



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Improving Breast Cancer Detection through Deep Learning and Digital Breast Tomosynthesis Leveraging Open-Source Data

TESI DI LAUREA MAGISTRALE IN
MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

Author: **Chiara Mocetti**

Student ID: 995451

Advisor: Prof.ssa Francesca Ieva

Company Advisor: Dr. Federico Comotto - LAIFE REPLY

Academic Year: 2022-23

Abstract

One in eight women will be diagnosed with breast cancer in her lifetime. The complexity of the disease is well-documented, yet its early-stage detection can significantly reduce mortality rates over the long term. As a universal public health issue, breast cancer is addressed through massive screening programs, demanding significant time and human resources. In this scenario, this thesis delves into the application of deep learning algorithms to streamline the detection of breast cancer lesions through Digital Breast Tomosynthesis (DBT), an advanced imaging technique reported to enhance diagnostic accuracy by providing a pseudo-3D view of the lesioned breast. While reducing women's recall rates of an estimated 30%, DBT requires twice as long reading time as the gold standard in breast screening, Digital Mammography (DM). Supporting radiologists in this resource-consuming process, automated lesion detection applications can make the difference.

The entire research is conducted navigating through the complexities of leveraging open-source data, with the objective of contributing to the body of knowledge by offering a methodological foundation for future applications on consistent and curated datasets. With the vast majority of the studies present in the literature relying on private datasets, this thesis aims to bridge the existing gap by developing and comparatively analyze two end-to-end pipelines. The first approach relies on Detectron object detection algorithm and explores the application of transfer learning from DM to DBT images. The second approach consists of a two-phased strategy, involving the classification of DBT patches followed by the lesion localization on full DBT slices. This solution produces an intuitive heatmap output, meeting the interpretability requirements of medical applications.

This analysis demonstrates the effectiveness of both the approaches, while drawing the attention to the challenges introduced by the utilization of open-source data and pointing to the advantages of transitioning to 3D modeling to exploit DBT diagnostic potential to the fullest.

Keywords: breast cancer detection; deep learning; digital breast tomosynthesis; open-source data

Abstract in lingua italiana

Il cancro al seno viene diagnosticato su una donna su otto, nel corso della sua vita. La complessità della malattia è ben documentata, tuttavia la diagnosi precoce può ridurre significativamente i tassi di mortalità a lungo termine. Le autorità sanitarie dei diversi paesi affrontano tale grave patologia con l'implementazione di programmi di screening su larga scala, che richiedono un significativo impiego di tempo e risorse umane. In questo scenario, la tesi approfondisce l'applicazione di algoritmi di deep learning per semplificare la rilevazione di lesioni da cancro al seno attraverso la Tomosintesi Mammaria Digitale (DBT), una tecnica di imaging avanzata che ha dimostrato di migliorare l'accuratezza diagnostica fornendo una vista pseudo-3D del seno lesionato. Sebbene riduca i tassi di richiamo delle donne di circa il 30%, la DBT richiede un tempo di lettura raddoppiato rispetto alla Mammografia Digitale (DM), che costituisce, oggi, la tecnica di riferimento nello screening del seno. Le applicazioni di rilevamento automatico delle lesioni, supportando i radiologi in questo dispendioso processo di analisi, possono fare la differenza.

L'intera ricerca si propone, pur con i vincoli e le complessità indotte dallo sfruttamento dei dati open-source, di contribuire al patrimonio delle conoscenze offrendo una base metodologica per le future applicazioni su dataset privati di più facile gestione. Sviluppando ed analizzando comparativamente due pipeline end-to-end, questa tesi mira a colmare le lacune derivanti dal fatto che la stragrande maggioranza degli studi presenti in letteratura si affidano a dataset privati. Il primo approccio si basa sull'algoritmo di rilevamento di oggetti Detectron ed esplora l'applicazione del transfer learning dalle immagini di DM alle DBT. Il secondo approccio consiste in una strategia articolata su due fasi, che prevedono la classificazione delle patches di DBT seguita dalla localizzazione delle lesioni su intere slices di DBT. Questa soluzione produce un output sotto forma di heatmap intuitivo, in linea con i requisiti di interpretabilità delle applicazioni mediche.

L'analisi dimostra l'efficacia dei due approcci e, focalizzando l'attenzione sulle sfide introdotte dall'utilizzo di dati open-source, fa emergere i vantaggi della transizione alla modellazione 3D, che consente di sfruttare appieno il potenziale diagnostico della DBT.

Parole chiave: rilevamento del cancro al seno; deep learning; tomosintesi mammaria digitale; dati open-source

Contents

Abstract	i
Abstract in lingua italiana	iii
Contents	v
1 Introduction	1
1.1 Research Question	4
1.2 Thesis Outline	6
2 AI for Breast Cancer Detection in DBT on Open-Source Data	7
2.1 Problem Introduction	7
2.2 DM versus DBT: Technologies Comparison	11
2.3 Motivation to the Development of Deep Learning Algorithms for DBT . . .	13
2.4 State-of-the-Art Analysis	16
2.5 X-RAIS: the Inspiration for a Further Exploratory Analysis on DBT	19
3 Data and Datasets	21
3.1 Data in Medical Imaging	21
3.2 Deep Learning Perspective	22
3.3 Data Sources	23
3.3.1 Breast Cancer Screening - Digital Breast Tomosynthesis (BCS-DBT)	23
3.3.2 Curated Breast Imaging Subset of the Digital Database for Screen-	
ing Mammography (CBIS-DDSM)	25
4 Methods	27
4.1 Approach 1: Detectron	31
4.1.1 Data Preparation and Image Pre-Processing	31
4.1.2 Model	35
4.1.3 Performance Metrics and Model Selection	38

4.2	Approach 2: Patches Model	41
4.2.1	Data Preparation and Image Pre-Processing	41
4.2.2	Model	46
4.2.3	Performance Metrics and Model Selection	52
5	Evaluation	55
5.1	Results	55
5.2	Discussion	55
6	Conclusion	65
6.1	Limitations	68
6.2	Further Developments	69
	Bibliography	71
	A Appendix A	75
	List of Figures	77

1 | Introduction

Breast cancer is a universally challenging public health issue, manifesting as a leading cause of cancer-related mortality among women worldwide.

The estimated statistics for 2024, projecting 2,001,140 new diagnoses and 611,720 deaths to occur exclusively in the United States [27], markedly highlight the urgency for effective prevention, detection and treatment strategies.

At the forefront of the thesis is the intention to develop deep learning algorithms for the automated detection of breast cancer lesions in Digital Breast Tomosynthesis (DBT) images, against the backdrop of the convoluted open-source data scenario.

The research delineates the exploration, development and assessment of two distinct approaches, handling the entire process from data acquisition and pre-processing to model designing, training, fine-tuning and evaluation.

The ambition is to establish an end-to-end methodological pipeline, which, once augmented with consistent and structured private data, holds the potential to significantly streamline the diagnostic process of the disease, thereby contributing to reduce its clinical burden.

The complexity and heterogeneity of breast cancer are well-documented, although numerous studies have consistently demonstrated that its identification at early-stage, when it presents minimal risk of death, can significantly lighten the load of the disease impact over the long term, thereby highlighting the indispensable role of comprehensive screening protocols capable of promptly detecting the tumor presence, and thus mitigating its implications [4].

Screening activity, if conducted efficiently and extensively, has the concrete potential to diminish the premature mortality associated with breast cancer. Anticipating the timing of the diagnosis, in fact, it enables for precise intervention directed at improving the prognosis.

Although it invests a primary role in alleviating the overall burden of breast cancer, the screening process is concurrently intricate, demanding and time-consuming. But, for women dealing with breast cancer, time is critical. With one in eight of them being diagnosed with the disease during her lifetime [4], there is an impelling need to expedite the

current workflow, reducing the time to treatment initiation.

Within this context, Artificial Intelligence (AI), in the form of deep learning models, is set to play a crucial role. By assisting the medical staff during the screening phase, the ability of deep learning models to automatically detect tumor lesions in imaging examinations can substantially optimize the diagnostic routine.

Screening programs adopt different approaches, yet the standard technique remains the conventional 2D X-ray mammography. However, this method is not effective in women under 40 years old and with dense breast tissue, where it exhibits reduced sensitivity to tumors smaller than 1 mm [14].

Mammography primary limitations include its tendency to produce a relatively high number of false positives and the associated phenomenon of breast tissue superimposition. In fact, the compression of the breast necessary to acquire the Digital Mammography (DM) image can lead to tissue overlap, which may falsely present normal areas as anomalies or obscure abnormalities.

In response to these drawbacks, the past decades have seen the advent of DBT, a low-dose X-ray imaging technique which operates with a pseudo-tomographic modality, generating a stack of 2D slices of the breast that collectively provide a vertical resolution. This feature significantly reduces the superimposition phenomenon and confers on DBT the capability of detecting traces of malignancies when they are most treatable. Indeed, DBT has been reported to reduce women recall rates by an estimated 30% [24]; this finding highlights its potential for early cancer detection.

However, a significant restriction to the widespread adoption of DBT in screening programs is posed by the considerable time required by its interpretation; reading a DBT volume takes roughly twice as long as a DM image [24]. The comprehensive analysis of up to 80 slices per single DBT examination [1] considerably extends the complexity of the reading process. Therefore, integrating this time-intensive practice into routine screening would place an undue workload on radiologists, challenging them to keep up with the steadily increasing volume of patient requests.

As a result, the advancements introduced by DBT can be harnessed on a large scale exclusively if coupled with strategies to expedite the examination process. Deep learning, wielding its revolutionary force in the interpretation of diagnostic imaging studies, if applied to this innovative technique, can establish it as the gold standard of breast screening.

To be effectively implemented and deployed, deep learning models require access to large-scale patient data. The employment of big data in medicine yet comes with significant privacy concerns and challenges.

The integration of advanced AI systems in healthcare involves a delicate balance between harnessing its transformative potential for clinical practice and addressing the multi-faceted legal and ethical complexities related to confidentiality constraints, which limit data acquisition.

Beyond general restrictions, the gathering of comprehensive and high quality medical data is further hampered by the variety of regulations in force across different jurisdictions, specifically the US and the EU.

The US's approach to health data privacy is predominantly custodian-centric, focusing on the entities handling the data [23]. Here, for the guidance of protecting health-related privacy, the Health Insurance Portability and Accountability Act (HIPAA) specifies 18 categories of protected health information (PHI) which must be removed before the health data is released to a third party [32]. This poses inevitable constraints and slowdowns to data collection, distribution and utilization.

The EU's General Data Protection Regulation (GDPR), conversely, adopts a more unified regime, setting out a single broadly defined regime for health data, independently of their source, format, or custodian [23]. The regulation is nevertheless stringent as the traditional data protection principles - purpose limitation, data minimization, special treatment of "sensitive data" – counteract big data deployment [10].

Regulatory restriction and divergence impact consistent and standardized data acquisition, essential for training robust deep learning models.

In addition, in the public sector, where the pace of data adoption lags behind that of the private sector, pronounced constraints are placed on data reliability. Open-source datasets do not inherently guarantee transparency and quality. Moreover, the voluntary nature of public data acquisition introduces uncertainty about the sustainability and continuous improvement of the related projects, as contributors can disengage at any time [28].

Health big data have already emerged as the most significant big data category for the huge potential value of secondary use and the serious privacy disclosure concerns involved. According to [32], the best course of action is to make such data available in accordance with the principle of minimum necessary. This perspective clearly stands in stark contrast to deep learning models demand for extensive and diverse datasets.

As a consequence, conducting research within an open-source framework presents substantial impediments. Simultaneously, it marks the critical starting condition for academic and practical investigations.

By relying exclusively on the restricted public data availability, the thesis is intended to develop a structured deep learning pipeline for breast lesion detection on DBT, with the ambition to provide a solid basis for breast cancer diagnosis advancement.

1.1. Research Question

Research on innovative approaches for the early detection of breast cancer is necessary. Physicians' expertise alone, without external support, is inadequate to handle the rising incidence of cases combined with the deployment of advanced techniques, foremost among them the DBT, which brings enhanced precision, albeit at the cost of increased reading times.

Concurrently, the screening activity should be expanded, in light of its reported effectiveness in mitigating the clinical burden of the disease, and should embrace the adoption of the newly developed imaging technologies into standard protocols.

The breakthrough driven by deep learning and Convolutional Neural Networks (CNNs) in diagnostic imaging draws the attention to the potential of leveraging these methods to streamline the DBT interpretation process.

The development of automated lesion detection systems, providing substantial support to radiologists during the initial demanding screening phase, is well-positioned to be the strategy to effectively implement the improvement of breast cancer early diagnosis.

Considerable research efforts have been conducted in this area, leading to the development of several mature commercial products. However, the vast majority of the present studies rely on private datasets. Thus, a methodology developed by exploiting open-source data exclusively is lacking.

The thesis aims to bridge this gap by establishing an end-to-end pipeline dedicated to deep learning-based lesion detection in DBT employing publicly available data.

Research in the medical field typically encounters significant hurdles, primarily in the form of scarce accessible, well-organized and consistent data, which is crucial to develop structured models and achieve reliable outcomes.

The intersection of stringent data privacy laws and clinical investigation introduces a complex level of challenges to the development of medical analytic models.

The strict regulations on data processing significantly limit the public availability of datasets annotated and systematically organized. Meanwhile, deep learning technologies, with their data appetite, face substantial obstacles in medical research contexts, due to the onerous regulations on data acquisition and elaboration. Although these ethical and legal restrictions are intended to safeguard patient privacy, they frequently slow down the rapid progress of healthcare research.

To effectively contribute to the improvement of breast cancer diagnosis leveraging automated detection models, researchers necessitate the access to extensive and curated privately-held, institutional data, allowing them to conducting in-depth analyses. How-

ever, in order to apply for such permissions, they must first establish the viability of their analytical approaches using the constrained scope of publicly available datasets.

A practical exemplification of the present circumstances is evident in the domain of deep learning solutions for DBT.

DBT, as an advanced imaging modality, offers significant benefits over traditional DM in detecting breast cancer at early stages. Yet, the adoption of DBT technique is not as widespread, and its availability is inconsistent across different imaging facilities. This disparity not only complicates the efforts of collecting a broad spectrum of data, but also highlights the existing divergences in research infrastructures, hindering the compilation of large and diverse datasets, necessary for robust analyses.

In this context, the Breast Cancer Screening-Digital Breast Tomosynthesis (BCS-DBT) database stands out as the unique resource offering open-source access to DBT data. This is emblematic of the challenges faced by researchers who aims at developing and validating deep learning models for lesion detection on DBT images.

While addressing the complexities of comprehensive, high-quality data collection remains a key challenge for the improvement of breast cancer diagnosis, we attempt to integrate powerful computational methodologies with the available data.

In the described challenging scenario, the objective of the thesis is to design an end-to-end pipeline for applying innovative deep learning solutions to breast cancer detection in DBT, utilizing open-source data. This effort seeks to strike an intricate equilibrium between leveraging advanced technological tools for medical research and complying with the stringent privacy regulations governing the access to patient data.

The study is intended to contribute to the research field by laying the groundwork for future analyses, ensuring a streamlined transition to the application of the developed methodologies on private datasets. By strategically establishing the workflow in advance, the solution is well-positioned to deliver impactful results when broader data collections become accessible for investigation purposes.

1.2. Thesis Outline

The remainder of the thesis is organized as follows.

- In **Chapter 2** we delve into the main concepts related to the issues addressed in the thesis. We highlight the significance of DBT as a diagnostic method and explore how deep learning technologies can enhance its effectiveness. Additionally, this chapter provides a comprehensive review of the latest advancements and methodologies within this domain, setting the basis for the research presented.
- In **Chapter 3** we introduce the datasets employed in our study, detailing their sources, characteristics, and the rationale behind their selection.
- In **Chapter 4** we articulate the methodologies adopted in our analysis. We explore and compare two distinct approaches to address the task at hand, offering insights into the techniques and algorithms implemented.
- In **Chapter 5** we present the outcomes of our proposed solutions, analyzing and discussing the results in the context of the objectives set forth at the beginning of the thesis.
- In **Chapter 6** we conclude by summarizing the key findings and contributions of the thesis, commenting on the limitations encountered, reflecting on the knowledge acquired through the process and proposing potential further developments in the research.

2 | AI for Breast Cancer Detection in DBT on Open-Source Data

2.1. Problem Introduction

DBT marks a considerable leap forward in breast imaging technology, overcoming through its innovative approach some of the limitations inherent in traditional 2D mammography. DBT technology employs a sophisticated image acquisition process that offers a pseudo-3D volumetric reconstruction of the breast. This is achieved by the acquisition of a sequence of stacked 2D images, from various angles, using a moving X-ray tube around the compressed breast that generates multiple projections. Depending on the manufacturer, the X-ray tube rotates in an arc that varies between 15° - narrow range - and 60° - wide range - on a plane aligned with the chest wall. A greater angular range of motion for the X-ray tube produces more tomographic information and improves vertical resolution on the z-axis, or section separation. For sufficient sampling, when the angular range for the tube motion increases an additional number of projections is necessary [7].

Generally, the system acquires from 15 to 80 projections, which are then digitally reconstructed into a series of thin slices of the breast, each typically approximately 1 mm thick, into a pseudo-3D image [12].

The patient's experience during a DBT scan is similar to that of a traditional mammography, with the breast being compressed between two plates.

The stacked slices can be examined individually or in combination, enabling radiologists to navigate through the breast tissue layers with improved clarity. The high-resolution volume, in fact, allows for sequential detailed examination, slice by slice. This facilitates lesion localization and characterization and may reduce or eliminate the necessity for additional diagnostic work-up.

Breast Tissue Superimposition Phenomenon

Conventional mammography faces challenges in detecting tumors within dense breast tissue, where overlapping structures can mask lesions. This phenomenon, known as superimposition, refers to the radiographic overlay of breast tissues, which may significantly

obscure the visualization of underlying carcinomas, complicating the interpretation of mammography examinations.

Screen-film and digital mammography both require taking a single, 2D picture of the breast. Consequently, breast tissues that are separated in the direction of the projection converge onto the same spot in the mammography image. As a result, distinct normal tissues may resemble a suspicious lesion, decreasing specificity, and normal tissues may hide the existence of a malignant lesion, decreasing sensitivity [24].

