer: Age-old milestones underlying our current knowledge database," *International Journal of Cancer*, vol. 136, no. 9, pp. 2022–2036, 2015.

[6] R. García-Figueiras, S. Baleato-González, A. R. Padhani, et al., "How clinical imaging can assess cancer biology," *Insights into Imaging*, vol. 10, no. 1, 2019.

[7] M. F. Berger and E. R. Mardis, "The emerging clinical relevance of genomics in cancer medicine," *Nature Reviews Clinical Oncology*, vol. 15, no. 6, pp. 353–365, 2018.

[8] Y. Kuang, J. D. Nagy, and S. E. Eikenberry, "Introduction to mathematical oncology," *Introduction to Mathematical Oncology*, vol. 3, pp. 1–453, apr 2016.

[9] R. C. Rockne, A. Hawkins-Daarud, K. R. Swanson, et al., "The 2019 mathematical oncology roadmap," *Physical Biology*, vol. 16, no. 4, jun 2019.

[10] C. Luchini, A. Pea, and A. Scarpa, "Artificial intelligence in oncology: current applications and future perspectives," *British Journal of Cancer*, vol. 126, no. 1, pp. 4–9, 2022.

[11] D. X. Yang, P. R. Soulos, B. Davis, et al., "Impact of Widespread Cervical Cancer Screening," *American Journal of Clinical Oncology: Cancer Clinical Trials*, vol. 41, no. 3, pp. 289–294, 2018.

[12] R. C. Fitzgerald, A. C. Antoniou, L. Fruk, and N. Rosenfeld, "The future of early cancer detection," *Nature Medicine 2022 28:4*, vol. 28, no. 4, pp. 666–677, apr 2022.

[13] G. Barbany, C. Arthur, A. Liedén, et al., "Cell-free tumour DNA testing for early detection of cancer – a potential future tool," *Journal of Internal Medicine*, vol. 286, no. 2, pp. 118–136, aug 2019.

[14] K. P. Pritzker and H. J. Nieminen, "Needle biopsy adequacy in the era of precision medicine and value-based health care," *Archives of Pathology and Laboratory Medicine*, vol. 143, no. 11, pp. 1399–1415, 2019.

[15] N. Cabioglu, K. K. Hunt, A. A. Sahin, et al., "Role for intraoperative margin assessment in patients undergoing breast-conserving surgery," *Annals of Surgical Oncology*, vol. 14, no. 4, pp. 1458–1471, 2007.

[16] J. S. Poling, T. N. Tsangaris, P. Argani, and A. Cimino-Mathews, "Frozen section evaluation of breast carcinoma sentinel lymph nodes: a retrospective review of 1,940 cases," *Breast Cancer Research and Treatment*, vol. 148, no. 2, pp. 355–361, 2014.

[17] B. W. Maloney, D. M. McClatchy, B. W. Pogue, et al., "Review of methods for intraoperative margin detection for breast conserving surgery," *Journal of Biomedical Optics*, vol. 23, no. 10, pp. 1, oct 2018.

[18] J. Heidkamp, M. Scholte, C. Rosman, et al., "Novel imaging techniques for intraoperative margin assessment in surgical oncology: A systematic review," *International Journal of Cancer*, vol. 149, no. 3, pp. 635–645, 2021.

[19] M. Thill, "MarginProbe®: Intraoperative margin assessment during breast conserving surgery by using radiofrequency spectroscopy," *Expert Review of Medical Devices*, vol. 10, no. 3, pp. 301–315, 2013.

[20] M. Ragazzi, S. Piana, C. Longo, et al., "Fluorescence confocal microscopy for pathologists," *Modern Pathology*, vol. 27, no. 3, pp. 460–471, 2014.

[21] W. Jung and S. A. Boppart, "Optical coherence tomography for rapid tissue screening and directed histological sectioning," *Analytical Cellular Pathology*, vol. 35, no. 3, pp. 129–143, 2012.

[22] K. Fujii, R. Kawakami, and S. Hirota, "Histopathological validation of optical coherence tomography findings of the coronary arteries," *Journal of Cardiology*, vol. 72, no. 3, pp. 179–185, sep 2018.

[23] A. Dubois, L. Vabre, A. C. Boccara, and E. Beaurepaire, "High-resolution full-field optical coherence tomography with a Linnik microscope," *Applied Optics*, vol. 41, no. 4, pp. 805–812, feb 2002.

[24] H. Yang, S. Zhang, P. Liu, et al., "Use of high-resolution full-field optical coherence tomography and dynamic cell imaging for rapid intraoperative diagnosis during breast cancer surgery," *Cancer*, vol. 126, no. S16, pp. 3847–3856, aug 2020.

[25] C. Apelian, *Imagerie Optique Multimodale des tissus par Tomographie Optique Cohérente Plein Champ*, Ph.D. thesis, 2017.

