University of Brasilia at Gama – FGA/UnB
Biomedical Engineering Graduate Program

Explainable Artificial Intelligence Model for Mammogram Breast Cancer Classifiers

Geovanni Oliveira de Jesus

Advisor: Cristiano Jacques Miosso

University of Brasilia at Gama

Programa de
Pós-Graduação em
ENGENHARIA BIOMÉDICA

Explainable Artificial Intelligence Model For Mammogram Breast Cancer Classifiers

Geovanni Oliveira de Jesus

Advisor: Cristiano Jacques Miosso

Master Degree Thesis on
Biomedical Engineering

Publication: 180A/2023

Brasilia/DF, November of 2023

University of Brasilia at Gama

Graduate Program

Explainable Artificial Intelligence Model For Mammogram Breast Cancer Classifiers

Geovanni Oliveira de Jesus

Master Thesis submitted to the Biomedical Engineering Graduate Program, as a partial fulfillment of the requirements for the degree of Master in Biomedical Engineering

Approved by:

Cristiano Jacques Miosso

(Advisor)

Prof. Dr. Luciano Manhães de Andrade Filho

(External examiner)

Prof. Dr. Nilton Correia da Silva

(Internal examiner)

Brasilia/DF, November of 2023

Catalog Card

Jesus, Geovanni

Explainable Artificial Intelligence Model for Mammogram Breast Cancer Classifiers

Biomedical Engineering Graduate Program

[Distrito Federal], November of 2023.

56p., 210 × 297 mm (FGA/UnB Gama, Mestrado em Engenharia Biomédica, 2023).

Dissertação de Mestrado em Engenharia Biomédica, Faculdade UnB Gama, Programa de Pós-Graduação em Engenharia Biomédica.

1. Explainable Artificial Intelligence (XAI) 2. Deep Learning
3. Breast cancer 4. Classification
I. FGA UnB/UnB. II. Título (série)

Reference

Jesus, Geovanni (2023). Programa de Pós-Graduação em Engenharia Biomédica. Master thesis in Biomedical Engineering, Publication 180A/2023, Biomedical Engineering Graduate Program, University of Brasilia at Gama, Brasilia, DF, 56p.

Copyright

Autor: Geovanni Oliveira de Jesus

Título: Programa de Pós-Graduação em Engenharia Biomédica

Grau: Mestre

Ano: 2023

geovannirock@gmail.com

Brasília, DF – Brasil

Resumo

Modelo de Inteligência Artificial Explicavel para Classificadores de Câncer de Mama em Mamografias

Câncer de mama está associado à maior taxa de incidência de câncer entre as mulheres, em todo o mundo. Assim como em outros tipos de câncer, diagnósticos mais precoces levam a tratamentos potencialmente menos invasivos e a maiores taxas de sobrevida. Uma ferramenta que auxilie a análise de mamografias para descobrir lesões mamárias e sua classificação constitui portanto um importante instrumento para o tratamento eficaz.

Com desenvolvimento da tecnologia, a Inteligência Artificial passou a trazer impactos cada vez maiores e mais positivos em diversas áreas, e se destaca na Engenharia biomédica por prover ferramentas de auxílio a diagnóstico cada vez mais eficazes. Em particular *Machine Learning* (ML) e *Deep Learning* a partir das duas primeiras décadas do século XX, passaram a prover cada vez mais soluções em problemas considerados complexos nas áreas de visão computacional e processamento de sinais e imagens, como os problemas de classificação de imagens e detecção de objetos.

Neste contexto, a análise mamográfica, para fins de diagnósticos e classificação de câncer, pode ser descrito como um problema de classificação de imagens, e se beneficia de algumas abordagens de ML descritas na literatura. De fato, vários trabalhos já propõe a detecção de câncer e a classificação do tipo BIRADS de forma automatizada, a partir de imagens de ultrassom ou de tomografia por raios-X da mama. No entanto, a maioria das abordagens em ML para análise mamográfica encontradas focam em técnicas tidas como de *caixa preta*, em que não há justificativa direta, em formato compreensível para um analista humano, dos resultados de classificação ou de decisões de encaminhamento. Esta limitação reduz a aplicabilidade dos modelos, já que a ausência de explicabilidade tem impactos legais em procedimentos de autorização de tratamentos, por exemplo, e reduz o potencial de novos avanços na área de oncologia e radiologia mamária, tendo em vista que os conhecimentos adquiridos se tornam menos propagáveis e generalizáveis.