Superimposed breast tissue is estimated to mimic a lesion in about 5% to 12% of mammography screening exams. These cases require recalls for additional imaging evaluation and, occasionally, biopsy. Even more concerning, superimposed breast tissue potentially obscures a true lesion and miss cancer presence in up to 20% of cases [29].

For instance, most asymmetries, which theoretically indicate the presence of breast cancer, are caused by the summation of artifacts deriving from tissue superimposition in the DM. Their detection and diagnosis is challenged by the subtle presentation, similar to normal fibroglandular breast tissue [15].

The effect of superimposition is most pronounced in dense breast tissue, which is characterized by a higher proportion of fibroglandular tissue with respect to fatty tissue. The dense fibroglandular tissue appears radiopaque - white - on DM, similar to malignancies, thereby complicating the differentiation between non-lesioned and lesioned areas.

In dense breast tissue the superimposition phenomenon, which is present in around half of the screened cases, accounts for one-third of missed cancers. Additionally, in extremely dense breast tissue, it has been explicitly reported that only half of the tumors will be visible [8].

The American College of Radiology (ACR) Breast Imaging-Reporting and Data System (BI-RADS) Atlas provides systematic guidelines for breast density assessment, a critical factor in understanding the potential challenges posed by the phenomenon of superimposition to lesion detection.

Breast density is a measure of the volume of attenuating tissues within the breast which may obscure lesions, compromising the accuracy of mammography examinations.

The composition of the breast is distinguished into four categories, based on the visually estimated content of fibroglandular density tissue.

- Category A (almost entirely fatty): the breast is almost entirely composed of fatty tissue, with less than 25% of glandular tissue, which does not significantly attenuate X-rays. Mammography is highly sensitive in this setting, as the low density allows for a better visualization of structures and abnormalities.
- Category B (scattered areas of fibroglandular density): the breast presents scattered

areas of fibroglandular density tissue. Approximately 25-50% of the breast tissue is dense. Nonetheless, as the breast tissue is mostly fatty, it still has good sensitivity for mammography. It may be helpful, however, to distinguish breasts in which there are few scattered areas of fibroglandular density tissue from those in which there are moderate scattered areas of fibroglandular density tissue.

- Category C (heterogeneously dense): some areas of the breast may be relatively dense, potentially hiding non-calcified lesions within these denser regions. With 51-75% of dense tissue it becomes harder to detect tumors on mammographies, due to the phenomenon of superimposition, which in this cases might obscure small masses.
- Category D (extremely dense): the breast is almost entirely composed of dense glandular and fibrous tissue, with more than 75% of the breast being dense. This significantly lowers the sensitivity of mammography, due to the dense fibroglandular tissue masking effect on mammography depiction of non-calcified lesions.

Figures 2.1, 2.2, 2.3 and 2.4 report four different breast densities categorized according to BI-RADS criteria.

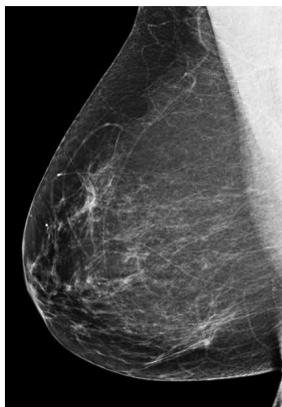


Figure 2.1: Category A

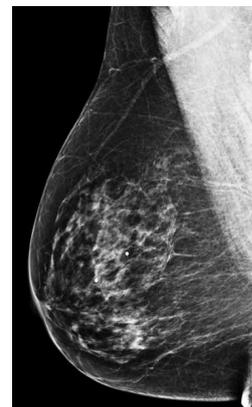


Figure 2.2: Category B

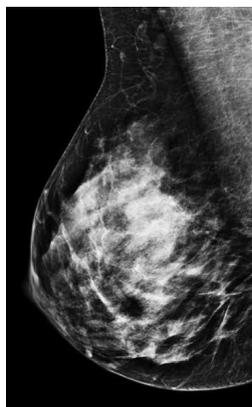


Figure 2.3: Category C

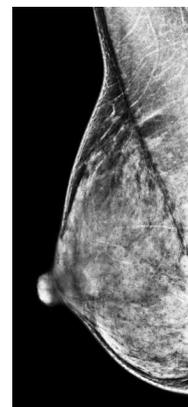


Figure 2.4: Category D

It must be remarked that visually estimating the breast density of any two adjacent categories implies considerable intra- and inter-observer variability. The densest breast in one density category and the least dense breast in the next higher density category only slightly and insignificantly differ in terms of mammography sensitivity. Due to these factors, the breast density categorization of individual cases has limits in terms of clinical relevance and must be used as a standard supporting tool for breast examination reporting [2].

The phenomenon of superimposition, particularly relevant in categories C and D, according to BI-RADS, introduces considerable impediments to mammography imaging. The tumor masking effect generated by the overlapping of fibroglandular tissue leads to a high rate of missed diagnoses.

DBT 3D imaging technology improves the visibility of abnormalities within dense breast tissue, thus enhancing the effectiveness of early cancer detection in high-risk groups. By providing a clearer and more detailed image resolution, it offers substantial advantages in lesion identification and characterization, facilitating a more detailed assessment of both the morphology and the architecture of suspicious areas within the breast.

The depth information provided by DBT enables radiologists to distinguish between benign and malignant lesions with higher confidence, reducing the likelihood of false-positive results.

Furthermore, DBT enhances the visualization of the lesion margins and the detection of architectural distortions and small calcifications, decisive for the precise diagnosis and staging of breast cancer [7].

Several study findings prove that DBT offers significant advantages across diverse breast compositions. Reduced recall rates and increased cancer detection capability are observed in patients with dense breast tissue, as well as those with fatty or scattered tissue.

An observational study comparing recall rates, biopsy rates, cancer detection rates, and positive predictive values for radiologists employing DBT versus those not integrating it in the examination process, published by Rose et al., demonstrated gains in all the assessed performance metrics when DBT was utilized, and in all breast density categories.

Margolies et al. examined the impact of incorporating DBT while examining a group of patients with a higher risk of breast cancer. The researchers discovered that patients with breast densities classified as categories C or D on the BI-RADS standards were substantially more likely to have DBT findings resulting in a treatment modification - 13% of patients with highly dense or heterogeneously dense tissue versus 9% of patients with fatty or dispersed tissue. The study assessed that three additional cancers detected on DBT were found in patients with heterogeneous or extremely dense breasts. Still, DBT

also improved treatment for participants with less dense breast patterns [8].

In summary, DBT improves diagnostic accuracy across all the categories of breast density, with advantages pronounced in cases of dense breast tissue.

Additionally, the prospective population-based study Malmö Breast Tomosynthesis Screening Trial highlights the relevance of DBT as a standalone modality for breast cancer screening. Results show that DBT was superior in terms of detection rate and equal in terms of positive predictive value compared to the current gold standard DM [16].

DBT represents a significant technological leap forward in the field of breast imaging and can be fundamental in the context of personalized screening strategies, being particularly beneficial for women with dense breasts, where, instead, traditional mammography might fall short. Nonetheless, its equally proven advantages for non-dense breasts indicate it as the perfect candidate for a universally effective screening tool.

The integration of DBT into clinical practice can facilitate a more efficient use of health-care resources. In fact, DBT increased precision compared to DM may reduce women recall rates of a significant amount, since fewer of them will be asked to repeat the examinations due to uncertain results. Avoiding unnecessary diagnostic procedures, not only the medical costs decrease, but also the anxiety and inconvenience for patients are minimized.

2.2. DM versus DBT: Technologies Comparison

The acquisition of DBT images exhibits similarities with DM.

In DM, the breast is compressed between two paddles, with the X-ray detector positioned underneath and the X-ray source situated above, perpendicularly aligned to the detector. Typically, two images for each breast are acquired: craniocaudal (CC) and mediolateral oblique (MLO). The significant compression of the breast against the detector plate allows these images to achieve higher resolutions and contrast levels than many traditional radiographic methods, all while administering a relatively low radiation dose, known as the average glandular dose (AGD).

However, the main limitation of DM, as previously underlined, is the tendency of 2D views to produce artifacts from the overlay of parenchymal tissue, which appears similar to malignant lesions, thereby obscuring them.

DBT addresses this issue by capturing a series of images as the X-ray source rotates around the compressed breast, which results in a collection of Projection Views (PVs). The specifics of reconstruction may vary, but the commonly used systems all reconstruct these PVs in a stack of parallel slices, known as a z-stack, in the standard CC and MLO

orientations. This technique offers an optimal balance: by maintaining the breast position and compression against the detector, it is possible to preserve the necessary resolution and contrast, despite the limited range of PVs [4]. Figure 2.5 illustrates the acquisition techniques of DM and DBT, respectively.

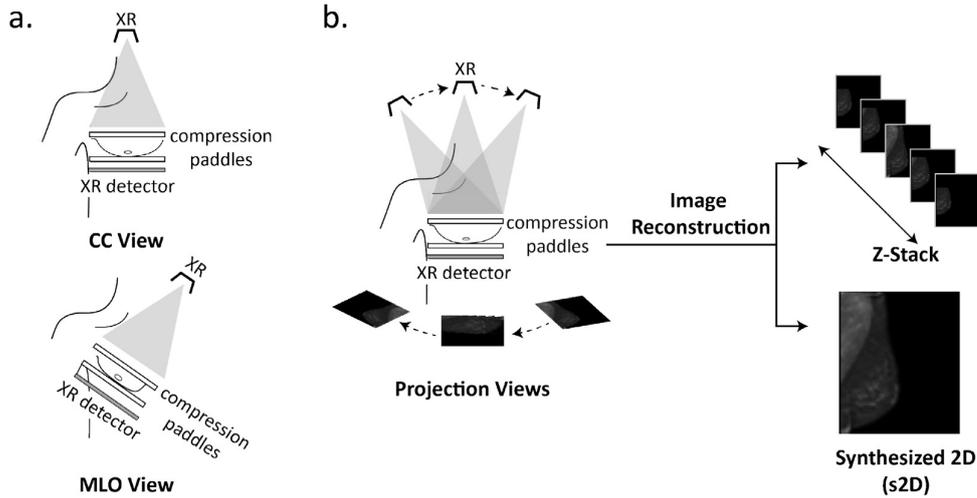


Figure 2.5: DM (a.) and DBT (b.) acquisition techniques

A number of retrospective studies demonstrates that DBT is an effective stand-in replacement for DM.

A first analysis, encompassing data from 23,149 patients screened with DBT and 54,684 with DM, reveals significant improvements in clinical performance metrics in favour of DBT. In details, the implementation of DBT resulted in a 16.1% reduction in recall rates, proving the technique efficacy in minimizing unnecessary follow-up procedures. Furthermore, the overall cancer detection rate for DBT was 28.6% higher with respect to DM, with a particularly striking 43.8% increase in the detection of invasive cancers. Additionally, the positive predictive value for recalls from screening was 53.3% greater with DBT, as an indicator of its substantial reliability [11].

A second research, including 5,703 DBT examinations and 80,149 DM examinations, interpreted by 10 breast-specialized radiologists, reports that DBT allowed 54.3% more carcinomas to be detected, compared to DM. The overall recall rate was 7.51% for DM and 6.1% for DBT. In addition, for patients with extremely or heterogeneously dense breasts and in the 40-49-year and 60-69-year age groups, this difference was even more significant [25].

A third retrospective analysis, evaluating a total of 44,468 examinations attributable to 23,958 unique women over a 3-year period, assessed the sustainability over time of DBT improved outcomes and its impact over multiple screenings at both population and indi-

vidual levels. Although a slight increase in recall rates was observed over the first three years of DBT screening, these recall values remained significantly reduced if compared to the values registered with DM. The ratio of cancer cases per recalled patients showed a notable increase with the employment of DBT over time, shifting from a DM baseline rate of 4.4% to 6.2%, 6.5%, and 6.7% for DBT for years 1 to 3, respectively. Moreover, women who underwent multiple DBT screenings revealed decreasing recall rates with each additional DBT screening. The interval cancer rates decreased from 0.7 per 100 women screened with DM to 0.5 per 1,000 women screened with DBT [20].

A fourth investigation on 84 women in whom mammography was negatively interpreted, showed that, after the supplementary screening through DBT, 54% of cancers previously not evident on DM were instead detectable on DBT [21].

These findings confirm that the advanced imaging capabilities of DBT lead to an improved diagnostic accuracy, advocating for its widespread adoption as the standard practice in screening protocols.

2.3. Motivation to the Development of Deep Learning Algorithms for DBT

The reading time required by DBT is a critical factor that must be considered in high-volume screening programs, when deciding whether to use the DBT as imaging modality. In contrast to a traditional DM exam, interpreting a DBT exam involves the visualization and analysis of tens of image slices. This significantly increases the complexity and implementation time of clinical radiology workflows.

A study by Astley et al. was conducted to evaluate the employment of DBT in screening scenarios by comparing the time required for interpreting DBT images to that necessary for traditional DM images, among four radiologists. The outcome revealed a significant difference in the median reading times: a value of 66 seconds was registered for DBT against the 17 seconds of DM. Interpreting DBT images required approximately four times as long as DM images. The study, while acknowledging that the complexity encountered in the cases exceeded typical screening scenarios and suggesting that the significantly extended reading time might have partly stemmed from radiologists' unfamiliarity with DBT interpretation, still stated that in a personalized screening environment DBT should be reserved to women at high risk or with an established dense breast condition [3].

A second observer study, which compared the diagnostic accuracy of DM with both two-view (MLO and CC) and single-view (MLO) DBT, analyzing the performances of accredited readers on 220 women with breast density of categories B, C and D, revealed a

significant extension in reading times for DBT images. The authors registered that the average reading time was 124 seconds for two-view DBT and 97 seconds for single-view DBT, in stark contrast to the 67 seconds required for DM [30].

When considering the possibility to effectively adopt the DBT technology as a routine screening tool for both dense and fatty breast compositions, it is necessary to be aware of the longer reading times involved. Although these times should decrease with radiologists' practice, through the introduction of DBT into standard clinical protocols, they will invariably remain longer compared to traditional DM, due to the inherent examination of multiple image slices.

Meanwhile, particularly in the initial stages of the screening process, extended reading times affect radiologists' workload, patient's scheduling, and may potentially culminate in an overall deadlock of day-to-day practice workflow.

Nevertheless, several studies collectively emphasize the benefits which DBT offers to the broad spectrum of the female population. Research points out the importance of integrating DBT into the standard screening protocols, not only for high-risk groups but for all women, to fully capitalize on the technology capabilities of advancing breast cancer screening.

Limiting DBT employment to women with a specific breast density category would signify missing valuable opportunities to leverage the advantages it provides in enhancing cancer detection rates, reducing recall rates, and potentially improving outcomes for a wider audience.

In [30], in addition to the extended reading times registered for DBT, a second factor emerges: the performance gap between DM and two-view DBT was significantly noticeable among readers with less experience.

In the determination of the variability in radiologists' interpretation of screening mammographies, also Duijm et al. observed large divergences in the outcomes depending on the nature and number of readers, as the recorded sensitivity ranged from 51.5% to 75% [9].

This observations highlight an additional challenge in breast imaging interpretation: the variability in reading times and diagnostic accuracy among different radiologists, a phenomenon known as inter-reader variance.

Inter-Reader Variance

Inter-reader variance in the interpretation of breast X-ray images affects both the accuracy and the consistency of the diagnosis.

The level of training and experience of radiologists can markedly influence the interpretation of the examinations. DBT studies, which are inherently more complex with respect

to traditional 2D DM images, need highly trained radiologists adept at comprehensively capturing the features of the entire volume.

The observed unevenness in diagnostic performances evidently call for the introduction of radiologists' specialized training aimed at improving interpretative skills. However, this approach may not be sufficient to fully tackle the issue.

The phenomenon of inter-reader variance is partially addressed by the process of double reading, where two radiologists independently review the imaging exams. The process is executed with various methods. Double-reading can be independent, if the second reader is not informed about the first reader's decision or non-blinded, in the opposite case. Alternative approaches for double interpretation include utilizing either a radiographer or Computer-Aided Detection (CAD) devices for the secondary review.

Screening programs employ a variety of techniques to solve readers' disagreements. If one reader deems a study abnormal, the patient may be recalled without any discussion of disagreements; alternatively, the examinations may be interpreted in consensus, allowing for recall only with the consent of the participating radiologists; as an additional case, a panel arbitration process may be used to resolve the disagreements [9].

Several researches emphasize the potential of double reading in DBT to mitigate inter-reader variance, by leveraging the complementary expertise of radiologists. However, this process can lead to conflicts which further complicate the determination of the most appropriate clinical action to undertake, highlighting the need for consensus strategies or additional imaging assessments to resolve the disagreements.

In conclusion, the improvement in accuracy is counterbalanced by longer interpretation times. Double reading divergences, indeed, contribute to potential delays in breast cancer diagnosis.

In this intricate scenario deep learning has the promising potential to transform the analysis of DBT images, addressing key challenges such as the long reading times and the inter-reader variance.

Deep learning algorithms can rapidly process and analyze vast amounts of DBT data identifying suspicious areas, thereby acting as a screening support tool to be associated to the further closer examination of radiologists.

Furthermore, the automatic localization of potential lesions can direct radiologists' attention to specific regions of interest within the images, thus reducing the time required for a complete exam review. This advantage is a turning point specifically in the analysis of DBT volumes, which require increased attention, experience and time.

A retrospective study by Rodriguez-Ruiz et al. demonstrated that AI-assisted reading can decrease the interpretation time without compromising the diagnostic accuracy, stream-

lining the screening process. This research, comparing commercial AI systems to average-trained breast radiologists reported that the case-level performance of the algorithms was statistically non-inferior to human-performance level. The authors, analyzing nine datasets collected from different sites in both the US and Europe, showed that AI was able to detect lesions correctly better than 61% of the 101 radiologists involved in the investigation [24].