[26] L. Pantanowitz, A. Sharma, A. B. Carter, et al., "Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives," *Journal of Pathology Informatics*, vol. 9, no. 1, jan 2018.

[27] M. N. Kent, T. G. Olsen, T. A. Feeser, et al., "Diagnostic accuracy of virtual pathology vs traditional microscopy in a large dermatopathology study," *JAMA Dermatology*, vol. 153, no. 12, pp. 1285–1291, dec 2017.

[28] M. D. Herrmann, D. A. Clunie, A. Fedorov, et al., "Implementing the DICOM standard for digital pathology," *Journal of Pathology Informatics*, vol. 9, no. 1, pp. 37, jan 2018.

[29] P. Bankhead, M. B. Loughrey, J. A. Fernández, et al., "QuPath: Open source software for digital pathology image analysis," *Scientific Reports*, vol. 7, no. 1, pp. 1–7, dec 2017.

[30] R. Marée, L. Rollus, B. Stévens, et al., "Collaborative analysis of multi-gigapixel imaging data using Cytomine," *Bioinformatics*, vol. 32, no. 9, pp. 1395–1401, may 2016.

[31] A. Goode, B. Gilbert, J. Harkes, et al., "OpenSlide: A vendor-neutral software foundation for digital pathology," *Journal of Pathology Informatics*, vol. 4, no. 1, pp. 27, jan 2013.

[32] E. Abels, L. Pantanowitz, F. Aeffner, et al., "Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association," *Journal of Pathology*, vol. 249, no. 3, pp. 286–294, 2019.

[33] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, "Digital pathology and artificial intelligence," *The Lancet Oncology*, vol. 20, no. 5, pp. 253–261, 2019.

[34] A. M. Turing, "Computing machinery and intelligence," *Machine Intelligence: Perspectives on the Computational Model*, vol. LIX, no. 236, pp. 1–28, oct 2012.

[35] J. Von Neumann, "The Computer and the Brain," *Philosophical Studies*, vol. 8, no. 0, pp. 182–185, 1958.

[36] M. Gardner, "Mathematical Games - The fantastic combinations of John Conway's new solitaire game "life"," *Scientific American*, vol. 220, no. 1, 1969.

[37] J. M. Springer and G. T. Kenyon, "It's Hard for Neural Networks to Learn the Game of Life," in *Proceedings of the International Joint Conference on Neural Networks*, sep 2021, pp. 1–8.

[38] E. H. Shortliffe, A. C. Scott, M. B. Bischoff, et al., "Oncocin: an Expert System for Oncology Protocol Management.," in *International Joint Conference on Artificial Intelligence*, 1981, vol. 2, pp. 876–881.

[39] T. M. Mitchell, *Machine Learning textbook*, 1997.

[40] S. B. Kotsiantis, "Decision trees: A recent overview," *Artificial Intelligence Review*, vol. 39, no. 4, pp. 261–283, apr 2013.

[41] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[42] C. Cortes, V. Vapnik, and L. Saitta, "Support-Vector Networks Editor," Tech. Rep., 1995.

[43] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[44] D. H. Hubel and T. N. Wiesel, *Brain and Visual Perception: The Story of a 25-year Collaboration*, 2012.

[45] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[46] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, 1986.

[47] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, 1998.

[48] J. Deng, W. Dong, R. Socher, et al., "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[49] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, vol. 32, 2012.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, 2017.

[51] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, pp. 411–418.

[52] S. Perincheri, A. W. Levi, R. Celli, et al., "An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy," *Modern Pathology*, vol. 34, no. 8, pp. 1588–1595, 2021.

[53] M. S. Zubairy, "A very brief history of light," in *Optics in Our Time*, pp. 3–24. Springer, Cham, jan 2016.

[54] S. L. Jacques, "Fractal nature of light scattering in tissues," *Journal of Innovative Optical Health Sciences*, vol. 4, no. 1, pp. 1–7, jan 2011.

[55] S. Jacques and S. Prahl, "Lecture notes in introduction to biomedical optics," 2002.

[56] A. H. Hielscher, A. A. Eick, D. Shen, et al., "Mechanisms of light scattering from biological cells relevant to noninvasive optical-tissue diagnostics," vol. 37, no. 16, jun 1998.

[57] J. R. Mourant, M. Canpolat, C. Brocker, et al., "Light scattering from cells: the contribution of the nucleus and the effects of proliferative status," *Journal of Biomedical Optics*, vol. 5, no. 2, pp. 131, 2000.

[58] J. D. Wilson, *Measurements and Interpretations of Light Scattering From Intact Biological Cells*, Ph.D. thesis, 2007.

[59] D. Huang, E. A. Swanson, C. P. Lin, et al., "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[60] J. G. Fujimoto, C. Pitris, S. A. Boppart, and M. E. Brezinski, "Optical coherence tomography: An emerging technology for biomedical imaging and optical biopsy," *Neoplasia*, vol. 2, no. 1-2, pp. 9–25, jan 2000.