Por outro lado, a maior parte das soluções encontradas na literatura científica para classificação de lesões mamárias foca em abordagens de ML com redes rasas. Foram encontradas poucas abordagens utilizando algoritmos de aprendizagem profunda, que demonstram desempenho mais alto em outras aplicações de classificação de imagens, desde que o treinamento explore uma base de imagens suficientemente representativa. Além disso, são soluções consideradas caixa preta, o que significa que é fornecida uma resposta a partir de uma imagem de entreda, e não é possível determinar diretamente quais características das imagens analisadas que mais influenciaram diretamente a decisão final fornecida pela rede, ainda que esse resultado já seja conhecido. Existem modelos de ML que provêem explicabilidade para as decisões, o que significa que são explicitadas as principais características das imagens ou sinais de entrada que levaram à decisão final do sistema. Essa característica é relevante no

contexto de aplicação em saúde, tanto pelo avanço de conhecimento que pode representar em termos dos aspectos relevantes ao diagnóstico. Entretanto, não foram encontrados na literatura muitos trabalhos avançados abordando o uso de modelos explicáveis em análise de imagens mamográficas com uso de ML.

Um modelo explicávei é capaz de se encaixar em leis e fazer com que a solução seja aplicada em um domínio real. O diagnóstico de cancer é um momento sensível, por conta disso saber como um modelo de Deep Learning ou Machine Learning chegou a um determinado resultado, pode direcionar melhor médicos a investigarem casos de maneira mais direcionada, dessa forma dando mais ênfase em algumas características da imagem, além de gerar mais confiabilidade nos resultados de predição de modelos.

O uso de Deep Learning para tarefas de classificação de imagens tem obtido resultados surpreendentes, que se igualam e em alguns casos e até mesmo superam a capacidade humana, por isso essa abordagem vai ser discutida nessa dissertação. Aliando uma poderosa ferramenta de classificação com técnicas que permitam deixar os modelos criados com predições explicáveis, assim tornar uma ferramenta de classificação de lesões mamárias com alto potencial de confiabilidade para seus usuários finais, os médicos especialistas. Para atingir esses objetivos são investigados arquiteturas de Aprendizagem Profunda como VGG16 e técnicas de explicabilidade de modelos treinados como LIME que é um framework de explicabilidade de bom desempenho e de maneira simples de utilização.

Essa dissertação tem o intuito de desenvolver um modelo de Deep Learning, que utilize técnicas de Ingeligência Artificial Explicável (XAI, do inglês Explainable Artificial Intelligence) ou seja tenha predições explicáveis que classifique lesões mamárias e identifique as características importantes que levaram o modelo a atingir tal resultado.

Após o treinamento do modelo usando arquitetura VGG16, as métricas analisadas foram acurácia, especificidade e sensibilidade, os resultados obtidos foram respectivamente 68% 77% e 65%. Resultados maiores foram encontrados na literatura, porém não são resultados que sejam reprodutíveis. Em muitos casos as bases de dados são particulares de hospitais que a equipe fez o levantamento de mammografias dos últimos 20 anos, criou-se o conjunto de dados e os testes foram feitos. Houve uma dissertação de mestrado feita por Adam Jaamour em 2020 na Universidade de *St Andrews* com uma abordagem semelhante e que obteve resultados próximos aos apresentados nessa dissertação, o autor reportou o resultado da acurácia de 67%.

Modelos de ML e DP têm um grande potencial, entretanto devem ser treinados com conjunto de dados *datasets* com grande quantidade de imagens, e imagens de qualidade. Com a performance melhorada, e atingindo métricas melhores do que as apresentadas neste trabalho, pode ser que esses modelos sejam aplicáveis em uso da vida real. O uso combinado de modelos de DL com frameworks de Inteligência Artificial Explicável, pode ajudar no direcionamento de lesões. As marcações de áreas suspeitas são diferentes do direcionamento que médicos radiologistas procuram, entretanto são marcações com potencial direcionador para uma área da lesão em análise.

Abstract

Breast cancer is associated with the highest cancer incidence rate among women worldwide. Just like other cancer types, early diagnosis leads to potentially less invasive treatments and higher survival rates. A tool that assists in the analysis of mammograms to discover breast lesions and their classification is, therefore, an important tool for effective treatment.