Additionally, to contrast the variability in the diagnoses, deep learning models of lesion detection can provide a consistent, robust and standardized assessment of DBT images. Indeed, the developed automated systems can be employed as a second reader, providing a level of review that is not influenced by the individual radiologist's experience, fatigue, or subjective interpretation, thus leading to more uniform and reliable diagnostic outcomes [17].

In conclusion, the adoption of the DBT in large-scale screening programs will rely on its effect on clinical outcomes, but only if it is preceded by the development of technologies to shorten the associated reading time. In this context, deep learning-based methods for DBT volumes analysis will undoubtedly make the difference on the possibility of effectively employing this advanced technique in screening routine.

The impact would be twofold. In the first place, the amount of time each radiologist spends visually searching for suspicious findings would be significantly reduced as a result of the computer-driven faster navigation of the DBT slices. Moreover, minimizing inter-reader variance, the AI techniques designed to assist the interpretation process of DBT images would contribute to diminish the variability recorded by the multiple studies conducted to analyze the impact of DBT on detection and recall rates in the screening activity.

2.4. State-of-the-Art Analysis

AI applications to DBT, leveraging deep learning models primarily based on CNNs, have demonstrated promising results, paving the way to their employment in assisting radiologists by overcoming the limitations of manual analysis, such as the high reading times and the potential for human error.

This section delves into an exploration of the cutting-edge researches most relevant in the specific domain of the thesis.

Among the lesion detection algorithms, we cite the three-stage approach proposed by Lotter et al., which leverages a pre-trained ResNet-50-based model for the initial classification of DM patches extracted from 24,253 images belonging to both public and private

datasets. This trained architecture is subsequently used as the backbone for the localization phase. In this step, the application of multiple-instance learning (MIL) for the classification of DM and optimized 2D images condensed from DBT z-stacks allowed the model to achieve an improved sensitivity of 14%, through the detection of previously negative cancers, and an AUC of 0.945. The proposed solution outperformed five out of five radiologists in the patch-level classification and subsequent detection on full images. Additionally, the algorithm includes a domain specialized strategy. By utilizing multi-scale CNNs for distinct lesion types, namely masses and calcifications, and afterwards aggregating the outputs in a final classification estimate, this multifaceted approach contrasts the challenge posed by the frequent variability in lesion characteristics.

Further contributions are provided by Fan et al., who introduces a Faster-RCNN-based method for efficient mass detection in DBT images, supplemented by a model for false positives reduction, which achieved an AUC of 0.96. This work was subsequently extended in 2020, with a 3D version of the Mask RCNN model, which demonstrated a sensitivity of 0.9 at 0.8 false positives per breast, a more relevant result with respect to the 2D methodologies. The research highlights two aspects which will be object of subsequent exploration within the thesis: addressing the domain present issue of the false positives and the prospective benefits introduced by the volumetric analysis.

Efforts to minimize training dataset requirements have been explored by Samala et al.. Their work exemplifies the effective application of transfer learning from pre-trained networks, AlexNet in this case, to counteract the limitations posed by poor, low quality lesioned image patches. Furthermore, the algorithm provides a successful application of CNNs pruning, which, without compromising the performance, partially addressed the hurdles posed by the inadequate training datasets.

Li et al. implemented a model specifically targeting the detection of architectural distortions using mammary glands distribution as a prior information. Results showed that the proposed solution has a sensitivity of 0.8 at 1.95 false positives per image. The authors claim that identifying distortions which are complex to annotate is one of the real strengths of 3D imaging, placing the attention on the potential for models trained on DBT to detect lesions previously not visible on DM.

For the development of their anomaly detection model, Swiecicki et al. employed Generative Adversarial Networks (GANs) to generate abnormal breast tissue images from normal DBT data. The rationale behind the research is that if the generated image is substantially different from the original image, then technically, the region is abnormal. However, the lesions produced by GAN appeared irrelevant compared to the original ones, as the average intensity of the pixels in the patches produced was twice as that of the normal ones. This study evidences the ongoing question for the design of new efficient

data augmentation techniques, since the standard transformations, although widely applicable, present limitations.

Kooi and Karssemeijer developed deep learning algorithms capable of analyzing bilateral views and incorporating asymmetry analysis, besides methods to compare current exams to prior analysis, marking a significant progress in the direction of a comprehensive examination.

The inability to analyze simultaneously diverse images, including controlateral and previous examinations, is among the primary limitations of the current AI systems developed for lesion detection. Strategies in this direction have the potential to produce results comparable to radiologists' assessment which involves the access to such priors.

Still leveraging on DBT bilateral views analysis, it is worth to mention Kim et al. divergent strategy, although not specifically targeted to lesion detection, but more to image classification. The authors devised an AI methodology which does not require annotated images for training. This method achieved a sensitivity of 0.761 using a dataset including 4,000 cancer cases and almost 25,000 normal cases of screening and diagnostic images from three DM system vendors. This development signifies a shift towards more scalable and less labor-intensive methods. However, it also points out the imperative requirement for even more extensive datasets, if unannotated.

Currently, several commercial AI applications approved by Food and Drug Administration (FDA) or Conformance Européenne (CE) marked have been integrated into clinical settings. Recent related evaluation studies have been conducted, such as the analysis of Conant et al. which compared the stand-alone performance of several AI systems to that of radiologists reading the examinations retrospectively, obtaining that the average sensitivity of the former, with a value of 0.91, was considerably higher with respect to the 0.77 obtained by the latter. Nevertheless, the algorithms achieved a sensitivity of 0.41, which, compared to the value of 0.627 recorded for the radiologists, highlights a major challenge in lesion detection algorithms, namely the model tendency to false positives.

In a second assessment study, Schaffter et al. included the analysis of the algorithms predictions merged with the radiologists decisions, revealing the winning strategy for a considerable accuracy improvement [24], [13], [4].

The researches discussed underscore a broad trend towards the improvement of the accuracy, efficiency, and comprehensiveness of diagnostic tools in breast cancer screening and diagnosis through the sophisticated application of deep learning technologies.

2.5. X-RAIS: the Inspiration for a Further Exploratory Analysis on DBT

The research question arises in the context of the continuous activity of the R&D Imaging team of Laife Reply, the Reply Group company that operates in the Health, Welfare and Pharma sectors, and is dedicated to the application of AI across a broad spectrum of domains, including Medical Imaging, Drug Discovery, Digital Therapeutics, and the analysis of unstructured data through Natural Language Processing.

The collaborative work of Data Scientists, AI Engineers, and Data Architects at Laife Reply led to the development of X-RAIS, a platform that has established its presence in the market over several years, and has gained recognition, forming research partnerships with the scientific directorates of leading Scientific Institute for Research, Hospitalization, and Healthcare (IRCCSs), both public and private, along hospital companies and research centers.

X-RAIS introduces an innovative AI tool designed to support radiologists in DM interpretation by precisely detecting and characterizing microcalcifications. After training, this tool can autonomously assess new DM exams through the detection and reliable categorization of microcalcifications, providing radiologists with a valuable second opinion to enhance diagnostic precision. Demonstrating high accuracy in the localization and characterization of microcalcifications in DM, X-RAIS underscores the significant potential of AI-based systems to augment the radiologists' interpretation of the examinations [22].

The successful development of this standardized and observer-independent system for identifying and classifying microcalcifications has sparked further interest in exploring the capabilities of deep learning models for DBT, which is seen as the 3D advancement of the traditional DM.

This study is conducted within an entirely open-source framework, with the intent of establishing a foundation that may induce the company's partnerships to collect and provide structured and comprehensive private data on which fine-tuning the pipeline designed and implemented. This would enable the enhancement of the X-RAIS functionalities with a DBT-based model that improves the solution diagnostic accuracy and efficiency.

3 | Data and Datasets

3.1. Data in Medical Imaging

Lesion identification through breast imaging techniques relies on a detailed analysis of precise visual features, critical for identifying potential abnormalities and distinguishing them between benign and malignant.

Malignancies recognition requires a structured approach, encompassing the inspection of size, shape, margins, intensity or density, texture and contrast, which collectively provide a comprehensive tumor characterization. BI-RADS criteria report the clinical features associated with different lesions types on DM, which can be extended to DBT.

Masses represent space-occupying lesions visible in two different projections, essential for distinguishing between benign and malignant entities. They present with round, oval or irregular shape and margins circumscribed, obscured, microlobulated, indistinct or spiculated. They can be high-density, equal-density, low-density or fat-containing. Notably, fat-containing lesions are generally benign.

In architectural distortions, instead, normal breast architecture appears disrupted with no visible mass, potentially indicative of scar tissue or carcinoma.

Asymmetries represent unilateral fibroglandular tissue not conforming to mass criteria, with variations including simple asymmetry, focal asymmetry, global asymmetry, and developing asymmetry, each necessitating careful evaluation for distinctions from masses.

Calcifications, which are typically benign, present skin vascular, coarse or "popcorn-like", large rod-like, round, rim, dystrophic, milk of calcium and suture. The morphology can be amorphous, coarse heterogeneous, fine pleomorphic, fine linear or fine-linear branching. The distribution patterns - diffuse, regional, grouped, linear, and segmental - offer insights into the likelihood of malignancy. The shape of calcifications, whether round and punctate or branching and linear, is relevant in the assessment of lesion malignity [33], [31].

Generally, malignant tumors have irregular, spiculated or lobulated edges, which suggest an invasive growth pattern. Benign lesions, as cysts or fibroadenomas, conversely, typically have smoother, well-defined and rounded borders.

A paramount feature allowing the identification of lesions in radiographic images is the intensity, or density. Malignant lesions may appear as areas of increased density, due to their composition, compared to the surrounding breast tissue. However, in denser breasts, the contrast between the lesions and the fibroglandular tissue may be less pronounced, posing challenges to accurate detection. As previously remarked, DBT, with its ability to provide a pseudo-3D perspective, can aid in overcoming this issue, by offering sliced images which reduce the impact of superimposition, allowing for a better visualization of the lesion density.

Finally, the contrast between a lesion and the surrounding breast tissue is crucial for lesion identification. High-contrast features are relatively easy to detect and can be indicative of the lesion nature. In fact, malignant lesions may alter the normal architecture of the breast, creating distortions or architectural distortions in the tissue that can be detected as areas of contrast on DM or DBT images.

3.2. Deep Learning Perspective

As DM and DBT images exhibit a highly heterogeneous array of shape features, a deep learning model strategically searches for hyperintense pixels indicative of potential abnormalities.

During the training of a CNN, the learning procedure executes feature selection and classification tasks, without the necessity for the intervention of an expert. The process, however, involves feeding the CNN with a vast array of annotated medical images, allowing the model to iteratively learn and refine its understanding of the data.

Given the heterogeneous characteristics of different lesion types and the frequently still-limited data availability, the task is highly complex. Training datasets, in fact, need to be sufficiently large and diverse to cover the broad spectrum of phenotypes representative of the lesions under investigation. However, this is not always the case, specifically in the open-source framework.

Initially, the CNN first layer captures basic attributes such as edge location and orientation. Subsequent layers build upon these foundational elements, learning to recognize specific combinations and configurations of the initial simple features. As the network delves deeper, it begins to comprehend increasingly complex patterns and arrangements derived from the previous feature sets. In the final layers, the CNN leverages these sophisticated imaging features, or representations, to accurately classify the images or identify other significant patterns [5].

The algorithms used to automatically identify breast cancer lesions in radiographic exams typically go beyond standard deep learning CNNs, due to the unique features of DM and

DBT images. Not only should the algorithm detect whether an image contains a lesion, but it should also identify the position of such suspicious finding. This calls for methods that extend deep learning CNNs typical classification algorithms.

3.3. Data Sources

For the purposes of the research we rely on two datasets of annotated images of tomosynthesis and mammography exams.

3.3.1. Breast Cancer Screening - Digital Breast Tomosynthesis (BCS-DBT)

Breast Cancer Screening - Digital Breast Tomosynthesis (BCS-DBT) is a comprehensive dataset including masses and architectural distortions annotated by two experienced radiologists.

Derived from the Duke University Health System, it encompasses DBT volumes, in DICOM format, acquired between August 26, 2014, and January 29, 2018. The dataset, approved by the Duke University Health System institutional review board with a waiver of informed consent due to its retrospective nature, comprises 16,802 DBT studies, with at least one reconstruction view available, from 13,954 patients, following the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline.

The curated dataset contains 22,032 reconstructed DBT volumes which belongs to 5,610 studies from 5,060 patients with a mean age of 55 years and standard deviation of 11 years, 5,059 of them women.

These studies were categorized into four distinct groups based on radiological findings and subsequent clinical actions: 5,129 normal studies (91.4%), without abnormal findings; 280 actionable studies (5%), necessitating further imaging examination; 122 benign studies (2%), with benign masses or architectural distortions biopsied based on the DBT examination; 89 cancer studies (1.6%), with cancerous masses or architectural distortions biopsied.

Each study included left craniocaudal (LCC), right craniocaudal (RCC), left mediolateral oblique (LMLO), and right mediolateral oblique (RMLO) reconstruction views, except where specific exclusions are applied, such as the presence of foreign objects or the visibility of findings only on spot compression views.

For annotation purposes, the study images alongside the corresponding radiology and pathology reports for biopsied cases were reviewed by two radiologists, with 18 and 25 years of experience, at Duke University, who identified and annotated the masses and

architectural distortions. This process yielded 190 bounding boxes for cancerous lesions and 245 for benign lesions, facilitating a detailed analysis of lesion characteristics across 336 masses and 99 architectural distortions.

The final composition includes a training set encompassing 4,838 studies from 4,362 patients, a validation set containing 312 studies from 280 patients, and a test set comprising 460 studies from 418 patients [6].

We compile all the 431 lesioned DBT volumes, presenting a total of 435 annotated lesions, into a unified dataset to perform a custom train-test split maintaining consistency with the split ratios established in the rest of the thesis. We notice that each volume presents generally a single lesion, with the exception of a minimal number of examples.

A single DBT examination is associated with the following information:

- PatientID: the patient identifier code;
- StudyUID: the code associated to the DBT study. Each exam, indeed, is composed of a maximum of four volumes, i.e. the CC and MLO views for the two breasts;
- View: an acronym encoding the view, among RCC, LCC, RMLO and LMLO reconstructions;
- Subject: an identification number encoding the radiologist who performed the annotation;
- Slice: the central slice index, across the full volume, of a biopsied lesion;
- X: the coordinate, on the horizontal axis, of the left edge of the predicted bounding box;
- Y: the coordinate, on the vertical axis, of the top edge of the predicted bounding box;
- Width: the predicted bounding box width, i.e. along the horizontal axis;
- Height: the predicted bounding box height, i.e. along the vertical axis;
- Class: the label indicating if the lesion is malignant or benign;
- AD: a binary value specifying if an architectural distortion is present;
- VolumeSlices: the total number of slices in the volume.

Among the metadata, we focus our attention on two information: the slice index highlighting the 2D image, across the stack, where the lesion appears more evident, according to the radiologists' knowledge; the Region of Interest (ROI), represented as a bounding

box delineating the lesion precise spatial domain, defined by its minimum and maximum coordinates on both axes: $(x_{min}, y_{min}, x_{max}, y_{max})$. To better understand this notation, please refer to the Figure A.1.

3.3.2. Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM)

Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) is an updated and standardized version of the original Digital Database for Screening Mammography (DDSM).

The DDSM, established as a repository of 2,620 scanned film mammography studies, encompasses a wide spectrum of mammography cases including normal, benign and malignant classifications, all associated with verified pathology information.

The CBIS-DDSM dataset represents a carefully selected subset of the DDSM, in DICOM format, curated by experienced radiologists to enhance its utility for academic and clinical research.

This subset has undergone images decompression and conversion into DICOM format. Additionally, the CBIS-DDSM includes updated ROIs, alongside pathological diagnoses for the training data.

Notably, despite the DICOM metadata suggest the presence of 6,671 patients due to multiple patient IDs for individual participants, the actual cohort consists of 1,566 participants.

Data related information are split according to lesion type: masses and calcifications.

The dataset comprises mammography full images, cropped images centered on the lesion, and mask images, providing the binary localization of the lesion on the full image. Note that each study generally includes multiple reconstruction views. We restrict the analysis to full images and corresponding mask images.

Images are associated with the following information:

- Patient_id: the patient identifier code;
- Breast_density: the breast density category according to BI-RADS;
- Left or right breast: the laterality of the exam;
- Image view: an acronym encoding the view, among CC and MLO;
- Abnormality id: a code associated to the specific abnormality;
- Abnormality type: if the abnormality is a calcification or a mass;

- Mass shape/Calc type: the mass shape or the calcification type;
- Mass margins, Calc distribution: the mass margins or the calcification distribution;
- Assessment: the risk assessment, according to BI-RADS;
- Pathology: the lesion characterization as malignant, benign or benign without call-back.
- Subtlety: the degree of difficulty in detecting the abnormality.

4 | Methods

To design an automated pipeline for tumor lesion identification in DBT images relying on open-source data, our methodology involves the development of a deep learning-based framework.

Firstly, we conceptualize the problem within a 2D scope, with the intent of subsequently further extending the model capabilities to embrace the intrinsic 3D nature of DBT images, thereby leveraging their diagnostic potential to the fullest.

Our primary dataset, the BCS-DBT, consists of DBT volumes, associated with a set of metadata, from which we retain only the information relevant to our investigation. The dataset is indexed following the lesion numbering within the images. Each lesion is explicitly linked to a DBT volume and associated with two essential information: the slice index highlighting the 2D image, across the stack, where the lesion appears more evident, according to the radiologists' knowledge, and the ROI, represented as a bounding box delineating the lesion precise spatial domain, defined by $(x_{min}, y_{min}, x_{max}, y_{max})$.

We extract the 2D images corresponding to the radiologists-identified slices. As a result, we obtain a 2D dataset on which we intend to perform an object detection task.

Given the restrictions posed by the limited dataset size, we adopt a transfer learning strategy, exploiting mammography images from the CBIS-DDSM dataset, in order to facilitate the model learning experience.