[61] C. Apelian, *Imagerie Optique Multimodale des tissus par Tomographie Optique Cohérente Plein Champ*, Ph.D. thesis, Université Paris sciences et lettres, nov 2017.

[62] C. Apelian, F. Harms, O. Thouvenin, and A. C. Boccara, "Dynamic full field optical coherence tomography: subcellular metabolic contrast revealed in tissues by interferometric signals temporal analysis," *Biomedical Optics Express*, vol. 7, no. 4, pp. 1511, apr 2016.

[63] P. J. Thul and C. Lindskog, "The human protein atlas: A spatial map of the human proteome," *Protein Science*, vol. 27, no. 1, pp. 233–244, 2018.

[64] R. McLendon, A. Friedman, D. Bigner, et al., "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061–1068, 2008.

[65] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, et al., "1399 HE-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset," *GigaScience*, vol. 7, no. 6, pp. 1–8, jun 2018.

[66] F. Xu, C. Zhu, W. Tang, et al., "Predicting Axillary Lymph Node Metastasis in Early Breast Cancer Using Deep Learning on Primary Tumor Biopsy Slides," *Frontiers in Oncology*, vol. 11, 2021.

[67] M. Veta, Y. J. Heng, N. Stathonikos, et al., "Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge," *Medical Image Analysis*, vol. 54, pp. 111–121, jul 2019.

[68] G. Aresta, T. Araújo, S. Kwok, et al., "BACH: Grand Challenge on Breast Cancer Histology Images," *Medical Image Analysis*, vol. 56, pp. 122–139, aug 2018.

[69] B. H. Menze, A. Jakab, S. Bauer, et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, oct 2015.

[70] K. Sirinukunwattana, J. P. Pluim, H. Chen, et al., "Gland segmentation in colon histology images: The GLAS challenge contest," *Medical Image Analysis*, vol. 35, pp. 489–502, jan 2017.

[71] S. Berg, D. Kutra, T. Kroeger, et al., "ilastik: Interactive Machine Learning for (Bio)Image Analysis," *Nature Methods*, vol. 16, no. 12, pp. 1226–1232, sep 2019.

[72] F. De Chaumont, S. Dallongeville, N. Chenouard, et al., "Icy: An open bioimage informatics platform for extended reproducible research," *Nature Methods*, vol. 9, no. 7, pp. 690–696, jun 2012.

[73] Y. Matsuda, T. Fujii, T. Suzuki, et al., "Comparison of fixation methods for preservation of morphology, RNAs, and proteins from paraffin-embedded human cancer cell-implanted mouse models," *Journal of Histochemistry and Cytochemistry*, vol. 59, no. 1, pp. 68–75, jan 2011.

[74] P. R. Roelfsema and R. Houtkamp, "Incremental grouping of image elements in vision," *Attention, Perception Psychophysics*, vol. 73, no. 8, pp. 2542, nov 2011.

[75] R. Achanta, A. Shaji, K. Smith, et al., "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2281, jun 2012.

[76] R. Achanta, A. Shaji, K. Smith, et al., "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2281, 2012.

[77] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary region segmentation of objects in N-D images," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 105–112, 2001.

[78] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, may 2002.

[79] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[80] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision 2004 59:2*, vol. 59, no. 2, pp. 167–181, sep 2004.

[81] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100, 1979.

[82] V. Anklin, P. Pati, G. Jaume, et al., "Learning Whole-Slide Segmentation from Inexact and Incomplete Labels Using Tissue Graphs," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 636–646, 2021.

[83] J. Huang and R. Li, "Fast Regions-of-Interest Detection in Whole Slide Histopathology Images," *Pathology - From Classics to Innovations*, nov 2021.

[84] S. Fouad, D. Randell, A. Galton, et al., "Unsupervised Superpixel-Based Segmentation of Histopathological Images with Consensus Clustering," *Communications in Computer and Information Science*, vol. 723, pp. 767–779, 2017.

[85] B. E. Bejnordi, G. Litjens, M. Hermsen, et al., "A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images," *Medical Imaging 2015: Digital Pathology*, vol. 9420, pp. 94200H, mar 2015.

[86] S. Sornapudi, R. J. Stanley, W. V. Stoecker, et al., "Deep Learning Nuclei Detection in Digitized Histology Images by Superpixels," *Journal of Pathology Informatics*, vol. 9, no. 1, jan 2018.

[87] B. Irving, "maskSLIC: Regional Superpixel Generation with Application to Local Pathology Characterisation in Medical Images," *arXiv*, jun 2016.

[88] N. Otsu, "Threshold Selection Method From Gray-Level Histograms.," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-9, no. 1, pp. 62–66, 1979.

[89] T. W. Ridler and S. Calvard, "Picture Thresholding Using an Iterative Slection Method.," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-8, no. 8, pp. 630–632, 1978.