With the development of technology, Artificial Intelligence has begun to bring increasingly greater and more positive impacts in several areas, and stands out in biomedical engineering for providing increasingly effective diagnostic aid tools. In particular, Machine Learning (ML) and Deep Learning (DP) from the first two decades of the 20th century began to increasingly provide solutions to problems considered complex in the areas of computer vision and signal and image processing, such as image classification problems, and object detection.

In this context, mammographic analysis, for cancer diagnosis and classification, can be described as an image classification problem, and benefits from some ML approaches described in the literature. Several studies already propose the detection of cancer and classification of the BI-RADS type in an automated way, based on ultrasound images or X-ray tomography of the breast. However, most ML approaches for mammographic analysis that were found, focus on techniques considered to be black boxes, in which there is no direct justification, in a format understandable to a human analyst, of the classification results or referral decisions. This limitation reduces the applicability of the models, since the lack of explainability has legal impacts on treatment authorization procedures, for example, and reduces the potential for new advances in the area of oncology and breast radiology, considering that the knowledge acquired makes them less propagable and generalizable.

On the other hand, most of the solutions found in the scientific literature for classifying breast lesions focus on ML approaches with shallow networks. Few approaches using deep learning algorithms were found, which demonstrate higher performance in other image classification applications, as long as the training explores a sufficiently representative image base. Furthermore, they are considered black box solutions, which means that an answer is provided from an input image, and it is not possible to directly determine which characteristics of the analyzed images most directly influenced the final decision provided by the network, even though this result is already known. There are ML models that provide explainability for decisions, which means that the main characteristics of the images or input signals that led to the system's final decision are explained. This characteristic is relevant in the context of health application, for the advancement of knowledge that it can represent in terms of aspects relevant to diagnosis. However,

not many advanced works were found in the literature addressing the use of explainable models in mammographic image analysis using ML.

The diagnosis of cancer is a sensitive moment, with this in mind, knowing how a Deep Learning or Machine Learning model has achieved certain results can better direct physicians to investigate cases in a more directed way, thus placing more emphasis on some characteristics of the image, and it can generate more reliably in a model prediction.

The use of Deep Learning for image classification tasks has obtained surprising results, some results have achieved human capacity, and in some cases, it has even surpassed human capacity, which is why this approach will be discussed in this dissertation. Combining a powerful classification tool with techniques that allow the created models to have explainable predictions, thus making a breast lesion classification tool with high potential reliability for its end users, specialist physicians. To achieve these goals, Deep Learning architectures such as VGG16 and explainability techniques for trained models such as LIME, which is an explainability framework with good performance and simple to use, are investigated.

This thesis aims to develop a Deep Learning model, which uses Explainable Artificial Intelligence (XAI) techniques so that it has explainable predictions that classify breast lesions and identify the important characteristics that lead the model to achieve such a result.

After training the model using VGG16 architecture, the metrics analyzed were accuracy, specificity, and sensitivity, the results obtained were respectively 68% 77%, and 65%. Greater results were found in the literature, but these are not reproducible results. In many cases, the databases are private to hospitals where the team collected mammograms from the last 20 years, created the dataset, and tested them. There was a master's thesis done by Adam Jaamour in 2020 at the University of St Andrews with a similar approach which obtained results close to those presented in this dissertation, the author reported an accuracy result of 67%.

ML and DP models have great potential, however, they must be trained with datasets with a large number of images, and quality images. By improving performance and achieving better metrics than those presented in this work, these models may be applicable in real-life use. The combined use of DL models with Explainable Artificial Intelligence frameworks can help direct physicians to important areas of the lesion. The highlights of the lesions are different from what is expected to be shown in lesions by physicians, however, these lesions highlights potentially guide physicians to important areas for the ML or DP model.

Keywords: Breast cancer, breast lesion, Deep Learning, VGG16, XAI, Explainable, AI.

Contents

List of Tables

List of Figures

1 Introduction

This thesis addresses methods of Explainable Artificial Intelligence on Machine Learning models in breast cancer classification. The proposal is that with the support of a lesion classification system, it is possible to get the features that led the model to give the final result. So it will be possible to get the given lesion classification result, malignant or benign, and the features of the lesion that the model used to classify it with the output prediction.

This first chapter introduces background knowledge about breast cancer its impacts on public health, and the possible consequences in women's lives. Also, the importance of an automated tool to classify breast lesions and show the features on the images that led the tool to conclude the final diagnosis.

This work aims to develop an explainable Deep Learning model so that it can classify breast lesions on a mammogram, and explain what are the features contained in it that resulted in the model's classification.