Transfer Learning from Mammography

Deep learning methods for lesion localization, as techniques using pixel- or patch-level analysis, typically call for large-scale annotated training datasets, where the abnormalities are outlined in the images, with the identification of a ROI or, alternatively, datasets of image patches containing the lesions. These requirements significantly increase the challenge of acquiring adequate training datasets, since image annotation is a time-consuming, laborious process which necessitates the expertise of subject-matter specialists and is still prone to inter-reader variability.

Considered that limited datasets can be a bottleneck to the further advancement of medical AI models and established that building progressively growing well-annotated datasets

is at least as important as developing new algorithms, in the meantime, it is of interest to reduce the quantity of training these algorithms are based on and, consequently, the size of the datasets sufficient to complete the training process [24].

Transfer learning from mammography is an efficient technique to accomplish this objective.

Leveraging on the similarities between DBT and DM imaging modalities, a great deal of research has been done to determine how the deep learning advancements for DM image processing connect to those in DBT and whether or not some of the techniques, training datasets, and knowledge acquired for one can be used for the other.

It has been discovered that networks trained for DM can be fine-tuned for DBT, with a process similar to the standard practice of model initialization through natural images. AI-based image recognition techniques, indeed, are generally built on the backbone of CNN models pre-trained on vast and heterogeneous datasets, including images belonging to completely different domains with respect to the specific field of interest. Large-scale, well-annotated datasets with representative features of the data distribution are the basis for data-driven learning to build more accurate and generalizable models. ImageNet provides a very comprehensive database for the pre-learning phase of object detection and image segmentation problems [26]. Since data collection is challenging and high-quality annotation is expensive, there is not a large-scale annotated medical image dataset equivalent to ImageNet. Therefore, the majority of deep learning models for healthcare applications are developed from the extensive and generalized feature representations derived from natural images.

Leveraging this principle and following the same conceptual process with similar motivations, models developed using DM images can undergo a process of transfer learning to adeptly acquire and fine-tune the specific features of DBT images. In this case, the adaptation is more fine-grained, given the images equal medical domain.

DM images and 2D individual slices extracted from DBT volumes share evident features, as it can be noticed from the Figure 4.1, which reports a CBIS-DDSM mammography image and a BCS-DBT tomosynthesis slice.

However, in these two datasets specifically, DM and DBT also present significant distinctions which add complexity to the transfer learning process, as we will further detail.

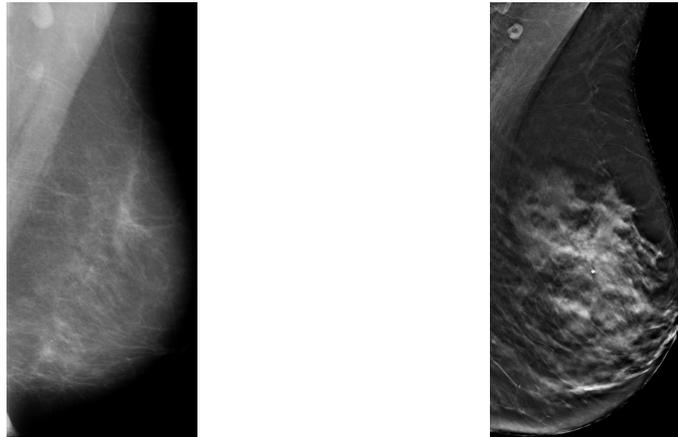


Figure 4.1: Comparison between a DM and a DBT slice

The transfer learning optimized network is reported to outperform conventional approaches specialized to DBT in terms of performance. Considering that there are far larger databases and clinical archives for DM images compared to DBT, this strategy is of significant importance.

A rigorous image pre-processing routine is instituted to adapt the images for compatibility with CNNs, recognizing that medical images, despite their standardized acquisition protocols, may exhibit variations which can potentially confound deep learning models. This adaptation is essential to mitigate non-diagnostic factors, as the non-uniform image laterality and presence of non-informative regions, which can disproportionately influence the model training efficacy.

Moreover, in this phase, we endeavor to reconcile the structural and feature-related discrepancies between the two datasets, BCS-DBT and CBIS-DDSM.

Employing a Facebook Research object detection model, Detectron, as a foundation, with custom implementation adaptations to the specific setting, we initially train it on pre-processed mammography images. Subsequently, we perform fine-tuning on the DBT dataset.

The performance evaluation of the algorithm is two-sided, incorporating both the conventional Average Precision (AP) metric at 50% of Intersection Over Union (IOU) threshold and a tailored criteria specifically designed for our specific research context.

Parallel to our primary methodology, we pursue an alternative strategy, in the attempt of addressing the limitations presented by employing a highly complex model such as Detectron on the restricted available dataset, acknowledging the complexities of tailoring this sophisticated architecture to fit our specific setting - predominantly, challenging implementation requirements and a significant computational demand.

This second approach focuses on a two-phase strategy, involving a first classification stage applied on lesioned and non-lesioned tissue patches extracted exclusively from the DBT dataset, followed by a lesion localization step performed on full images.

This method aims at simplifying and customizing the modeling process for the task of lesion detection targeted to the screening application, in order to develop an interpretable and tailored strategy that offers greater control over the implementation.

This shift from full-image-based to patch-based analysis involves training a binary classifier to distinguish between lesion-containing and lesion-free patches. The classifier is then deployed across a grid of patches derived from the entire DBT images, utilizing Class Activation Mapping (CAM) to identify lesion locations.

This comprehensive approach represents a multifaceted effort to exploit the capabilities of deep learning methods to enhance the precision and reliability of tumor lesion detection applied to breast imaging.

4.1. Approach 1: Detectron

In the first approach we perform lesion detection on full images using Facebook Research object detection model Detectron as a backbone, with a two-stage strategy, composed of the pre-training on mammography images and the fine-tuning on tomosynthesis images.

4.1.1. Data Preparation and Image Pre-Processing

BCS-DBT

Data Preparation

DBT volumes are acquired in DICOM format. Firstly, using the `pydicom` package, we read each volume and convert it into a pixel array of three dimensions, where the first dimension is the single slice height, the second dimension is the single slice width and the third dimension is the depth of the volume, in terms of slice numbers. The images are in grayscale, thus we have a single channel. The values of the pixels range from 0 - black to 255 - white.

As a second step, we check if the view information provided among the metadata coincides with the image laterality intercepted from taking the maximum value obtained summing up the two sides pixel values. A DBT slice, indeed, always presents a black band on the opposite side with respect to the breast imaged.

Subsequently, we rescale the intensity of pixel values to fit into the range `[low, high]`, with `low` and `high` obtained through the formulas:

$$\text{low} = \frac{2 * \text{window_center} - \text{window_width}}{2}$$

$$\text{high} = \frac{2 * \text{window_center} + \text{window_width}}{2}$$

Here, `window_center` and `window_width` are two pieces of metadata typically associated to DICOM images, obtained through the access to specific DICOM tags. These parameters suggest the proper displaying of medical images, providing the optimal brightness and contrast settings, which facilitate radiologists, and therefore deep learning models, to interpret the images and identify potential lesions. In details, the window center represents the center of the window, typically in Hounsfield (HU) units for computed tomography images or in arbitrary units for other modalities, and determines the brightness of the image. The window width defines the range of intensity values that should be mapped to the display, controlling the contrast of the image.

This process of image enhancement is standard for DBT images acquisition.

The next step consists in the extraction of the slice which, according to the information previously mentioned, represents the most evident location of the lesion across the full stack.

The procedure results in a dataset with the 2D DBT slices of interest and the associated relevant information for subsequent pre-processing and analysis. This information is the image laterality and the values $(x_{min}, y_{min}, x_{max}, y_{max})$, individuating the bounding box.

Image Pre-Processing

The image pre-processing procedure consists of image cropping and flipping.

To ensure compatibility with CNNs, we remove the non-informative black band on the opposite side with respect to the breast, which does not confer diagnostic value.

The process is conducted leveraging the connected components of the image. In details, we crop the image on the relevant content, that is the breast tissue, storing the resulting image pixel array. Accessing to its shape, then, we retain the updated image width and height values, which will be used in the subsequent stages of the workflow, specifically during the training loop.

Moreover, the images are flipped when necessary to maintain a consistent pattern across the dataset; in particular, we ensure that all the slices adhere to a uniform left orientation. We recall indeed that each study, although with some exceptions, contains the patient's two breasts DBT exams, acquired from CC and MLO views. Therefore, images will present heterogeneous lateralities. This standardization is crucial for eliminating biases related to image laterality and enhancing the model ability to learn from the data effectively. In fact, CNN learning is advantaged from input images with uniform patterns. Each transformation applied to the images is correspondingly mapped to the bounding boxes to maintain annotation consistency.

By addressing these non-diagnostic factors through the pre-processing, the images are optimized for training; this improves the model ability to learn from the diagnostic features within the images.

CBIS-DDSM

Data Preparation

The extraction process of DICOM images from the CBIS-DDSM dataset tracks the pipeline established for the BCS-DBT dataset.

In CBIS-DDSM, the `window_center` and `window_width` parameters are not explicitly available among the DICOM metadata. Therefore, these parameters are derived utilizing the minimum and maximum pixel intensity values, directly available among the DICOM

information, as follows:

$$\text{window_center} = \frac{\text{min_pixel_value} - \text{max_pixel_value}}{2}$$

$$\text{window_width} = \text{max_pixel_value} - \text{min_pixel_value}$$

This method enables the application of the image enhancement procedure consistently across both the BCS-DBT and CBIS-DDSM datasets, thus facilitating the effective implementation of the transfer learning strategy.

The CBIS-DDSM dataset presents a structural distinction from the BCS-DBT. It is organized according to the membership to the categories "full mammography images", "cropped images", and "ROI mask images".

Generally, each full image should be associated with one or more cropped images focusing on the lesion and the related binary masks. The mask segments the lesion area, assigning it pixel values of 255 - white, in contrast with the rest of the image, which is set to pixel values of 0 - black.

However, the delineation between these categories is not always clear-cut, as mask images can be found both in the "ROI mask images" category, alongside lesion cropped images, or within the "cropped" category.

Furthermore, numerous images contain multiple lesions, and the details regarding their association are not consistently specified. More specific data related to the lesions are stored in separate files, with a categorization based on lesion type, i.e. masses and calcifications, but the correspondence between images and information is not easily retrievable.

In order to obtain the annotations required to perform the object detection task we need to extract bounding boxes from the images. We have, therefore, to locate the mask images associated with each full image and define, in terms of pixels coordinates, the lesion location.

The absence of a systematic data organization structure requires a methodical approach to identify the corresponding masks for each full image. The identification process is based on the analysis of black pixel counts within the images, and necessitates the establishment of suitable thresholds for both the sum of pixel values and the identification of black pixels, due to the variability in lesion sizes, and the occasional inconsistency in the value associated to these black pixels, mainly attributable to variations in the image quality.

Throughout this procedure, multiple checks must be established to account for instances of either full images lacking an associated mask image or to ensure the inclusion of all the relevant mask images associated to a full image. To guarantee accurate pairing between images and their corresponding masks, we rely on verifying the dimensional match between

them. This step is necessary because, due to occasional data entry mistakes, the filenames may not always correctly indicate the association.

As a result of the standardization process, we obtain a total of 3014 DM images, including different views, with 3477 lesions annotated.

Upon locating each mask image, the next phase is to extract the minimum and maximum coordinates of the pixels valued at 255, indicating the lesion, across both axes. These coordinates serve to the construction of a rectangular bounding box identifying the lesion, recorded in the standard format $(x_{min}, y_{min}, x_{max}, y_{max})$, ensuring consistency with the BCS-DBT lesion annotation.

Image Pre-Processing

The image pre-processing procedure applied to the CBIS-DDSM dataset conceptually mirrors the one of the DBT images, coherently with the necessity of standardizing the dataset to match the BCS-DBT structure, in order to make it feasible to perform transfer learning from mammography to tomosynthesis images.

However, adapting the established pipeline to the CBIS-DDSM proves to be significantly more challenging due to the quality of the CBIS-DDSM data. Originally captured as film-screen mammographies and later digitized, these images inherently contain elements of noise not present in the DBT images. This noise and the digitization artifacts compromise the effectiveness of conventional methods for identifying the non-informative regions and the image laterality, and, in the meantime, if not addressed, could substantially impair the CNN performance.

Regarding the black bands cropping step, contrary to the straightforward application to the DBT images, the same connected components analysis is inadequate for CBIS-DDSM images. These digitized screen-film mammographies, indeed, often include textual annotations indicating the view and other examination details, which appear as clusters of white pixels and white bands on the image sides resulting from the scanning process. These noisy elements complicate the model learning process and the standard pre-processing steps.

Similarly, the standard procedure for extracting the image laterality is ineffective due to these peculiarities.

To address these challenges, a tailored approach is implemented, involving the exclusion of the top and bottom 5% of pixels in the connected component analysis to mitigate the impact of noise and scanning artifacts, in order to perform black bands cropping. Likewise, to determine the image laterality for flipping, we examine its edges excluding the left and right 5% of pixels.

By adopting this strategy, we are able to effectively pre-process the great majority of

the CBIS-DDSM images, ensuring that the dataset can be utilized efficiently for model training.

This adaptation underscores the necessity of flexible and problem-specific pre-processing techniques when dealing with the unique characteristics of open-source datasets, not always registered with a structured and precise method.

4.1.2. Model

Underlying Theory: Detectron

Detectron, developed by Facebook Research, is a cutting-edge object detection structure of models that leverages the power of deep learning through the application of CNNs. This framework is designed to efficiently detect, classify, and localize objects within images by employing a variety of state-of-the-art models, including Faster R-CNN, Mask R-CNN, and RetinaNet, each tailored to address specific challenges within the object detection domain.

The core mechanism of Detectron centers on integrating and optimizing CNNs for the task of identifying the presence and location of objects in digital images.

Detectron incorporates the Faster R-CNN architecture, which innovatively combines the process of generating region proposals directly within the CNN framework. This is achieved through a Region Proposal Network (RPN) that shares convolutional features across the entire image, allowing for the simultaneous proposal of candidate object locations and classification of these objects. This integration facilitates a streamlined and efficient detection process, by merging the steps of object proposal and detection into a unified network.

Building upon the capabilities of Faster R-CNN, Detectron also employs Mask R-CNN to extend its functionalities to instance segmentation tasks. Mask R-CNN adds a branch for predicting segmentation masks for each ROI, alongside the existing branches for object classification and bounding box regression. This enables precise pixel-level segmentation, identifying not just the objects but also delineating their exact shapes within the image. The segmentation branch utilizes a Fully Convolutional Network (FCN) applied to each ROI, producing a binary mask that outlines the object.

To address the challenge of class imbalance in object detection, Detectron includes RetinaNet, which introduces the focal loss function. This innovative loss function is designed to focus the learning process on hard-to-classify examples by reducing the relative loss for easy-to-classify negatives, thus ensuring that the model does not become overwhelmed by the abundance of easy negatives during training. RetinaNet leverages a Feature Pyramid Network (FPN) as its backbone, enabling the detection of objects at various scales by

providing a rich, multi-scale feature hierarchy.

Detectron architecture is highly modular, allowing for extensive experimentation with different network configurations, backbone architectures, and training techniques. The framework supports various backbone networks, such as ResNet and ResNeXt, for feature extraction, enabling researchers to tailor the object detection system to their specific requirements.

The comprehensive integration of advanced neural network architectures allows Detectron to be a robust, versatile and powerful tool for a broad spectrum of computer vision applications involving object detection tasks.

Training

The strategy applied in our study involves two phases: the first consists of pre-training the model on the mammographies dataset, the CBIS-DDSM; the second one involves fine-tuning it on the tomosynthesis dataset, the BCS-DBT.

We employed the "Detectron2" framework, specifically leveraging the "Faster R-CNN R 101 FPN 3x" architecture, to automate the identification of tumor lesions within both mammography and tomosynthesis images. The choice is motivated by its robust performance in various object detection tasks. The backbone of the model is a ResNet-101 integrated with a FPN, a combination allowing for effective hierarchical feature extraction across multiple scales, which is important for detecting lesions of varying sizes and appearances.

The CBIS-DDSM dataset is divided into a training set and a test set, with allocations of 90% and 10%, respectively.

Given Detectron extensive data requirements and considered that the effective evaluation of the model must be done, as the final step, on the BCS-DBT dataset, we use the test set as validation set, dedicating the entirety of the remaining data to model training.

The initial phase involves reformatting the images and the associated metadata into a structure compatible with Detectron input requirements.

Specifically, Detectron framework necessitates supplying images, annotations detailing the bounding box coordinates in the format $(x_{min}, y_{min}, x_{max}, y_{max})$, which identify the lesions, and the width and height of the single images. These dimensions were beforehand extracted during the image pre-processing stage, thus avoiding the need for individual image file access and reading during later phases, which would significantly increase the processing time. Detectron, indeed, requires to instantiate a "Dataset" object comprising all the mentioned information before the training loop begins. Without pre-saved image dimensions, this step would necessitate a redundant double reading of all the images -

once at the "Dataset" object initialization and again during the training iterations, consistently slowing down the overall process. By storing the image dimensions in advance, we efficiently streamline the workflow, ensuring a more expedited and resource-effective model training phase.

To ensure a solid starting point for training, we initialize the model weights from the "Detectron2 model zoo", benefiting from a base pre-trained on the COCO dataset, a large image dataset specifically designed for object detection and segmentation.

Given the nature of our task - lesion localization on DBT images - we configure the model to focus on a single class of objects, that we identify through the "lesion" label.

Training is approached with the objective of optimizing the performances while, meanwhile, accommodating the constraints typical of medical image analysis and taking into account the computational demands proper of Detectron complex architecture.

We configure the data loader with two parallel workers to streamline the process of feeding data into the model. We employ aspect ratio grouping and filtering of empty annotations to ensure data integrity.

Input data undergo augmentation and pre-processing to improve the model ability to generalize across diverse presentations of the breast lesions. We apply random vertical flip and crop of 95% of the original image size, aimed at introducing variability into the training process and reducing noise, without losing critical diagnostic features. It is important to note that we refrain from flipping images horizontally, coherently with our previous effort of standardizing the orientation of all the breasts to the same side, to provide the model with a consistent pattern, which assists it in learning the informative features.