[90] G. W. Zack, W. E. Rogers, and S. A. Latt, "Automatic measurement of sister chromatid exchange frequency," *Journal of Histochemistry and Cytochemistry*, vol. 25, no. 7, pp. 741–753, jan 1977.

[91] C. H. Li and P. K. Tam, "An iterative algorithm for minimum cross entropy thresholding," *Pattern Recognition Letters*, vol. 19, no. 8, pp. 771–776, jun 1998.

[92] M. Roser and H. Ritchie, "Cancer," *Our World in Data*, 2015, https://ourworldindata.org/cancer.

[93] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, dec 2019.

[94] H. He and Y. Ma, "Imbalanced learning: Foundations, algorithms, and applications," *Imbalanced Learning: Foundations, Algorithms, and Applications*, pp. 1–210, jan 2013.

[95] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Computing Surveys*, vol. 49, no. 2, aug 2016.

[96] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority oversampling technique," Tech. Rep., 2002.

[97] E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *International Statistical Review / Revue Internationale de Statistique*, vol. 57, no. 3, 1989.

[98] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[99] R. Liu, L. O. Hall, K. W. Bowyer, et al., "Synthetic minority image over-sampling technique: How to improve AUC for glioblastoma patient survival prediction," *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017*, vol. 2017-Janua, pp. 1357–1362, nov 2017.

[100] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data 2019 6:1*, vol. 6, no. 1, pp. 1–48, jul 2019.

[101] American Cancer Society, "Cancer Facts and Figures 2021," 2021.

[102] S. Z. Aasi, D. J. Leffell, and R. Z. Lazova, *Atlas of Practical Mohs Histopathology*, 2013.

[103] G. Sauer, H. Deissler, K. Strunz, et al., "Ultrasound-guided large-core needle biopsies of breast lesions: Analysis of 962 cases to determine the number of samples for reliable tumour classification," *British Journal of Cancer*, vol. 92, no. 2, pp. 231–235, jan 2005.

[104] H. Y. Ji, E. K. Kim, J. K. Min, et al., "Missed breast cancers at us-guided core needle biopsy: How to reduce them," *Radiographics*, vol. 27, no. 1, pp. 79–94, 2007.

[105] G. K. Malhotra, X. Zhao, H. Band, and V. Band, "Histological, molecular and functional subtypes of breast cancers," *Cancer Biology and Therapy*, vol. 10, no. 10, pp. 955–960, nov 2010.

[106] L. Solorzano, G. M. Almeida, B. Mesquita, et al., "Whole Slide Image Registration for the Study of Tumor Heterogeneity," Tech. Rep., 2018.

[107] E. V. Lang, K. S. Berbaum, and S. K. Lutgendorf, "Large-core breast biopsy: Abnormal salivary cortisol profiles associated with uncertainty of diagnosis," *Radiology*, vol. 250, no. 3, pp. 631–637, mar 2009.

[108] S. Krishnamurthy, A. Contreras, C. T. Albarracin, et al., "Breast pathology," in *Oncological Surgical Pathology*, A. Sapino and J. Kulka, Eds., pp. 921–1047. Springer International Publishing, 1 edition, 2020.

[109] B. Weigelt, F. C. Geyer, and J. S. Reis-Filho, "Histological types of breast cancer: How special are they?," *Molecular Oncology*, vol. 4, no. 3, pp. 192–208, 2010.

[110] J. Makki, "Diversity of breast carcinoma: Histological subtypes and clinical relevance," *Clinical Medicine Insights: Pathology*, vol. 8, no. 1, pp. 23–31, 2015.

[111] A. C. Bateman, "Pathology of benign breast disease," *Women's Health Medicine*, vol. 3, no. 1, pp. 6–8, jan 2006.

[112] E. W. Elston and I. O. Ellis, "Method for grading breast cancer.," *Journal of Clinical Pathology*, vol. 46, no. 2, pp. 189, 1993.

[113] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep Learning for Whole Slide Image Analysis: An Overview," *Frontiers in Medicine*, vol. 6, nov 2019.

[114] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[115] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[116] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 2016.

[117] L. S. Fei-Fei, R. N. Fergus, and A. M. Torralba, "A Short Course Matlab Code: Recognizing and Learning Object Categories," in *IEEE International Conference on Computer Vision*, 2009.

[118] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157.

[119] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[120] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005, vol. I, pp. 886–893.

[121] I. Austvoll, "Filter banks, wavelets, and frames with applications in computer vision and image processing (a review) invited," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2749, pp. 907–921, 2003.

[122] S. Russell and P. Norvig, "Artificial Intelligence A Modern Approach (4th Edition)," 2021.

[123] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, no. 4, pp. 303–314, 1989.

[124] S. Ruder, "An overview of gradient descent optimization algorithms," Tech. Rep., 2016.

[125] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Networks*, vol. 12, no. 1, pp. 145–151, 1999.

[126] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, dec 2015.