1.1 Breast cancer and its challenges

Cancer starts when cells grow out of control. These cells can form tumors that can be seen in X-ray images or felt as lumps. Tumors can be classified as malignant, which is a more invasive form, and cells can grow into surrounding tissues. In this stage, the tumors can be called cancer [67]. There is also the benign classification, it is not considered cancer because cells appear to be almost normal. Their growth is not as fast as a malignant tumor's cells, and they do not invade nearby tissues nor spread to other body parts.

Breast cancer is the cause of the biggest amount of death among women around the world. Studies in 2010 have shown that it is estimated that in 2030 there will be 2.7 million cases of breast cancer in the world [23]. The most common kind of cancer around the globe in 2018 is illustrated in Figure 1.1, and breast cancer is the most common cancer among women.

Figure 1.2 shows a distribution of different cancer cases among women around the world in 2018, with approximately 8.6 million new cases. Breast cancer is the most

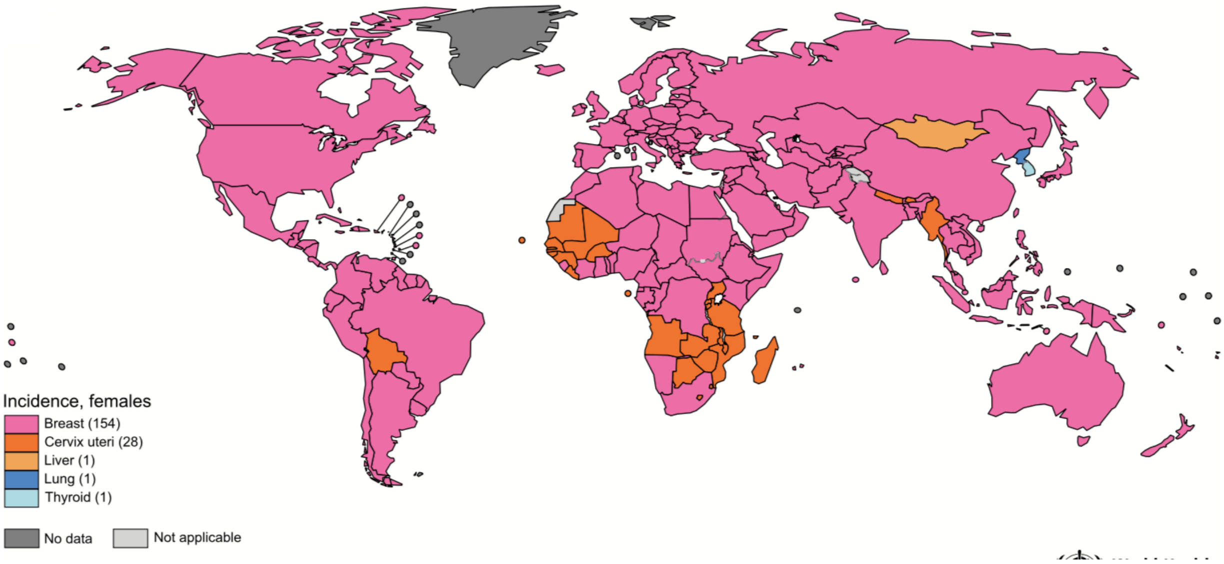


Figure 1.1. Global Map presenting the Most Common Type of Cancer Incidence in the world in 2018 in Each Country. The most cancer incidence is breast cancer - Source: GLOBOCAN 2018

common among them. The chart of deaths caused by cancer around the world indicates that breast cancer has the highest percentage of deaths caused by cancer for a single kind in women.