The data augmentation procedure is intentionally simplified to manage the substantial memory requirements, acknowledging that each transformation applied to the images must also be mapped to the bounding boxes. Furthermore, given the intensity-based characterization of lesions, transformations affecting brightness, color, and contrast are excluded to prevent potential model confusion. This limits the scope of applicable augmentation.

The validation set remains non-augmented to ensure a consistent and reliable model training tracking and performance evaluation.

Additionally, images are resized to ensure consistency in input dimensions, with a maximum size of 1333 pixels for both training and validation sets and a variable minimum size, ranging from 640 to 800 pixels, for training set, while validation set is resized at a fixed minimum size of 800 pixels. Images are converted to BGR format, aligning with the common structure of models pre-trained on natural images, from which weights initialization is performed.

The batch size is set to 2 images per training batch, striking a balance between the

computational demand and the necessity for meaningful gradient updates during back-propagation.

In the training configuration, the Stochastic Gradient Descent (SGD) optimizer is chosen for its simplicity and efficiency in navigating the high-dimensionality of deep neural networks.

We set the base learning rate at 0.00025. This value is selected through rigorous testing to identify a rate that facilitates steady convergence towards the optimal model weights. The learning rate scheduler we adopt, the "WarmupMultiStepLR", is step-based, and provides gradual learning rate adjustments. This strategy involves an initial warm-up phase over 1000 iterations to stabilize the learning process, where the learning rate is increased with the value of the base learning rate as a top threshold. Subsequently, the learning rate is reduced at predetermined intervals, namely each 5000 iterations. This adaptive approach ensures a balanced exploration of the solution space.

The model, thus defined, is trained for 30,000 iterations, during which we monitor the validation set losses.

After an intermediate assessment of the model performance, executed computing the AP metric at 50% of IOU threshold on the validation set, we proceed with the successive phase, the fine-tuning of the model on the target dataset BCS-DBT.

Therefore, we initialize the model weights using those derived from the training on the CBIS-DDSM dataset, and we repeat the training for 9,000 iterations.

In this step, the BCS-DBT dataset is divided into 90% for training, with 20% of that allocated for validation, and the remaining 10%, reserved for testing. The available validation set is utilized during the model development phase, while the test set is exclusively used for the final evaluation of the model performance.

4.1.3. Performance Metrics and Model Selection

As evaluation techniques, we first use the AP metric at 50% of IOU threshold, which provides a quantifiable measure of the object detector accuracy.

This metric calculates the precision, by considering detections to be true positives if they have an IOU of 50% or more with the corresponding ground truth object. The IOU is a measure of the overlap between the predicted bounding box and the ground truth bounding box; the value of 50% indicates that at least half of the predicted bounding box overlaps with the ground truth one. This threshold is commonly used in object detection challenges as it strikes a balance between being permissive enough to recognize approximately correct predictions and strict enough to ensure predictions to be reasonably accurate.

The AP at 50% of IOU threshold provides insights into how well the model can identify and localize objects within the images, reflecting the precision, namely the model ability to return more true positives and fewer false positives. In practical terms, a high AP at 50% of IOU suggests that the model is effective at detecting objects with a reasonable degree of localization accuracy. It indicates that the detected bounding boxes align well with the actual lesions in the images.

As a second evaluation procedure, we introduce a tailored metric devised considering the specific context of the research.

Our goal, indeed, is to design a lesion detection pipeline utilizing open-source data. The model is developed with the ambition of being extended through the future integration of consistent, high-quality, proprietary datasets. Most importantly, we acknowledge its specific application to the screening phase, with the aim of optimizing the process, which demands significant time and human resources.

Given this context, the model serves as a preliminary analysis tool, implying that the ultimate assessment of the findings will always be conducted by a radiologist. The application aims at efficiently directing the radiologists' attention to the areas of potential presence of lesions, and which, consequently, require further examination.

The newly designed evaluation metric is constructed to quantify the model ability to highlight the existence of a true lesion in the full image, facilitating a streamlined review process for the radiologists.

We consider all the predicted bounding boxes per single image: such predictions are deemed accurate if the intersection over the minimum area between the predicted bounding box and the true bounding box exceeds 50%, and inaccurate otherwise.

For instances where the model does not produce any prediction, the metric value in correspondence of the specific image, is clearly set to 0.

Considering that our dataset primarily consists of images featuring a single lesion, as outlined in the "Data and Datasets" section, and the model typically outputs non-overlapping predictions, the average of this metric, across all the test samples, serves as an indicator of the model recall. This measure, indeed, effectively quantifies the ratio of true positive detections over the combined count of true positives and false negatives, providing a clear view of the model ability to correctly identify lesions.

In addition, to monitor the occurrence of false positives, we count the number of incorrectly predicted lesions per image, defined as the total predicted bounding boxes minus the exact predictions. Averaging this value, we obtain the mean amount of false positives per image.

Overall, the goal is for the model to maximize recall while controlling the average num-

ber of false positives, thus reducing potential confusion and avoiding counterproductive results.

This approach simplifies the assessment process, allowing for a more intuitive understanding of the model performance in detecting lesions, which we consider of crucial importance for the practical application of the research findings. We intentionally streamline the evaluation strategy, to favor interpretability.

Detectron outputs the bounding boxes along with their associated confidence scores. This information enables us to determine the ideal confidence threshold for our model, which is the minimum confidence level at which a prediction is considered valid.

We evaluate the recall and average number of false positives for various confidence threshold levels using the validation set.

At the standard confidence threshold of 0.5, the model yields a recall value of 0.692, and maintains a notably low average of false positives, recorded as 0.346.

Aiming to increase the recall, we therefore systematically explore confidence thresholds ranging from 0.05 to 0.5, excluded, in increments of 0.05, to identify the optimal setting. Analyzing results for the value 0 is useless since it would result in accepting the entire set of model predictions.

Figure 4.2 reports the recall and corresponding average false positives count obtained at the mentioned threshold values.

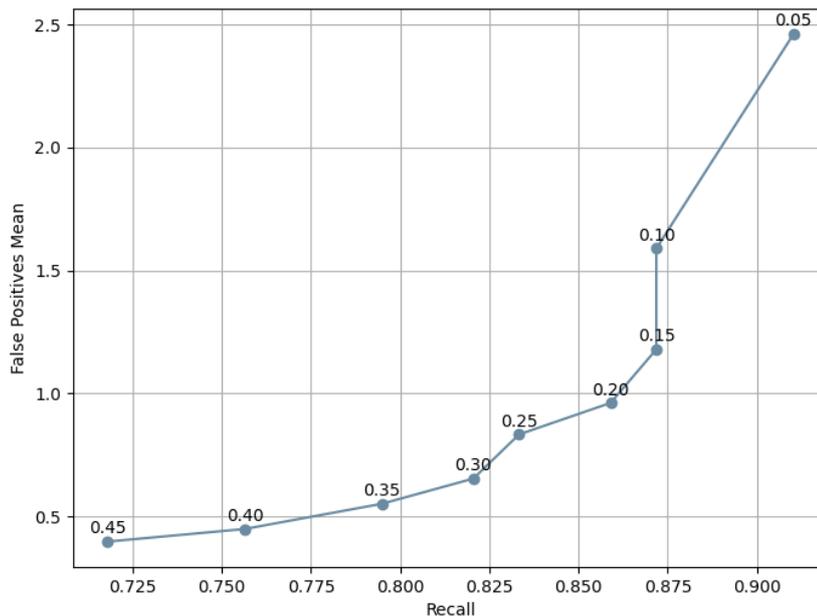


Figure 4.2: Recall versus false positives mean at different confidence thresholds

We settle on a threshold of 0.2 as it balances achieving a high recall while maintaining the mean number of false positives restricted.

The same threshold is assigned to the confidence level selected for the computation of the AP at 50% of IOU metric.

4.2. Approach 2: Patches Model

To address the challenge of the limited image availability, we explore a second approach, where we develop a classification model that differentiates between lesioned and non-lesioned tissue patches, followed by a phase of lesion localization, performed applying the newly developed classification model on a grid of patches generated from the full images. This methodology not only aims to mitigate the scarcity of data, by augmenting the number of usable images through the generation of patches from the full DBT slices, but also streamlines and tailors the modeling process.

As previously underlined, indeed, Detectron intricate architecture presents challenges in terms of customization, control, and extensibility. In response, we implement a strategy that prioritizes interpretability and manageability, essential aspects in a medical environment. Focusing on a patch-based approach facilitates the development of models that are more adaptable to our specific setting and also yield results that are clear and meaningful. Through this approach, we aim to develop a tool which provides insights easily understandable and actionable by medical professionals, ensuring a seamless integration into the clinical workflow.

4.2.1. Data Preparation and Image Pre-Processing

In this second approach, we exclusively utilize the extracted tomosynthesis slices of the BCS-DBT dataset.

Our aim is to construct a dataset composed of image patches categorized into two distinct groups: lesioned patches, derived from areas of breast tissue exhibiting lesions, and non-lesioned patches, originating from normal breast tissue.

To generate the "lesion" class patches, it is critical to take into consideration the variability of the lesion dimensions observed in the dataset.

Despite a limited number of samples, there is a significant variation in the lesion sizes, and this introduces additional complexity to the task.

The distribution of the lesion dimensions within the BCS-DBT dataset, reported in Figure 4.3, illustrates this point.

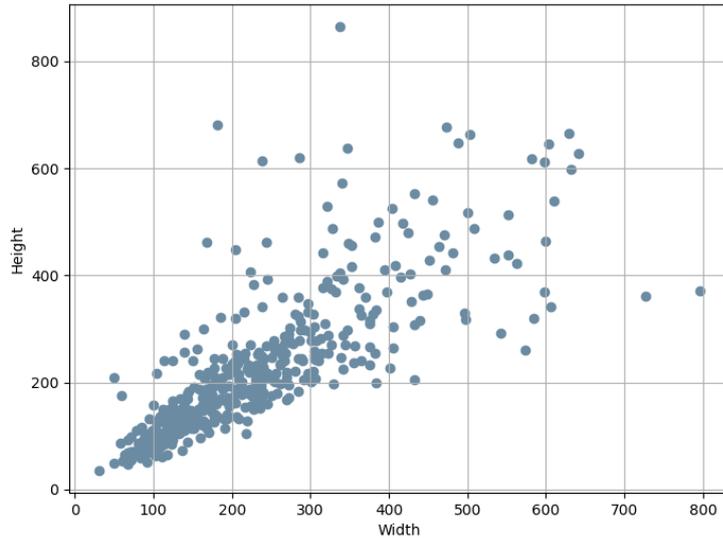


Figure 4.3: Lesion dimensions distribution

From the analysis on lesion shape values, we obtain a median height and width of 201 and 212 pixels, respectively.

A straightforward approach might suggest to crop the images to a standard size of 224x224, around the lesion bounding box center, as this input size is common across several models developed on patches. However, further data investigation reveals that 40% of lesions exceeds the value of 224 pixels in height, and 46% in width. This discrepancy indicates that, while a substantial portion of the lesions aligns closely to the standard 224x224 size, directly cropping all the lesions to these dimensions could lead to a significant information loss and may potentially compromise the localization accuracy of larger lesions. In addition, training the model exclusively on central portions of lesions, obtained through a fixed size cropping, could hinder its ability to generalize across the dataset varied cases.

Therefore, we opt for the development of a cropping strategy that accounts for the variable dimensions of the lesions. In details, we design an algorithm which, given an image and a corresponding lesion bounding box, automatically extracts the maximal number of 224x224 - a selected standard dimension - patches on the basis of the larger dimension of the lesion itself.

However, we must recall that the classification model we aim to develop is functional to the subsequent localization phase. Therefore, it is important to consider that dividing the full images in a grid may result in patches where the lesion is not centrally located, possibly leading to its division across corners or edges of multiple patches. This suggests

that, if the classification model is trained exclusively on patches centered on the lesion, it may struggle to correctly identify the lesion locations within the entire images.

To provide contextual information surrounding the lesion, addressing this potential complexity, we extract an additional number of patches per lesion, depending on its dimensions, by cropping the full image around the lesion periphery.

The steps we apply to perform such extraction are described in Algorithm 4.1.

Algorithm 4.1 Patches Extraction Algorithm

- 1: **Input:** Image, bounding_box_center, boundix_box_sizes,
 - 2: **Set:** $l = 224$, dimension of the patches,
 - 3: Select the maximum between the height and width of the bounding box.
 $size = \max(\text{boundix_box_sizes})$,
 - 4: $Depth = \lceil \frac{size-1}{l} \rceil$,
 - 5: Initialize `center_list` as a list with only the center of the bounding box.
 - 6: **for** i in Depth **do**
 - 7: Initialize `angle_list` as an empty list,
 - 8: **for** `center` in `center_list` **do**
 - 9: $angles = [(x_{center} - \frac{l}{2}, y_{center} + \frac{l}{2}), (x_{center} - \frac{l}{2}, y_{center} - \frac{l}{2}),$
 $(x_{center} + \frac{l}{2}, y_{center} - \frac{l}{2}), (x_{center} + \frac{l}{2}, y_{center} + \frac{l}{2})]$
 - 10: Add these angles to the `angle_list`.
 - 11: **end for**
 - 12: Add the angles of the `angle_list` in the `center_list`,
 - 13: Avoid repetition using a `set`.
 - 14: **end for**
 - 15: **for** `center` in `center_list` **do**
 - 16: Create the patch, cutting the Image given the angles computed from `center` as in line 9,
 - 17: Save the patch as `.png` file.
 - 18: **end for**
-

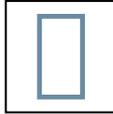
The idea is to implement a sort of recursive approach, selecting, for the different levels of depth, the angles as the centers of the next depth value patches.

The Depth defined in line 4 is used to understand how many patches are needed to fully cover the bounding box. Indeed we can compute the necessary number of patches as $(Depth - 1)^2 + (Depth)^2$.

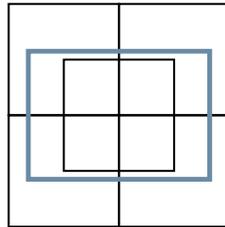
To better understand the rational behind the algorithm we provide a visual representation.

$\text{size} < 224$

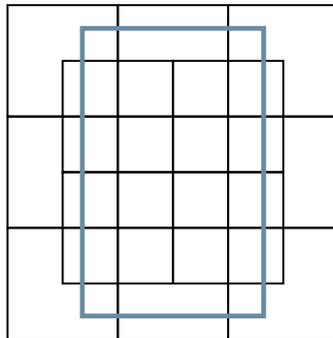
Depth = 1

 $224 < \text{size} < 448$

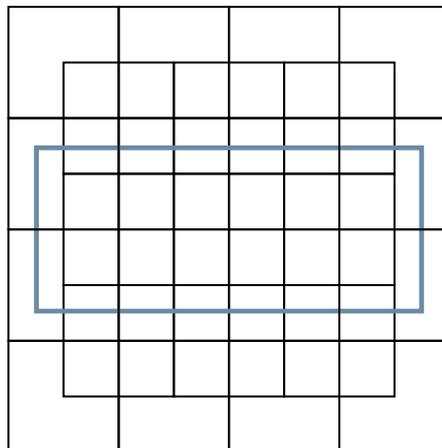
Depth = 2

 $448 < \text{size} < 672$

Depth = 3

 $672 < \text{size} < 896$

Depth = 4



This process, in addition, allows to augment the number of lesioned patches from 435 to 1637. Notice that this operation is not always successful; for example, when the lesion is on the borders of the images some of the context patches are not extracted.

We underline that, in the whole procedure, we maintain the same train-test split established in the first approach. Specifically, we obtain the training set "lesion" class patches from the first approach training set full lesioned images and the test set "lesion" class patches from the first approach test set full lesioned images. Maintaining this consistency is decisive for enabling a fair and systematic comparison between the two approaches, which is one of the objectives of the study.

For the generation of the "no lesion" class patches, we must consider that, in the subsequent phase of lesion localization, data will be inherently imbalanced, namely the number of lesioned patches obtained from the grid division of a single full image will be by far less than the number of non-lesioned patches. We recall, indeed, that each full image presents a single lesion, except for a singular case, and that, generally, the lesion extent is limited. To accurately reflect this imbalance in our model, we evaluate the proportion of lesioned patches with respect to non-lesioned patches - in this analysis we exclude the non-informative background patches - within the set of test images, computing the median imbalance ratio. Notice that this step is performed approximating the identification of the lesioned patches on the basis of the portion of the ground truth bounding box area included in each patch generated from the grid division. As a result, we obtain that the median imbalance ratio is across 1 over 20.

This examination suggests to sample a substantial number of non-lesioned patches to tackle the problem imbalance. Therefore, to obtain a number of normal breast tissue patches that enables us to implement a dataset imbalance of 1 over 20 we restore a part of the normal DBT volumes within the BCS-DBT.

From each volume, we extract the central slice through the same procedure explained for the first approach and we apply the usual pre-processing procedure, cropping out the non-informative regions and adjusting the laterality, as previously detailed.

We extract patches of standard dimensions 224x224, consistently to what has been performed for "lesioned" patches, to obtain a standard input size, required for the training of a deep learning model.

Having established the non-lesioned patches extraction specifications, we proceed by sampling 150 random points per normal image and, taking each of them as a reference corner point, we extract a patch of size 224x224. Then, if the sum of black pixels is less than 90% of the pixel number of the image, meaning that the extracted patch is not a background non-informative part but a normal breast tissue one, we retain it, otherwise we get rid of

it. We recall, indeed, that, even after cropping out the non-informative black band of the DBT, due to the shape of the breast, we still have some non-informative areas of black pixels.

The non-lesioned patches resulting from this procedure are then divided into train and test set with the usual proportions of, respectively, 90% and 10%, used across the whole study.

We obtain a train set composed of a total of 42,227 non-lesioned patches and 1,503 lesioned patches, which implements nearly a 1 over 28 imbalance ratio and a test set containing 4,692 non-lesioned patches and 134 lesioned patches, which results into a nearly 1 over 35 imbalance ratio. We intentionally extract more "no lesion" patches than necessary because the sampled points are random, and we are not able to guarantee the success of the generation of every patch at each single iteration. Our aim, yet, is to complete the extraction process in a single run.