[127] A. C. Wilson, R. Roelofs, M. Stern, et al., "The marginal value of adaptive gradient methods in machine learning," in *Advances in Neural Information Processing Systems*. may 2017, vol. 2017-Decem, pp. 4149–4159, Neural information processing systems foundation.

[128] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *33rd International Conference on Machine Learning, ICML 2016*, 2016, vol. 3, pp. 1868–1877.

[129] A. Araujo, W. Norris, and J. Sim, "Computing Receptive Fields of Convolutional Neural Networks," *Distill*, vol. 4, no. 11, pp. e21, nov 2019.

[130] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv*, 2017.

[131] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[132] C. Szegedy, V. Vanhoucke, S. Ioffe, et al., "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 2818–2826.

[133] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[134] J. Hu, L. Shen, S. Albanie, et al., "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, sep 2017.

[135] D. Chicco, "Siamese Neural Networks: An Overview," *Methods in Molecular Biology*, vol. 2190, pp. 73–94, 2021.

[136] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[137] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 3, no. January, pp. 2672–2680, 2014.

[138] M. Power, G. Fell, and M. Wright, "Principles for high-quality, high-value testing," *Evidence-Based Medicine*, vol. 18, no. 1, pp. 5–10, 2013.

[139] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *International Joint Conference of Artificial Intelligence*, 1995.

[140] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8689 LNCS, pp. 818–833, 2014.

[141] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Bernoulli*, , no. 1341, pp. 1–13, 2009.

[142] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[143] A. Esteva, B. Kuprel, R. A. Novoa, et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature 2017 542:7639*, vol. 542, no. 7639, pp. 115–118, jan 2017.

[144] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Journal of Machine Learning Research*, 2010, vol. 9, pp. 249–256.

[145] S. Mannor, B. Peleg, and R. Rubinstein, "The cross entropy method for classification," *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pp. 561–568, 2005.

[146] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[147] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[148] A. Cichocki, R. Zdunek, A. H. Phan, and S. I. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*, John Wiley and Sons, oct 2009.

[149] A. Vahadane, T. Peng, A. Sethi, et al., "Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[150] R. Maruyama, K. Maeda, H. Moroda, et al., "Detecting cells using non-negative matrix factorization on calcium imaging data," *Neural Networks*, vol. 55, pp. 11–19, 2014.

[151] S. Ullman, *High-level vision: object recognition and visual cognition*, vol. 34, MIT Press, 1997.

[152] Y. Freund and R. E. Schapire, "A Short Introduction to Boosting," *Society*, vol. 14, no. 5, pp. 771–780, 2009.

[153] H. Chahal, H. Toner, and I. Rahkovsky, "Small Data's Big AI Potential," Tech. Rep. September, Center for Security and Emerging Technology, sep 2021.

[154] M. Kohl, C. Walz, F. Ludwig, et al., "Assessment of Breast Cancer Histology Using Densely Connected Convolutional Networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 903–913, 2018.

[155] M. L. Richter, J. Schoning, A. Wiedenroth, and U. Krumnack, "Should You Go Deeper? Optimizing Convolutional Neural Network Architectures without Training," *Proceedings - 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021*, pp. 964–971, jun 2021.

[156] G. Apou, N. S. Schaadt, B. Naegel, et al., "Detection of lobular structures in normal breast tissue," *Computers in Biology and Medicine*, vol. 74, pp. 91–102, 2016.

[157] B. Helpap, "Nucleolar grading of breast cancer - Comparative studies on frequency and localization of nucleoli and histology, stage, hormonal receptor status and lectin histochemistry," *Virchows Archiv A Pathological Anatomy and Histopathology*, vol. 415, no. 6, pp. 501–508, nov 1989.

[158] J. E. Quin, J. R. Devlin, D. Cameron, et al., "Targeting the nucleolus for cancer intervention," *Biochimica et Biophysica Acta - Molecular Basis of Disease*, vol. 1842, no. 6, pp. 802–816, jun 2014.

[159] J. Yan and D. Tang, "The nucleolar aspect of breast cancer," in *Proteins of the Nucleolus: Regulation, Translocation, Biomedical Functions*, pp. 275–304. Springer Netherlands, sep 2012.

[160] M. Derenzini, L. Montanaro, and D. Treré, "What the nucleolus says to a tumour pathologist," *Histopathology*, vol. 54, no. 6, pp. 753–762, may 2009.

[161] R. C. Pezo and R. H. Singer, "Nuclear microenvironments in cancer diagnosis and treatment," in *Journal of Cellular Biochemistry*. aug 2008, vol. 104, pp. 1953–1963, J Cell Biochem.

[162] H. Lee, K. Chong, D. Giron, et al., "Automated image based prominent nucleoli detection," *Journal of Pathology Informatics*, vol. 6, no. 1, pp. 39, 2015.

[163] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10541 LNCS, pp. 379–387, jun 2017.