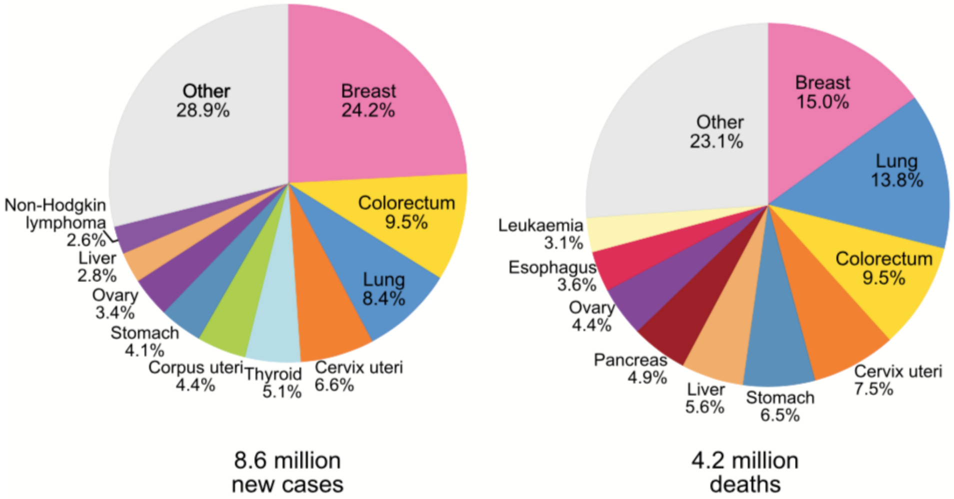


Figure 1.2. Distribution of cancer cases around the world, and deaths caused by cancer around the world for the 10 most common cancers in 2018. Source: GLOBOCAN 2018

Early treatment is a very important approach to increase a patient's chance of being healed with a less invasive approach. In order to achieve it, it is necessary to diagnose some suspicious characteristics in exams. Screening Mammography is an exam that has lower cost compared to other methods such as Computed Tomography, and it is an effective method to detect breast abnormalities and cancer in its stage [18].

Women after a certain age, even asymptomatic it is recommended to perform Screening Mammography to detect early lesions such as microcalcifications and masses, that can be a sign of cancer [7]. Mammogram analysis is mainly made manually by radiologists, it's time-consuming, subjective, and prone to human errors because it is a very difficult task because of the low contrast of the images [18]. To support the difficult diagnosis of a mammogram, studies show that radiologists have an error rate between 10% to 30% on screening mammograms [26].

Misclassifying a mammogram is very costly considering the emotional damage to the patients with a false positive cancer diagnosis, undergoing an unnecessary biopsy, or even a loss of life for a false negative case. A strategy to reduce this risk is to have a double mammogram reading, two radiologists read the mammogram to reduce the misclassification risk, however the cost increases, and the human error still remains [7].

In order to classify breast lesions, Breast Imaging Reporting and Data System (BI-RADS) classification is used as a standard measure developed by the American College of Radiology (ACR) [19]. BI-RADS contains six categories and each one has a follow-up plan that is associated with each one to help radiologists in handling the patient's situation, table 1.1 describes each category and the follow-up plan for each one [58].

1.2 Current research on automatic breast cancer diagnostic tools

Machine Learning (ML) is a subarea of Artificial Intelligence (AI) that has been becoming more and more popular over the last 10 years. Mainly because of the computational power that the latest hardware has achieved recently. The use of ML techniques to help with daily tasks such as self-driving cars, detecting road speed limits, and self-driving cars. Our daily lives have plenty of examples of ML's application, and there are other fields in which ML can help with tasks, like the medical field.

The medical field is a promising one, there are areas like screening images to diagnose diseases that can be helped and medical performance can be even more reliable. A second reading on a mammogram, for example, increases effectiveness [27], it can be done by another physician or a Computer Aided Detection System (CAD) that can be a Machine Learning Application, built to help physicians' daily work and reduce costs of having a second physician and reduce human error chances.

Computer Aided Detection (CAD) Systems are solutions designed to help physicians in analyzing results and exams. CAD solutions were developed to help radiologists in reading medical images [65]. The current approaches used in CAD are based on describing an X-ray image, and machine learning for classification [15][43].

Table 1.1. BI-RADS categories and their follow-up plan. Source: [58]

Category	Assessment	Follow Up
0	Need additional imaging evaluation	Additional imaging needed before a category can be assigned
1	Negative	Continue regular screening mammograms
2	Benign (noncancerous) finding	Continue regular screening mammograms
3	Probably benign	Receive a 6-month follow-up mammograms
4	Suspicious abnormality	May require biopsy
5	Highly suggestive of malignancy (cancer)	Requires biopsy
6	Known biopsy-proven malignancy (cancer)	Biopsy confirms presence of cancer before treatment begins

X-Ray exams are currently read by physicians, or when available by CAD system for a second reading. Some of them use Machine Learning models and their approach is a black box. There is an input image and an output answer about the image, and no explanation about what features on the image lead the model to give its answer. Making explainable models solves this problem and it also helps the daily use of the model, as a guide to how it has achieved its conclusions.