From the set of images obtained through this procedure, we build multiple datasets characterized by varying degrees of imbalance between "lesion" and "no lesion" patches. These datasets will be critical in the development of a classification model able to effectively manage the challenge of distinguishing between lesioned and non-lesioned tissues under the imbalance condition it will encounter during the later phase of lesion localization.

4.2.2. Model

Classification

From the data preparation and pre-processing procedures previously delineated, we obtain a dataset already divided into training and test subsets, exhibiting imbalance ratios of 1 over 28 and 1 over 35, respectively, for lesioned versus non-lesioned patches. We recall that our prior analysis identified an average imbalance ratio of approximately 1 over 20 in the localization phase. Consequently, to align the imbalance ratio across both training and test sets to 1 over 20, we adjust the quantity of non-lesioned patches in our dataset accordingly.

In addition, we allocate 20% of the training set for validation purposes, coherently with the proportion defined for the first approach, preserving the 1 over 20 imbalance ratio through the stratification with respect to the train labels.

Data Augmentation

The first step of our methodology involves data augmentation, specifically targeting the lesioned patches due to their scarcity compared to the non-lesioned patches. The objective, here, is to enrich the dataset with a broader spectrum of lesioned patch samples, thereby

enhancing the model capability to learn lesion features.

Given the complexity of constructing a model capable of distinguishing between lesioned and non-lesioned patches within a training set characterized by a 1 over 20 imbalance ratio, we employ several augmentation techniques.

In line with the first approach, we avoid adjustments to brightness, color and contrast to preserve the intrinsic characterization of the lesioned tissue. Additionally, we refrain from applying transformations that significantly alter the shape of the patches, to prevent extensive interpolated areas, which could introduce noise and potentially mislead the model. The augmentation strategies we implement include:

- random rotations within a range of $[0, 180]$ degrees;
- random zoom up to 20%;
- random horizontal and vertical flips.

This procedure is controlled to obtain an adjusted imbalance ratio of 1 over 5 for the training patches.

We remark that the augmentations are applied ensuring the inclusion of the original lesioned patches alongside the augmented ones, to avoid relying uniquely on artificially constructed images.

As a result, the training dataset comprises all the original non-lesioned patches, along with the original and augmented lesioned patches, which allow to attain the target imbalance ratio.

The validation set, conversely, retains its original composition, presenting a distribution of 1 lesioned patch every 20 non-lesioned ones. Consequently to numerous experiments across different configurations, we conclude that, in order to maximize the model recall, on such an imbalanced dataset, it is most convenient to provide it with the actual proportions of lesioned versus non-lesioned patches, in the validation phase. This is particularly important as, besides monitoring the model learning, we also depend on the validation set for implementing learning rate adjustments and hyperparameters selection.

Additionally, we resize all the patches, including those designated for testing, to a resolution of $224 \times 224 \times 3$. This standardization ensures compatibility with the original image cropping performed in the pre-processing phase and the standard input size of the model we intend to apply, beyond the pre-trained models we will employ for weights initialization. It is, indeed, common practice to leverage on pre-trained models, which are typically obtained from images with the dimension of $256 \times 256 \times 3$ - the "3" indicates the three channels (RGB) representing color images. Weights initialization from pre-trained models is widely adopted in scenarios where the dataset is relatively small, to avoid training models

from scratch. This concept will be further explored in the following "Training" section of our work.

Finally, both the training and validation data are shuffled to ensure a randomized distribution with respect to the classes, enabling a more robust learning process.

Training

The classification task is conducted using a ConvNeXt model, adhering to a transfer learning methodology. Specifically, we employ the simplest version of the architecture, providing the minor number of parameters, the ConvNeXt-Tiny, to comply with the scarcity of data.

Underlying Theory: ConvNeXt

The ConvNeXt architecture signifies a leap in the evolution of Convolutional Neural Networks (ConvNets). Rooted in the tried-and-tested framework of ResNets, ConvNeXt transitions through a transformative process inspired by the Swin Transformer structure to enhance its own design. This innovation has given rise to the ConvNeXt models, which are entirely constituted from conventional ConvNet modules but exhibit competitive performance, matching and in some scenarios, surpassing, that of Vision Transformers.

The ConvNeXt model, particularly the Tiny version, has been trained on ImageNet at a resolution of 224x224, presenting itself as a purely convolutional model. It abandons the typical ResNet layout, adopting a Transformer-like architecture without fully embracing the Transformer philosophy. Instead, it maintains the pure ConvNet approach, leading to notable achievements in accuracy and scalability.

This model holds the unique distinction of outperforming Swin Transformers in certain applications, notably image classification on ImageNet, object detection on COCO dataset, and segmentation on ADE20K. Its construction does not diverge from standard ConvNet modules, yet it manages to incorporate elements from the Transformer design space, such as Layer Normalization (LN) and the use of Gaussian Error Linear Unit (GELU) nonlinearities, which contribute significantly to its enhanced performance.

The streamlined design of ConvNeXt enables an efficient feature extraction and, in conjunction with its scalable nature, makes the Tiny version suitable for tasks where a balance between performance and computational efficiency is critical [18].

Initially, the convolutional layers are kept frozen, while the top layers are customized and trained to adapt to our specific task. This architecture includes a Global Average Pooling layer following the base model output, and a Dense layer with 1,024 units, employing the linear unit (ReLU) activation function and incorporating an L2 kernel regularizer with a rate of 0.01 to mitigate overfitting. Finally, a Dense output layer with 2 neurons employing a Softmax activation function is added for the binary classification.

Model weights are initialized using pre-trained weights from ImageNet.

Weights Initialization on ImageNet

ImageNet is the most extensive annotated natural image dataset currently available. It boasts over 1.2 million images, categorized into 1,000 distinct object classes, with each class containing upwards of a thousand training images.

This comprehensive dataset is meticulously organized according to the WordNet hierarchy, emphasizing a wide variety of object categories such as "sea snake", "sandwich", "vase" and "leopard". The sheer volume and diversity of ImageNet, coupled with its significant intra-class variation - characterized by occlusions, partial views, and small object sizes - offer a robust foundation for data-driven learning systems. These systems are designed to not only accurately classify instances within the dataset but also generalize effectively to new, unseen data, making ImageNet a benchmark for object recognition challenges [19]. While natural images and medical images differ substantially in their characteristics, the success of conventional image descriptors and ImageNet pre-trained CNNs in medical image analysis cannot be understated. Techniques such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) have found applications in detecting and segmenting objects within medical imagery. More recently, CNNs pre-trained on ImageNet have shown promising results in identifying and detecting pathologies in chest X-rays and Computer Tomography (CT) scans, outperforming traditional methods [26]. Our methodology builds upon this foundation, utilizing weights from ImageNet pre-trained CNN models as the starting point for the model training on our medical image dataset. This approach takes advantage of the pre-existing knowledge embedded in these models, gained from the extensive and diverse set of natural images within ImageNet, to enhance the feature extraction and recognition capabilities in the context of medical images, which reveals particularly convenient in our condition of data scarcity.

The ConvNeXt model is designed to directly process input images with pixel values ranging from 0 to 255, since it has an internal layer accounting for normalization. Therefore, we do not rescale the images in advance.

Given the relatively small size of our dataset, we select a batch size of 16.

The model is trained over 50 epochs, with the number of training iterations and validation steps determined by dividing the training and validation set sizes by the batch size, respectively.

For loss computation, we employ the focal loss function, specifically designed to address the class imbalance, by modifying the standard cross-entropy loss such that it down-

weights the loss assigned to well-classified examples. It is mathematically defined as:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t),$$

where:

- p_t is the model estimated probability for the class with the true label t ;
- α_t is a balancing factor applied to the t -th class, often set to be inversely proportional to the class frequency;
- γ is the focusing parameter that smoothly adjusts the rate at which easy examples are down-weighted. When γ is set to 0, focal loss is equivalent to cross-entropy loss.

The focal loss function ensures that the contribution of easy-to-classify examples to the overall loss is small, allowing the model to focus on the hard, misclassified examples, improving the sensitivity to the less frequent, yet most important, "lesion" category.

The parameters α and γ are selected based on dataset characteristics and common practice. Specifically, α is set to 0.05, which aligns with the actual frequency of the "lesion" class, being the 0.05% of the "no lesion" class. The parameter γ is assigned to a value of 2, adhering to the default choice.

The training procedure incorporates several optimization strategies.

The Adam optimizer is selected for its efficiency, with an initial base learning rate of 0.001.

Additionally, a customized learning rate scheduling strategy is implemented, designed to dynamically adjust the learning rate on the basis of the validation loss computed at the end of each training epoch. We start by monitoring the validation loss, maintaining a record of its best value observed. If the validation loss fails to decrease for a consecutive number of epochs defined by a patience parameter, set to 5, indicating no improvement in the model performance, the learning rate is reduced by a predetermined factor, 0.1 in our implementation. This adjustment is performed until the learning rate reaches a minimum threshold of 10^{-6} , preventing it from dropping to values not allowing the model to learn effectively. Simultaneously, the reduction is applied restoring to the model best weights, thus maintaining its optimal performance level.

An early-stopping mechanism is also employed, halting the training if the validation loss fails to decrease by at least 10^{-7} over 7 consecutive epochs.

Throughout the training, special attention is given to monitoring and optimizing the recall and false positives, which are crucial metrics in the medical imaging context. The model, indeed, should be able to detect true positives while, in the meantime, maintaining predictive power by not classifying all instances as positives.

After this initial training phase, the entire network is unfrozen to allow for fine-tuning of all the layers.

This second phase uses the weights obtained from the initial training as a starting point, with a reduced learning rate of 0.0001 to refine the model performance further.

We retain the same model structure and hyperparameters. Moreover, the same training criteria, including the learning rate scheduler and early stopping are applied during this fine-tuning phase, to ensure a consistent optimization process.

Localization

After obtaining the classification model, the localization phase is implemented through a Gradient-weighted Class Activation Mapping (Grad-CAM) sliding window approach.

The following steps are applied to each lesioned full image of the BCS-DBT test set defined at the beginning of the chapter.

First, we resize the image to ensure that its dimensions are divisible by the patch size 224x224, applying nearest neighbour interpolation. The scaling process enables us to obtain, from each full image, an integer number of patches matching the input dimensions expected by the classification model.

This step involves adapting also the bounding box coordinates to the resized image dimensions, to ensure that the lesion location is accurately represented in the resized image. The resized image is then divided into patches of fixed dimensions 224x224.

The classification model is employed to obtain the predictions for each of these patches, resized to a resolution of 224x224x3 for compatibility with the expected input, excluding the non-informative ones. Indeed, for each patch generated from the grid division we count the number of black pixels and if this number is higher than 90% - coherently with the criterion applied previously in the thesis, we automatically label the patch as non-lesioned. The analysis, therefore, is restricted to the breast tissue regions.

For every patch obtaining a prediction score for the "lesion" class exceeding a threshold - initially set at the standard value of 0.5 - a heatmap is computed following the method outlined below.

To adapt the CNN model for Grad-CAM, we eliminate the activation function of the last fully connected layer. This adjustment is necessary for generating gradient-based class activation maps.

We calculate the gradients of the predicted class with respect to the output feature map of the last convolutional layer, merging them to highlight the areas of the image most influential for the model prediction. The heatmap generated from this process individuates the patch regions which most reveal a lesion, providing an insight into the areas where

the model concentrates its attention.

Upon calculating the heatmaps for each patch, we aggregate them to form a comprehensive visualization overlaid on the original full image.

This visualization technique allows us to assess the model performance in localizing lesions within the broader context of the entire image. The final output includes both the original image, with the ground truth bounding box drawn for reference, and the aggregated heatmap.

To determine the predicted bounding boxes based on the heatmap, we first convert it into a binary matrix, by setting to 1 all the pixels contributing to the lesion classification, identified by the heatmap values higher than 0.

Subsequently, contours are identified within this binary representation, serving as the outlines for the predicted lesions. For each contour detected, a rectangular bounding box is determined, encompassing its area.

The pixel coordinates of these bounding boxes, obtained on the basis of the heatmap dimensions, are subsequently scaled back to match to the resized image shape, from which we initially generated patches.

The collection of these adjusted bounding boxes represents the final prediction of the lesion locations within the image.

4.2.3. Performance Metrics and Model Selection

To evaluate the performance of the second approach, we employ the custom metric, already designed in the first approach, for the specific context of this analysis.

The determination of a patch being classified as lesioned depends on whether the model prediction score, for that label, exceeds a predetermined threshold, which we initially set at the standard value of 0.5.

The computation of the recall and average count of false positives, on the validation set, using this threshold, yields values of 0.717 and 3.212, respectively.

This result aligns with the criteria applied in the development of the classification model, characterized by a strong inclination towards the maximization of the recall, although at the expense of a consistent number of false positives.

Consequently, in this case, we examine the confidence thresholds spanning from 0.55 to 1, excluded, in 0.05 increments, to select the value that optimally balances the recall with the control of the average false positives value. Notice that, since we accept model predictions presenting scores strictly higher than the threshold, checking results for the value 1 is meaningless.

Figure 4.4 reports the recall and corresponding false positives average obtained at the mentioned threshold values.

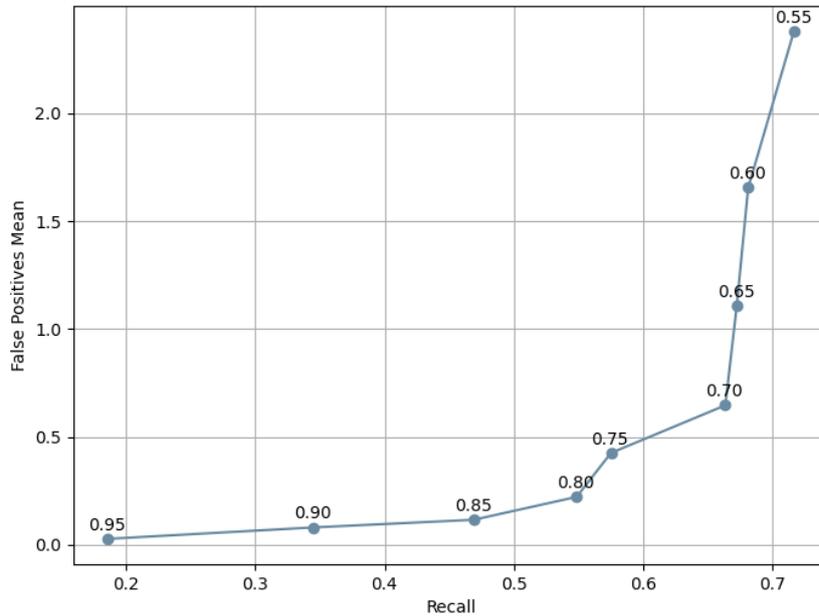


Figure 4.4: Recall versus false positives mean at different confidence thresholds

We opt for a threshold value of 0.6, which enables us to maintain a sufficient high recall rate, while keeping the average number of false positives in check. In fact, in our assessment, we consider the occurrence of fewer than 2 false positives per image to be an acceptable value, reflecting our primary commitment to maximize the model sensitivity. This perspective is based on the principle that, in the medical framework, the error of missing a positive detection - Type I error - is considered more significant with respect to the occurrence of an additional false positive.

It is worth noticing that, in this approach, the localization model is constructed on top of a classification model. Therefore, applying the first metric considered across the thesis, the AP at 50% of IOU threshold, specifically applied to the object detection models, is not straightforward. This is because we lack direct confidence scores for each predicted bounding box and we instead rely on a threshold score from the lesion classification phase, with confidence inferred from the heatmap intensity values.

Concerning the instances where the model outputs incorrect bounding boxes for the lesions, we compute a metric to analyze the distance between the bounding boxes erroneously predicted and the ground truth ones. The metric is computed as the Euclidean

distance of the centers of the two bounding boxes, but normalizing the value with respect to the x-axis and y-axis, taking into account the image shape, as expressed by the Equation 4.1.

$$\text{distance} = \sqrt{\left(\frac{x_{\text{center_pred_bb}} - x_{\text{center_true_bb}}}{\text{image_width}}\right)^2 + \left(\frac{y_{\text{center_pred_bb}} - y_{\text{center_true_bb}}}{\text{image_height}}\right)^2}, \quad (4.1)$$

where `center_pred_bb` and `center_true_bb` refer, respectively, to the incorrectly predicted and ground truth bounding boxes centers, while `image_height` and `image_width` are the dimensions of the image.

5 | Evaluation

In this section we present and discuss the results of our work, reporting the values of the evaluation of the metrics detailed in the "Model" chapter, on both the developed approaches.

5.1. Results

Approach 1: Detectron

This approach is evaluated through two distinct evaluation methodologies. Here, we present the outcomes of these metrics, in correspondence of the selected value for the confidence threshold at which a given prediction is considered valid. We recall that this threshold, tuned on the validation set, was established at a value of 0.2.

The AP at 50% of IOU criterion is 0.554.

According to the specific metric outlined in the previous section, tailored to the research question, the recall is 0.837 and the average false positives count per image is 0.977.

Approach 2: Patches Model

The second approach is assessed exclusively through the metric we specifically devised. The results are obtained by setting the acceptance threshold for a "lesion" class prediction at 0.6, a value tuned on the validation set.

The outcome for the recall is 0.627, and the average false positives count per image is 1.581.

5.2. Discussion

In this section, we delve into the discussion of the results obtained through our investigation, analyzing and reporting the different instances of predictions the two models yield.

The first approach implements a pipeline based on a sophisticated object detection architecture, which demonstrates promising capabilities in identifying lesions accurately.

Based on the tailored metric designed for the thesis specific application, we developed a model able to correctly detect the vast majority of the lesions within the test set. The model sensitivity, which measures the detection accuracy of the relevant instances - the lesions, is altogether high. This outcome is particularly impressive given the domain complexity, coupled with the limited scope of the dataset, and the broad range of lesion sizes and types it presents.

It is worth underlining that this performance is realized alongside a remarkably minimal number of false positives, averaging at less than one per image. Considering the relative size of lesions compared to the overall image area, this rate is deemed far acceptable.

The interpretation of these results requires the consideration of two factors.