[164] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari, "Cost-sensitive learning with noisy labels," *Journal of Machine Learning Research*, vol. 18, pp. 1–33, 2018.

[165] B. E. Bejnordi, M. Veta, P. J. Van Diest, et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA - Journal of the American Medical Association*, vol. 318, no. 22, pp. 2199–2210, dec 2017.

[166] J. Keeler, D. Rumelhart, and W. Leow, "Integrated Segmentation and Recognition of Hand-Printed Numerals," *Advances in Neural Information Processing Systems*, vol. 3, 1991.

[167] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[168] Y. LeCun, C. Cortes, and C. J. Burges, "MNIST handwritten digit database," 1998.

[169] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., University of Toronto, 2009.

[170] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," *Caltech mimeo*, vol. 11, no. 1, pp. 20, 2007.

[171] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-Instance Learning for Medical Image and Video Analysis," *IEEE Reviews in Biomedical Engineering*, vol. 10, pp. 213–234, 2017.

[172] Y. Liu, K. Gadepalli, M. Norouzi, et al., "Detecting Cancer Metastases on Gigapixel Pathology Images," *arXiv*, 2017.

[173] M. A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, may 2018.

[174] O. Maron and T. Lozano-Pérez, "A Framework for Multiple-Instance Learning," *Advances in Neural Information Processing Systems*, vol. 10, 1997.

[175] T. Durand, N. Thome, and M. Cord, "WELDON: Weakly supervised learning of deep convolutional neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 4743–4752.

[176] M. Sun, T. X. Han, Ming-Chang Liu, and A. Khodayari-Rostamabad, "Multiple Instance Learning Convolutional Neural Networks for object recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. dec 2016, vol. 0, pp. 3270–3275, IEEE.

[177] P. Courtiol, E. W. Tramel, M. Sanselme, and G. Wainrib, "Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach," *arXiv*, feb 2018.

[178] O. Dehaene, A. Camara, O. Moindrot, et al., "Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology," *arXiv*, 2020.

[179] K. He, H. Fan, Y. Wu, et al., "Momentum Contrast for Unsupervised Visual Representation Learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735, nov 2020.

[180] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2021.

[181] P. J. Sudharshan, C. Petitjean, F. Spanhol, et al., "Multiple instance learning for histopathological breast cancer image classification," *Expert Systems with Applications*, vol. 117, pp. 103–111, 2019.

[182] M. Y. Lu, D. F. Williamson, T. Y. Chen, et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[183] P. Tourniaire, M. Ilie, P. Hofman, et al., "Attention-based Multiple Instance Learning with Mixed Supervision on the Camelyon16 Dataset," *MICCAI Computational Pathology (COMPAY) Workshop*, vol. 156, pp. 216–226, 2021.

[184] J. Foulds and E. Frank, "A Review of Multi-Instance Learning Assumptions," *The Knowledge Engineering Review*, vol. 25, no. 1, pp. 1–25, 2010.

[185] P. Courtiol, C. Maussion, M. Moarii, et al., "Deep learning-based classification of mesothelioma improves prediction of patient outcome," *Nature Medicine*, vol. 25, no. 10, pp. 1519–1525, oct 2019.

[186] D. Pathak, P. Krahenbuhl, J. Donahue, et al., "Context Encoders: Feature Learning by Inpainting," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, nov 2016, vol. 2016-Decem, pp. 2536–2544.

[187] H. D. Couture, J. S. Marron, C. M. Perou, et al., "Multiple instance learning for heterogeneous images: Training a CNN for histopathology," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11071 LNCS, pp. 254–262.

[188] S. Li, Y. Zhao, R. Varma, et al., "PyTorch Distributed: Experiences on Accelerating Data Parallel Training," Tech. Rep., 2020.

[189] T. Y. Lin, P. Goyal, R. Girshick, et al., "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, aug 2020.

[190] J. Mukhoti, V. Kulharia, A. Sanyal, et al., "Calibrating deep neural networks using focal loss," *Advances in Neural Information Processing Systems*, vol. 2020-Decem, 2020.

[191] C. M. Focke, T. Decker, and P. J. van Diest, "The reliability of histological grade in breast cancer core needle biopsies depends on biopsy size: a comparative study with subsequent surgical excisions," *Histopathology*, vol. 69, no. 6, pp. 1047–1054, dec 2016.

[192] D. Wang, A. Khosla, R. Gargeya, et al., "Deep Learning for Identifying Metastatic Breast Cancer," *arXiv*, 2016.

[193] M. Lu, Y. Pan, D. Nie, et al., "SMILE : Sparse-Attention based Multiple Instance Contrastive Learning for Glioma Sub-Type Classification Using Pathological Images," *MICCAI Computational Pathology (COMPAY) Workshop*, vol. 1, pp. 1–8, 2021.

[194] A. Pirovano, H. Heuberger, S. Berlemont, et al., "Improving Interpretability for Computer-Aided Diagnosis Tools on Whole Slide Imaging with Multiple Instance Learning and Gradient-Based Explanations," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12446 LNCS, pp. 43–53, 2020.

[195] G. Campanella, V. W. K. Silva, and T. J. Fuchs, "Terabyte-scale Deep Multiple Instance Learning for Classification and Localization in Pathology," *arXiv*, 2018.

[196] M. J. Bissell and D. Radisky, "Putting tumours in context," *Nature Reviews Cancer*, vol. 1, no. 1, pp. 46–54, 2001.

[197] D. Erhan, Y. Bengio, A. Courville, et al., "Why does Unsupervised Pre-training Help Deep Learning?," Tech. Rep., 2009.

[198] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.

[199] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[200] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 1096–1103.

[201] M. Germain, K. Gregor, I. Murray, and H. Larochelle, "MADE: Masked autoencoder for distribution estimation," in *32nd International Conference on Machine Learning, ICML 2015*, 2015, vol. 2, pp. 881–889.

[202] B. Chandra and R. K. Sharma, "Exploring autoencoders for unsupervised feature selection," in *Proceedings of the International Joint Conference on Neural Networks*, 2015, vol. 2015-Septe.

[203] M. Alberti, M. Seuret, R. Ingold, and M. Liwicki, "A Pitfall of Unsupervised Pre-Training," *arXiv*, 2017.

[204] L. Zhou, H. Liu, J. Bae, et al., "Self Pre-training with Masked Autoencoders for Medical Image Analysis," *arXiv*, 2022.

[205] E. B. Asiedu, S. Kornblith, T. Chen, et al., "Decoder Denoising Pretraining for Semantic Segmentation," *arXiv*, 2022.

[206] K. He, X. Chen, S. Xie, et al., "Masked Autoencoders Are Scalable Vision Learners," *arXiv*, 2021.

[207] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*. jun 2017, pp. 5999–6009, Neural information processing systems foundation.

[208] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[209] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 1422–1430, 2015.

[210] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9910 LNCS, pp. 69–84.

[211] A. Taleb, C. Lippert, T. Klein, and M. Nabi, "Self-supervised Learning for Medical Images by Solving Multimodal Jigsaw Puzzles," *Ieee Transactions on Medical Imaging*, vol. 12729 LNCS, no. Xx, pp. 661–673, 2020.

[212] R. Zhang, P. Isola, A. A. Efros, and B. A. Research, "Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction," Tech. Rep., nov 2017.

[213] A. Jaiswal, A. R. Babu, M. Z. Zadeh, et al., "A Survey on Contrastive Self-Supervised Learning," *Technologies*, vol. 9, no. 1, pp. 2, 2020.

[214] X. Liu, F. Zhang, Z. Hou, et al., "Self-supervised Learning: Generative or Contrastive," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[215] M. Zheng, F. Wang, S. You, et al., "Weakly Supervised Contrastive Learning," *arXiv*, 2021.

[216] P. Khosla, P. Teterwak, C. Wang, et al., "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, apr 2020.

[217] B. Kulis, "Metric learning: A survey," *Foundations and Trends in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2012.

[218] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *37th International Conference on Machine Learning, ICML 2020*, vol. PartF16814, pp. 1575–1585, 2020.

[219] X. Chen, H. Fan, R. Girshick, and K. He, "Improved Baselines with Momentum Contrastive Learning," Tech. Rep., 2020.

[220] J. B. Grill, F. Strub, F. Altché, et al., "Bootstrap your own latent: A new approach to self-supervised Learning," *Advances in Neural Information Processing Systems*, vol. 2020-Decem, jun 2020.

[221] Y. Tian, X. Chen, and S. Ganguli, "Understanding self-supervised Learning Dynamics without Contrastive Pairs," Tech. Rep., 2021.

[222] D. Dwibedi, Y. Aytar, J. Tompson, et al., "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9568–9577, 2021.

[223] T. Chen, S. Kornblith, K. Swersky, et al., "Big self-supervised models are strong semi-supervised learners," *Advances in Neural Information Processing Systems*, 2020.

[224] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020.

[225] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *arXiv*, 2020.

[226] A. Ramesh, P. Dhariwal, A. Nichol, et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," apr 2022.

[227] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," 2019.

[228] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, vol. 2015 Inter, pp. 2623–2631.

[229] R. Gomez, L. Gomez, J. Gibert, and D. Karatzas, "Self-supervised learning from web data for multimodal retrieval," in *Multimodal Scene Understanding: Algorithms, Applications and Deep Learning*, pp. 279–306. 2019.

[230] V. Udandarao, A. Maiti, D. Srivatsav, et al., "COBRA: Contrastive Bi-Modal Representation Algorithm," Tech. Rep., 2020.

[231] Y. Zhang, H. Jiang, Y. Miura, et al., "Contrastive Learning of Medical Visual Representations from Paired Images and Text," oct 2020.