Firstly, the dataset used in the study is significantly restricted with respect to those typically employed in the most successful lesion detection applications. This limitation presents a considerable challenge, given the data-intensive nature of deep learning models, and particularly Detectron complex architecture, which is characterized by a multitude of parameters.

Secondly, some constraints on the computational resources necessitated a straightforward approach to model implementation. The inability to incorporate advanced data augmentation techniques or more elaborated learning rate schedules impacted the model potential performance. Furthermore, the complex structure of the model layers was fixed, eliminating the possibility of adjustments in the model definition.

Despite these limitations, the results remain creditably satisfactory, especially if framed within the precise context of our application. A key goal of this research is to design a pipeline leading to an application capable of supporting radiologists in the screening process. This can be achieved enhancing the model ability to clearly highlight the potential presence of lesions, by maintaining an high recall rate and minimizing the false positives count.

Remarkably, the reported accomplishments were achieved exclusively through the use of open-source data, a notable departure from the norm in the literature of breast lesion detection research, whose models often rely on meticulously annotated, consistent and proprietary datasets.

Although the BCS-DBT dataset stands as the unique publicly accessible tomosynthesis dataset, most researchers, even if they lack access to proper proprietary datasets, still benefit from transfer learning on extensive semi-private mammography datasets. These datasets, which are more widespread with respect to the tomosynthesis ones, are generally exclusive to well-established research institutions and their acquisition frequently requires

a considerable amount of time.

All considered, the image pre-processing pipeline, combined with the transfer learning strategy, enabled by the systematic standardization of the intricate CBIS-DDSM mammography dataset, yields the desired outcomes.

In our evaluation, the AP at 50% of IOU threshold serves as an alternative method, and it quantifies the proportion of accurately predicted lesions with respect to the whole amount of the detected lesions. A decrease in the performance under this metric offers insight into the model occasional propensity towards false positive predictions, as a drawback. This, however, is anticipated in the context of lesion detection tasks.

Nonetheless, it is crucial to stress out, once again, that within the specific objectives of this research recall emerges as the primary metric of interest.

In addition, the 50% of IOU threshold, measuring the bounding boxes overlapping, is presumably stringent for the intended application of assisting radiologists in streamlining the screening process. Achieving the exact lesion delineation is less critical than the ability to reliably identify a higher number of true lesions.

Furthermore, the aforementioned significant variability in lesion dimensions can lead to instances where the overlap is less than 50%, not due lesion misidentification, but rather because the lesion size is atypically more extended compared to the standards commonly encountered by the model, and the prediction results in the correct location but has a smaller area with respect to the ground truth object.

Additionally, it is important to acknowledge that annotations may not always be precise; they are sometimes larger than the actual lesion to accommodate the required rectangular contours of the bounding box, further complicating the accurate measurement of the model performance, if based on the AP at 50% of IOU.

Taking these considerations into account and observing the still minimal average number of true positives per image, the model performance is correctly assessed through the tailored metric we designed, aligning to the objective of enhancing the efficiency and accuracy of lesion detection in a medical context, where the ability to identify as many true positives as possible is of paramount importance, even if it requires to accept a slightly higher rate of false positives.

To illustrate the practical implications of the model performance, we present two illustrative case studies that encapsulate both the successes and the limitations of our proposed solution.

The Figure 5.1 shows an example of lesion correctly and accurately identified by the algorithm. We note that, although if the instance is of very small dimensions, the model is capable of detecting its presence. This is indicative of its scalability.

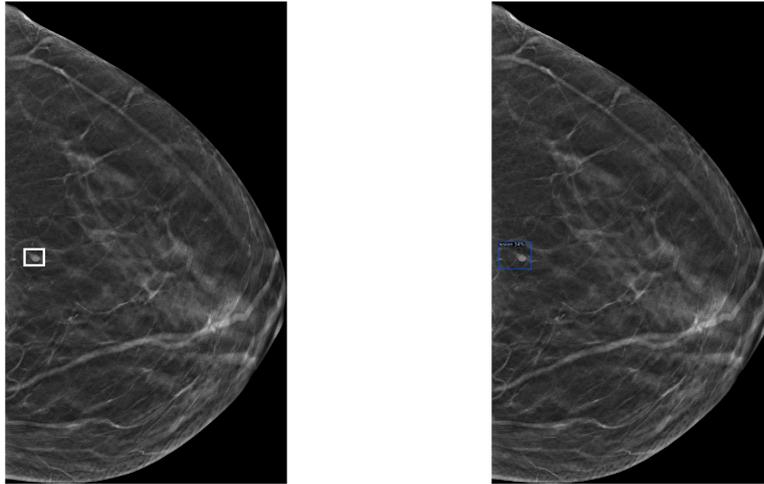


Figure 5.1: Ground truth and prediction of a lesion correctly detected

Conversely, the Figure 5.2 represents a particularly challenging case, where the model struggle to correctly identify the lesion. In this instance, the actual lesion is accurately identified but the model is simultaneously misled by two hyperintense areas with irregular shapes adjacent to the lesion.

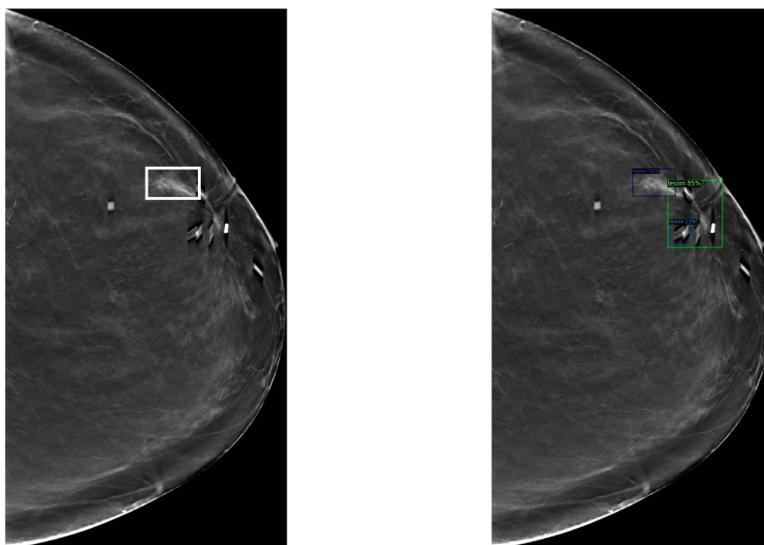


Figure 5.2: Ground truth and prediction of a lesion partially incorrectly detected

It is worth noticing that this serves as a representative example of the most consistent type of errors made by the model. Typically, the partial inaccuracies occur in the presence of atypical examples.

The performance observed in the study, which is completely satisfactory but still not comparable to the state-of-the-art lesion detection algorithms, can largely be attributed to the constraints posed by the limited size of the dataset available to perform the analysis. This is further justified by noticing that the exploration of various training configurations did not yield improved results.

With only 431 images and 435 lesions for training, validation and testing, the complexity and data requirements of Detectron model raised several challenges.

The low-quality and characteristics of the open-source mammography images used for transfer learning also contributed to the slight uncertainty in the model performance, given their divergent features compared to DBT images. But still, the transfer learning approach allowed to leverage the complex Detectron architecture on the limited dataset. Despite the inherent challenges, we were able to find a balance between computational complexity and model implementation, leading to outcomes that exceeded our initial projections.

In the second approach, we aimed at developing a model able to perform lesion detection through a two-phase process applied on image patches.

This strategy offered us greater control over the implementation.

Firstly, it allowed us to apply transformations directly to the images without the need to adjust the annotations correspondingly. This flexibility enabled us to experiment with a broader range of augmentations, exploring various configurations to achieve the optimal setup.

Additionally, the approach afforded us the opportunity to experiment with different models and learning strategies, improving our control over the implementation process.

A key phase of this approach was the design of an effective procedure for generating the input for the model, which led to the development of a patch extraction algorithm. This algorithm is easily generalizable to diverse datasets and broadly applicable to tasks requiring the classification of tumor lesions beyond breast cancer.

It potentially accepts input images with lesions of any size, producing patch extractions which adjust to varying dimensions while incorporating the portion of context necessary to train classification models that effectively transition to localization tasks.

This strategy addresses the question of how to crop lesioned patches, preventing from the information loss deriving from direct cropping, and circumventing the necessity of the

resize application for the alignment to the standard input sizes required by CNN models. As a last thing, it is worth noticing that the test set we relied on to extract the metrics, comprising only 43 images, reflected a wide variability in lesion characteristics, which likely influenced the observed outcomes.

In conclusion, training a model with such a constrained number of images and pronounced class imbalance presents significant challenges. The proposed solution, developed through extensive experimentation on diverse settings with the aim of figuring out the most suitable model features for the task at hand, demonstrated acceptable outcomes. The achievements are particularly significant considering the data open-source origin and the intended application of the developed strategies, specifically, to assist radiologists in the screening process.

To further illustrate these findings, we provide two visual examples of both successful and unsuccessful lesion detections through the heatmaps generated in the localization phase. These heatmaps, overlaid on the original DBT images, offer a visual representation of the model diagnostic precision, highlighting the identified lesioned areas.

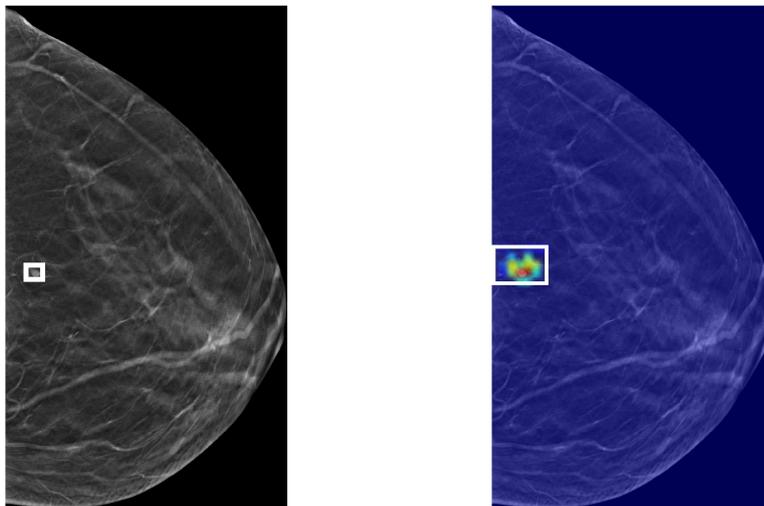


Figure 5.3: Ground truth and prediction of a lesion correctly detected

Figure 5.3 reports the second approach model correct prediction on the same test image used to provide an example of correct prediction for the first approach model. As previously remarked, this is a particularly challenging case, which, however, the model handles accurately.

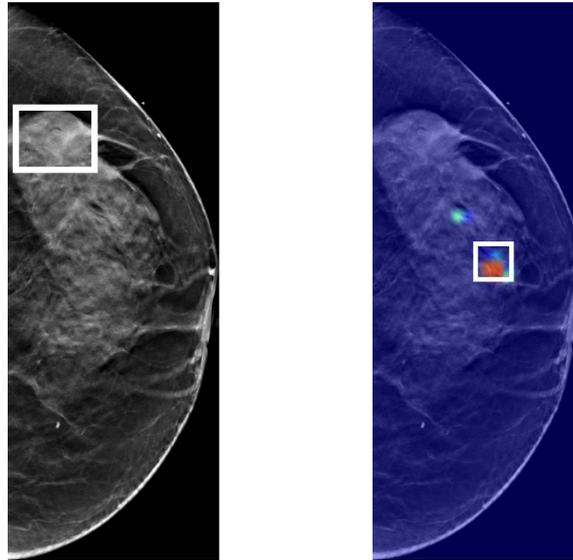


Figure 5.4: Ground truth and prediction of a lesion incorrectly detected

Figure 5.4, instead, reports an instance of incorrect prediction. We note that this case was instead correctly predicted by the model developed in the first approach. This is in line with the improved performance it demonstrated, compared to the second approach.

From the presented outputs emerges the high interpretability of the results that this approach yields. The predictions obtained for each patch are visualized through a heatmap that highlights the areas, within the single patch, that most significantly contribute to the prediction. Aggregating these patch-level heatmaps we obtain a comprehensive view of the specific lesioned areas, according to the model predictions. Such an intuitive application proves extremely valuable in assisting radiologists to the streamlined review of numerous exams, enabling them to focus with near-pixel accuracy directly on the areas of interest.

We extend the analysis to a finer examination of the incorrect predictions generated by the model, focusing specifically on the nature of these errors to gain deeper insights into the model performance.

We note that in 9 cases the model fails to produce any predicted bounding box.

A closer inspection reveals that these instances often involve particularly challenging samples, characterized by extremely small lesions, which exhibit features distinctly different from the standards on which the model is trained.

Even if the model was able to correctly identify a lesion instance with limited dimensions,

as shown in Figure 5.4, it generally struggles with this cases. The Figure 5.5 shows an example of a lesion with very limited dimensions not captured by any model prediction.

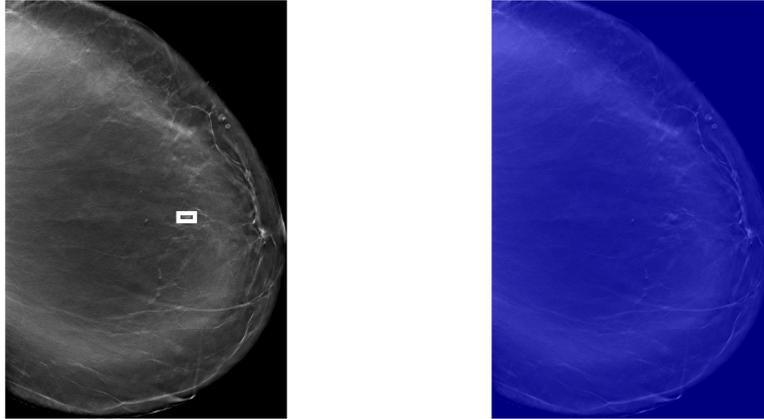


Figure 5.5: Ground truth and prediction of a not detected lesion

Shifting the focus to the instances where the model outputs incorrect lesion predictions, we report the results obtained through the computation of the distance 4.1 between the incorrectly predicted and the ground truth bounding boxes centers. An insight into this distance distribution is provided in Figure 5.6.

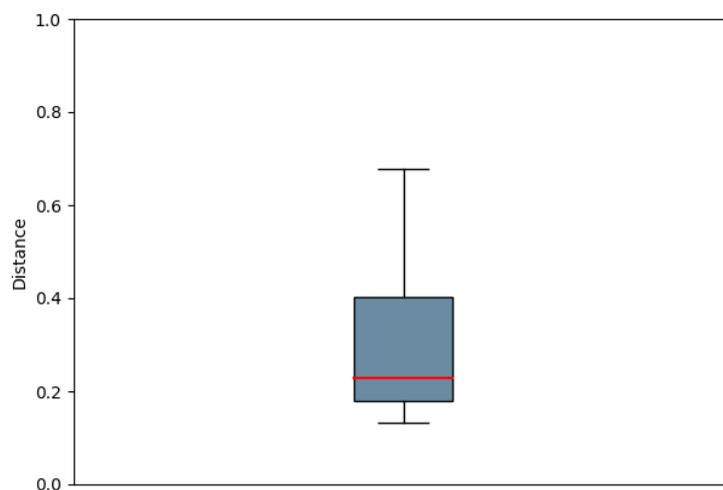


Figure 5.6: Distribution of the normalized center distance between incorrectly predicted and ground truth bounding boxes

The boxplot in Figure 5.6 shows that the median of the distribution is nearly the value of 0.2. The interquartile range (IQR) is narrow, indicating that most of the data points are clustered around the median. The model incorrect predictions are not drastically off-target, suggesting that even in its missteps, the model outputs are not significantly distant from the actual lesion locations.

In conclusion, the second two-phased approach, which enables increased control over the model design, demonstrates the capability of detecting lesions with a satisfactory level of recall. This is most evident in light of the challenges faced, ranging from class imbalance to variability in lesion sizes and shapes which, coupled with the limited number of images, complicates the model ability to generalize across different lesion characteristics effectively. To achieve an acceptable rate of recall, in this case, we must tolerate a marginally higher rate of false positives.

In Figure 5.7, the model successfully identifies the lesion; however, it also generates a second false-positive prediction, alongside the accurate detection.

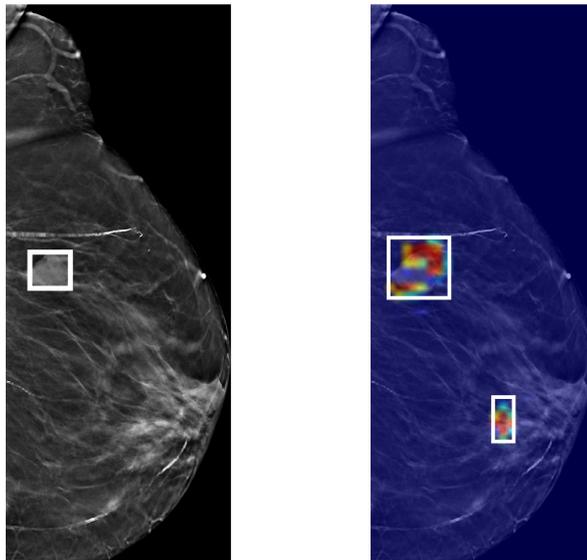


Figure 5.7: Ground truth and prediction of a lesion correctly detected, with a false positive

False positives represents a recurrent challenge in the domain of lesion detection algorithms. Within this research context, the emphasis is often placed the model sensitivity for the identification of lesions. Yet, as in real-world data, even of widely spread diseases,

negative instances prevail over positive instances, the incidence of false positives emerges as a considerable hurdle in models application to practical settings.

Nevertheless, caution in medical scenarios is critical. Specifically, the capacity to detect an additional lesion outweighs the risk of a higher false positives rate.

6 | Conclusion

In the thesis, we delved into the application of deep learning for the automatic detection of breast cancer lesions in DBT images on open-source data, evaluating two distinct methodologies through a comprehensive process spanning from data acquisition and pre-processing to model assessment.