[232] Y. Zong, T. Yu, X. Wang, et al., "conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics," *bioRxiv*, , no. Cci, pp. 2022.01.14.476408, 2022.

[233] J. Kaur and C. Shekhar, "Multimodal medical image fusion using deep learning," in *Advances in Computational Techniques for Biomedical Image Analysis*, pp. 35–56. Academic Press, jan 2020.

[234] N. Pielawski, E. Wetzer, J. Öfverstedt, et al., "CoMIR: Contrastive multimodal image representation for registration," *Advances in Neural Information Processing Systems*, vol. 2020-Decem, 2020.

[235] B. Zhang, "Stain based contrastive co-training for histopathological image analysis," *arXiv*, 2022.

[236] M. Khokhlova, V. Gouet-Brunet, N. Abadie, and L. Chen, "Cross-Year Multi-Modal Image Retrieval Using Siamese Networks," in *Proceedings - International Conference on Image Processing, ICIP*, 2020, vol. 2020-Octob, pp. 2361–2365.

[237] P. Baldi and Y. Chauvin, "Neural Networks for Fingerprint Recognition," *Neural Computation*, vol. 5, no. 3, pp. 402–418, may 1993.

[238] J. Bromley, J. W. Bentz, L. Bottou, et al., "Signature Verification Using a "Siamese" Time Delay Neural Network," Tech. Rep. 04, 1993.

[239] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1735–1742, 2006.

[240] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[241] M. Boudiaf, J. Rony, I. M. Ziko, et al., "A Unifying Mutual Information View of Metric Learning: Cross-Entropy vs. Pairwise Losses," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12351 LNCS, pp. 548–564.

[242] W. Liu, Y. Wen, Z. Yu, et al., "SphereFace: Deep hypersphere embedding for face recognition," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 6738–6746, 2017.

[243] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive Margin Softmax for Face Verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[244] H. Wang, Y. Wang, Z. Zhou, et al., "CosFace: Large Margin Cosine Loss for Deep Face Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

[245] A. Zhai and H. Y. Wu, "Classification is a strong baseline for deep metric learning," in *30th British Machine Vision Conference 2019, BMVC 2019*, 2020.

[246] B. Kim and J. C. Ye, "Mumford-shah loss functional for image segmentation with deep learning," Tech. Rep., 2020.

[247] I. Elezi, S. Vascon, A. Torcinovich, et al., "The Group Loss for Deep Metric Learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12352 LNCS, pp. 277–294.

[248] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3235–3244.

[249] Q. Qian, L. Shang, B. Sun, et al., "Softtriple loss: Deep metric learning without triplet sampling," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6449–6457.

[250] P. Chikontwe, M. Luna, M. Kang, et al., "Dual attention multiple instance learning with unsupervised complementary loss for COVID-19 screening," *medRxiv*, 2020.

[251] J. Zou and L. Schiebinger, "AI can be sexist and racist — it's time to make it fair," *Nature*, vol. 559, no. 7714, 2018.

[252] A. J. Larrazabal, N. Nieto, V. Peterson, et al., "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 23, pp. 12592–12594, 2020.

[253] C. Ekelem, J. Yu, D. Heydarlou, et al., "The effect of melanin on in vivo optical coherence tomography of the skin in a multiethnic cohort," *Lasers in Surgery and Medicine*, vol. 51, no. 5, 2019.

[254] M. A. Wilk, A. L. Huckenpahler, R. F. Collery, et al., "The Effect of Retinal Melanin on Optical Coherence Tomography Images," *Translational Vision Science Technology*, vol. 6, no. 2, pp. 8, 2017.

[255] Y. Kim, J. H. Lee, S. Choi, et al., "Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records," *Scientific Reports*, vol. 10, no. 1, pp. 20265, 2020.

[256] R. J. De Berardinis and N. S. Chandel, "Fundamentals of cancer metabolism," may 2016.

Glossary

AI Artificial Intelligence.

CL Contrastive Learning.

CNN Convolutional Neural Network.

CV Cross Validation.

DCI Dynamic Cell Imaging (commercial denomination of DFFOCT).

DFFOCT Dynamic Full Field Optical Coherence Tomography.

DL Deep Learning.

FFOCT Full Field Optical Coherence Tomography.

FOV Field of View.

FS Fully-Supervised Learning.

GPU Graphics Processing Unit.

Grad-CAM Gradient-weighted Class Activation Mapping.

H&E Haematoxylin & Eosin Staining.

HPC High Performance Computing.

IHC Immunohistochemistry.

MIL Multiple Instance Learning.

ML Machine Learning.

MRI Magnetic Resonance Imaging.

NMF Non-negative Matrix Factorization.

OCT Optical Coherence Tomography.

ROI Region of Interest.

SLIC Simple Linear Iterative Clustering.

WSI Whole Slide Imaging.

List of Figures

List of Tables