We developed an end-to-end pipeline for lesion identification exploring and comparing the two proposed alternative strategies: a direct full image object detection technique taking advantage of transfer learning from mammographies and a patch-level two-phase method based on a classification model extended to a localization task, to leverage a broader dataset.

Our investigation encompassed various pre-processing and modeling techniques, including the use of Detectron for object detection and CNNs complemented by activation maps for lesioned patches identification.

The first approach, consisting of a unified lesion detection method leveraging transfer learning from mammography data, provides an example of the application of transfer learning from mammography to tomosynthesis images and, concurrently, points out the challenges inherent in working in the intricate setting of open-source data, a common starting point for many researchers, particularly in the medical field, where data access is restricted due to privacy concerns and datasets are inherently unbalanced.

In response to the limited availability of the BCS-DBT images and the extensive parameters requirements of the Detectron algorithm, we adopted a transfer learning approach from a related domain to enhance the model learning capabilities. Although it enabled us to take advantage of a more consistent dataset, this strategy introduced a new hurdle due to the insufficient quality of the publicly available mammography dataset we employed. Furthermore, exploiting this dataset necessitated the unraveling of its disorganized structure. A significant effort was undertaken to standardize the CBIS-DDSM, the principal publicly available DM dataset, and to adapt it to the format of the BCS-DBT, the unique open-source DBT dataset. This required to establish a consistent correspondence among the full images and associated mask images files, and to extract the lesion annotations. This preliminary adjustment requirement aligns with the underlined inherent nature of

open-source data, often lacking the quality and systematic organization proper of privately acquired datasets, which are collected with uniform and standardized procedures.

In the second approach, designed as a two-phased methodology, based on a DBT patches classification model, followed by a localization step on full images, two significant outcomes emerge.

Firstly, we developed a patch extraction algorithm adept at handling images presenting lesions of any size. This solution tackles the prevalent challenges in developing models for lesion detection that employ a two-stage process of classification and localization - a methodology frequently adopted by researchers in this field.

One significant hurdle is managing the variability in lesion sizes: by dynamically adapting to lesion dimension ranges, the algorithm avoids cropping the images to the standard input size required by the models or, alternatively, directly cropping the images on their original dimensions to then resizing them. These procedures, indeed, can distort the image aspect ratio and introduce noise through interpolation.

In addition, the question of extracting an adequate quantity of contextual tissue, surrounding the lesion, is addressed as well. This phase, which itself depends on lesion dimensions, is critical for training classification models which are expected to generalize to detection tasks.

This strategy has transformed our approach, enabling us to cope with the high variability of lesions on the limited number of images, and can be extended to datasets beyond just the breast domain.

The second relevant aspect of this approach is that it provides radiologists with an intuitively understandable output, a not only valuable but necessary condition for medical applications.

The efficacy of deep learning models, while proven, hinges on their ability to offer results that radiologists can interpret and trust. With the employment of heatmaps as the final output of the whole detection pipeline, we are able to overlay on each full DBT slice the areas and the related lesion features that most strongly contributed to the prediction.

This validation strategy builds confidence among physicians and, meanwhile, provides researchers with immediate directions for model improvement. Indeed, it serves as a direct tool for analyzing model inaccuracies. For instance, if the model demonstrated limited efficacy in identifying smaller lesions, it could be the case to incorporate additional microcalcification samples into the training set.

In conclusion, the thesis presents two viable solutions for lesion detection on DBT, which can be employed according to the specific requirements of future research efforts in the domain.

The first approach, based on the object detection model Detectron, demonstrated the most impressive performance. These preliminary results, obtained leveraging exclusively open-source data, indicate that significant achievements have the potential to be attained, through the inclusion of consistent and quality data in the study. Nevertheless, this method poses challenges due to the limited control over the model implementation. The complexity and computational intensity of Detectron architecture restrict the ease of modifications and experimentation with diverse settings, outlining the necessity to strike a balance between computational efficiency and problem-specific implementation requirements.

To circumvent these limitations, a second strategy was devised, built upon a preliminary analysis of image patches. This approach decomposes the task into a dual-phase process, significantly increasing the volume of usable images extracted from the relatively restricted dataset. The results obtained through this method were achieved employing exclusively 431 DBT slices. This suggests a promising pathway for future fine-tuning of the approach, particularly if in combination with the application of transfer learning from DM patches and the introduction of extensive accurate data.

The comprehensive methodologies developed and validated through the research exhibit distinct merits and generate reliable outcomes. Their comparison enables the verification of results consistency, ensuring confidence in their applicability.

Differently to the prevailing presence in the literature of lesion detection algorithms utilizing privately curated datasets, the study achieved respectable outcomes entirely leveraging open-source data, developing two valuable approaches. These methods facilitate comparative analysis and can be adapted to meet the specific requirements of the investigation objectives.

The research addresses the balance between performance and interpretability, still bearing in mind that further enhancements in both aspects can be achieved with the access to more consistent and higher-quality data.

Remarkably, the analysis succeeded in yielding viable solutions under non-trivial circumstances, characterized by the limited data availability and simultaneous significant variability in the lesion nature.

Through extensive experimentation to determine the optimal configuration, the development of this pipeline aims to establish a reliable foundation for future research in this domain, in a complete open-source framework.

Achieving a 84% of success rate without relying on consistent, high-quality data underscores the potential of the approaches pursued and emphasizes the critical need of more extensive and curated datasets in the field of medical imaging analysis for breast cancer.

6.1. Limitations

The limitations encountered throughout the development of the thesis predominantly stem from the scarcity of data, a challenge that affects various aspects of our implementation. Detectron, along with other models designed for lesion detection, is complex and data-hungry. State-of-the-art algorithms in this field typically rely on datasets comprising thousands of images, a stark contrast to the 431 images at our disposal. This issue is exacerbated by the fact that most researches in this area leverage private datasets, whether directly of DBT or DM, to facilitate transfer learning.

In the first approach, the structure and image characteristics of the mammography dataset used for transfer learning introduced unavoidable complexities into the model learning process. The CBIS-DDSM, comprising digitized screen-film mammographies, consists of images with features significantly divergent from the higher-quality slices of the BCS-DBT tomosyntheses. This discrepancy is exacerbated by the low quality of the mammography images, which are marred by noise and artifacts.

In addition, it is worth to note that the CBIS-DDSM dataset adaptation process, detailed in the "Methods" section, is subject to errors due to the dataset chaotic nature, joint to the manual extraction of the bounding boxes, and the image pre-processing decisions based on arbitrary thresholds that may not uniformly apply across all the dataset instances. The correctness of this adaptation was verified through random sampling on a significant number of images, albeit not the entirety, indicating the potential scattered inaccuracies in the pre-processing procedure.

The task at hand, specifically, is further complicated by the minimal number of samples available for detecting both masses and architectural distortions, which can vary significantly in appearance.

Additionally, there is considerable variability in the lesion sizes, and this poses a substantial challenge, especially in our second approach.

Although the patch extraction algorithm we developed addressed most of the challenges deriving from the transition from classification to localization setting, the full image grid division necessary to actually perform detection still results in non-centered lesions, and, despite the inclusion of context patches mitigated the issue, expanding the dataset is crucial to effectively handle lesion diversity.

As detailed in the "Evaluation" section, through the analysis of the results, the model faced difficulties with micro-lesions. This scenario typically underlines the necessity for distinct algorithms tailored to specific lesion types - a strategy unfeasible for us due to the further reduction it would imply in our already limited dataset.

Furthermore, the second approach is prone to the false positives issue, which, as already

mentioned, is a well-documented challenge in the literature, and represented a concrete hurdle also in the development of the X-RAIS platform, from which the research is inspired.

Training a classification model on a dataset with a 1 over 20 imbalance ratio presented significant impediments. Simultaneously achieving an high recall rate and a low false positives count proved particularly hard. This concern is even more pronounced in the subsequent localization phase.

6.2. Further Developments

The primary path for further development, beyond the application of our model to more consistent, quality, private datasets, is to transition from 2D to 3D modeling. This shift would enable us to fully leverage the capabilities of the DBT, whose pseudo-3D nature offers a comprehensive view of the breast tissue.

Currently, the model processes individual slices extracted from the entire volumetric data, focusing on the 2D projections which, according to the radiologists' indication, contain the most evident sign of the lesion. The next step involves applying the model to every slice within a given examination, with the objective of generating predictions that include both the bounding box of the detected lesion and the corresponding slice index where it has been identified.

This advancement would enhance the model diagnostic precision, by enabling a more comprehensive analysis of the entire volume, and providing a multidimensional understanding of the spatial relationships and morphological features of the lesions across the breast tissue.

Moving to a 3D framework would align our efforts with the latest advancements in medical imaging technology, allowing for more sophisticated analyses that could potentially identify lesions undetectable in both the traditional 2D DM and the most relevant projection of the DBT volume.

This advancement entails leveraging the complete pseudo-3D examination and applying the trained models across all its slices. The objective is to generate a synthesized comprehensive prediction that aggregates single lesion detections at various depths within the volume.

Extending our model to the 3D setting represents a promising direction for future research, with the potential to furthermore advance breast cancer detection through deep learning solutions.

Bibliography

- [1] A. Aizenman, T. Drew, K. A. Ehinger, D. Georgian-Smith, and J. M. Wolfe. Comparing search patterns in digital breast tomosynthesis and full-field digital mammography: an eye tracking study. *Journal of Medical Imaging*, 4(4):045501–045501, 2017.
- [2] American College of Radiology. *BI-RADS® Mammography Reporting*, 2023. URL <https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/Mammography-Reporting.pdf>.
- [3] S. Astley, S. Connor, Y. Lim, C. Tate, H. Entwistle, J. Morris, S. Whiteside, J. Sergeant, M. Wilson, U. Beetles, et al. A comparison of image interpretation times in full field digital mammography and digital breast tomosynthesis. In *Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*, volume 8673, pages 189–196. SPIE, 2013.
- [4] J. Bai, R. Posner, T. Wang, C. Yang, and S. Nabavi. Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review. *Medical image analysis*, 71:102049, 2021.
- [5] A. Bhowmik and S. Eskreis-Winkler. Deep learning in breast imaging. *BJR/ Open*, 4(1):20210060, 2022.
- [6] M. Buda, A. Saha, R. Walsh, S. Ghate, N. Li, A. Świącicki, J. Y. Lo, and M. A. Mazurowski. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. *JAMA network open*, 4(8):e2119100–e2119100, 2021.
- [7] A. Chong, S. P. Weinstein, E. S. McDonald, and E. F. Conant. Digital breast tomosynthesis: concepts and clinical practice. *Radiology*, 292(1):1–14, 2019.
- [8] S. V. Destounis, R. Morgan, and A. Arieno. Screening for dense breasts: digital breast tomosynthesis. *American journal of roentgenology*, 204(2):261–264, 2015.
- [9] L. Duijm, M. Louwman, J. Groenewoud, L. Van De Poll-Franse, J. Fracheboud, and

- J. W. Coebergh. Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome. *British journal of cancer*, 100(6): 901–907, 2009.
- [10] European Parliamentary Research Service. The impact of the general data protection regulation (gdpr) on artificial intelligence. Study by the Panel for the Future of Science and Technology, June 2020. URL [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU\(2020\)641530_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf).
- [11] J. S. Greenberg, M. C. Javitt, J. Katzen, S. Michael, and A. E. Holland. Clinical performance metrics of 3d digital breast tomosynthesis compared with 2d digital mammography for breast cancer screening in community practice. *AJR Am J Roentgenol*, 203(3):687–693, 2014.
- [12] M. A. Helvie. Digital mammography imaging: breast tomosynthesis and advanced applications. *Radiologic Clinics*, 48(5):917–929, 2010.
- [13] S. Hussain, Y. Lafarga-Osuna, M. Ali, U. Naseem, M. Ahmed, and J. G. Tamez-Peña. Deep learning, radiomics and radiogenomics applications in the digital breast tomosynthesis: a systematic review. *BMC bioinformatics*, 24(1):401, 2023.
- [14] B. S. Idrees, G. Teng, A. Israr, H. Zaib, Y. Jamil, M. Bilal, S. Bashir, M. N. Khan, and Q. Wang. Comparison of whole blood and serum samples of breast cancer based on laser-induced breakdown spectroscopy with machine learning. *Biomedical Optics Express*, 14(6):2492–2509, 2023.
- [15] B. Johnson. Asymmetries in mammography. *Radiologic Technology*, 92(3):281M–298M, 2021.
- [16] K. Lång, I. Andersson, A. Rosso, A. Tingberg, P. Timberg, and S. Zackrisson. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the malmö breast tomosynthesis screening trial, a population-based study. *European radiology*, 26:184–190, 2016.
- [17] C. D. Lehman, R. D. Wellman, D. S. Buist, K. Kerlikowske, A. N. Tosteson, D. L. Miglioretti, B. C. S. Consortium, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11):1828–1837, 2015.
- [18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

- [19] L. Lu, Y. Zheng, G. Carneiro, and L. Yang, editors. *Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Large-Scale Datasets*. Advances in Computer Vision and Pattern Recognition. Springer International Publishing, illustrata edition, 2017. ISBN 9783319429984.
- [20] E. S. McDonald, A. Oustimov, S. P. Weinstein, M. B. Synnestvedt, M. Schnall, and E. F. Conant. Effectiveness of digital breast tomosynthesis compared with digital mammography: outcomes analysis from 3 years of breast cancer screening. *JAMA oncology*, 2(6):737–743, 2016.
- [21] K. J. Nam, B.-K. Han, E. S. Ko, J. S. Choi, E. Y. Ko, D. W. Jeong, and K. S. Choo. Comparison of full-field digital mammography and digital breast tomosynthesis in ultrasonography-detected breast cancers. *The Breast*, 24(5):649–655, 2015.
- [22] F. Pesapane, C. Trentin, F. Ferrari, G. Signorelli, P. Tantrige, M. Montesano, C. Ciccala, R. Virgoli, S. D’Acquisto, L. Nicosia, et al. Deep learning performance for detection and classification of microcalcifications on mammography. *European Radiology Experimental*, 7(1):69, 2023.
- [23] W. N. Price and I. G. Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019.
- [24] I. Sechopoulos, J. Teuwen, and R. Mann. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. In *Seminars in cancer biology*, volume 72, pages 214–225. Elsevier, 2021.
- [25] R. E. Sharpe Jr, S. Venkataraman, J. Phillips, V. Dialani, V. J. Fein-Zachary, S. Prakash, P. J. Slanetz, and T. S. Mehta. Increased cancer detection rate and variations in the recall rate resulting from implementation of 3d digital breast tomosynthesis into a population-based screening program. *Radiology*, 278(3):698–706, 2016.
- [26] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.
- [27] R. L. Siegel, A. N. Giaquinto, and A. Jemal. Cancer statistics, 2024. *CA Cancer J Clin*, 74(1):12–49, 2024.
- [28] A. Tetreault-Laroche, G. Misuraca, and F. Lipianez-Villanueva. Challenges and limits

- of an open source approach to artificial intelligence. Study PE 662.908, European Parliamentary Research Service, May 2021. URL [https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662908/IPOL_STU\(2021\)662908_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662908/IPOL_STU(2021)662908_EN.pdf).
- [29] University of Maryland Medical System. 3d mammography exams. URL <https://www.umms.org/umgccc/cancer-services/cancer-types/breast/diagnostic-treatment/3d-mammography-exams>.
- [30] M. G. Wallis, E. Moa, F. Zanca, K. Leifland, and M. Danielsson. Two-view and single-view tomosynthesis versus full-field digital mammography: high-resolution x-ray imaging observer study. *Radiology*, 262(3):788–796, 2012.
- [31] Y. Weerakkody, T. Manning, P. Lemos, et al. Breast imaging reporting and data system (bi-rads). *Radiopaedia.org*, 2010. URL <https://radiopaedia.org/articles/10003>.
- [32] D. Xiang, W. Cai, et al. Privacy protection and secondary use of health data: strategies and methods. *BioMed Research International*, 2021, 2021.
- [33] H. Zonderland and R. Smithuis. BI-RADS for Mammography and Ultrasound 2013. <https://radiologyassistant.nl/breast/bi-rads/bi-rads-for-mammography-and-ultrasound-2013>, 2013.

A | Appendix A

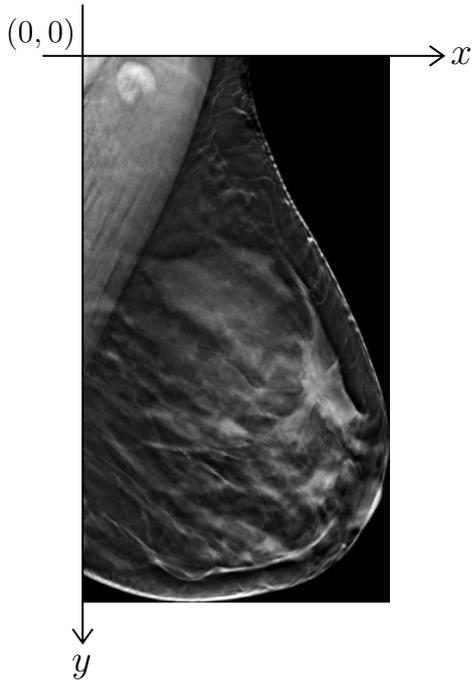


Figure A.1: Image axis orientation

List of Figures

2.1	Category A	9
2.2	Category B	9
2.3	Category C	9
2.4	Category D	9
2.5	DM (a.) and DBT (b.) acquisition techniques	12
4.1	Comparison between a DM and a DBT slice	29
4.2	Recall versus false positives mean at different confidence thresholds	40
4.3	Lesion dimensions distribution	42
4.4	Recall versus false positives mean at different confidence thresholds	53
5.1	Ground truth and prediction of a lesion correctly detected	58
5.2	Ground truth and prediction of a lesion partially incorrectly detected . . .	58
5.3	Ground truth and prediction of a lesion correctly detected	60
5.4	Ground truth and prediction of a lesion incorrectly detected	61
5.5	Ground truth and prediction of a not detected lesion	62
5.6	Distribution of the normalized center distance between incorrectly predicted and ground truth bounding boxes	62
5.7	Ground truth and prediction of a lesion correctly detected, with a false positive	63
A.1	Image axis orientation	75