Machine Learning models can also help the medical field get authorization for exams by health insurance. Sometimes it depends on a medical report, so when applying algorithms it is important to make it able to explain the diagnosis to justify the need for getting an exam or not, so it can be used as a resource in the medical report. When building the ML Model it is essential to make it capable of explaining how it has achieved the result presented.

This work has some difficulties, and they need to be faced. The lack of medical images to build a robust dataset is a tough challenge to overcome, not just for this thesis, but for most researchers that need to work with medical images. The public datasets are very scarce, and the number of images is not much. The lesions on mammograms are very difficult to find and to distinguish among the possibilities of being a mass or calcification. The BIRADS classification is very hard to classify, and datasets that contain this annotation are very rare. Training a model to classify a breast lesion as a BIRADS classification is hard, and may be prone to errors because of the number of close characteristics of each lesion class.

1.3 Scientific proposal

The use of Machine Learning models to create Computer Aided Detection (CAD) Systems has already been in use for some systems, however, there are improvements that could be done in order to increase their performance. They are black box models, that do not have Explainability, they do not explain the features that were important to the final result. Also, the use of Deep Learning models has achieved the same performance as humans in object detection and classification problems. So there is a great opportunity of proposing an explainable Deep Learning model that classifies breast lesions.

Explainable Artificial Intelligence (XAI) is used in this work so that it fulfills the goal of explaining the output of an AI model. The features of a breast lesion can be classified as benign or malignant, and also the elements of the image that have influenced the model prediction. The combination of a trained Deep Learning model with different techniques of XAI is investigated in this thesis.

Metrics like sensitivity, recall, accuracy, and F1-score can be used to verify the performance of the model. Analyzing breast lesions is very challenging because of the features of the mammograms. When evaluating explanations it is necessary to have a ground truth annotation or an expert to check the explanation and evaluate if it is correct or not. The proposal of this work is to implement a Deep Learning Model to classify breast lesions, the Region Of Interest (ROI) of a mammogram, evaluate the model, and investigate explainability with XAI methods that can make the model prediction to be explained.

This thesis has its limitations, such as the size of the publicly available datasets, and the images are difficult to analyze and distinguish the lesions from benign and malignant. It is important to have a classification of a lesion from the model with an explanation of what are the features that were important to achieve the given result, it can make the end-users trust more when using a cancer classification system based on Deep Learning models.

1.4 Objectives

1.4.1 General Objective

The objective of this thesis is to investigate the use of Explainable AI in Deep Learning approaches, to classify breast lesions as malignant or benign, and, to highlight the features of the input image that have influenced the model to predict its result. Model explainability is how a model takes decisions based on parameters of detecting breast lesions and classifies them is the question of how it will be the outcome of this experiment and investigation.

1.4.2 Specific Objectives

This work aims to build and study explainable techniques to be applied in a Deep Learning model. This model will use Deep Learning which is a Machine Learning sub-field instead of traditional methods, its performance in classifying breast lesions mammograms will be verified, and how well the explanations are made, will also be investigated. It will be a CAD system that can help physicians in diagnosing cancer through image exams like the Region of Interest (ROI) of a Screening Mammogram, a cropped area that contains only the lesion. After classifying the breast lesions with the model, the XAI techniques will be able to show what features in the input image have influenced its classification. The model output will be the classification of the lesion found in the mammogram, benign or malignant. After the model output, it will be time to apply XAI techniques, and it will be able to show how the model has achieved the final classification.

1.5 Thesis Structure

This thesis is structured in chapters, chapter 2 is background knowledge, that is the information about breast cancer, breast lesions, mortality of breast cancer, mammogram analysis, and the known resources that can be used in detecting breast lesions and their classification, and Explainable AI (XAI). Chapter 3 is about the used materials for this work, the explanation of each material, and how they were used and prepared. The mammogram datasets that were used are listed and their features are described, the raw mammogram must be preprocessed, such as splitting, resizing, and applying filters in order to improve image quality, increasing the number of mammogram samples. The CNN architecture is also explained, that is, the network type that the model will be implemented. Chapter 4 describes implemented algorithms and the approaches used in this thesis. Chapter 5 is where the results are discussed and presented, and Chapter 6 describes the conclusion and the next steps that will be done.

2 Foundation and state-of-the-Art on Breast cancer automatic detection and Artificial Intelligence Explanation

This section aims to provide information about breast cancer, breast lesions, the existing lesion types, lesion incidence, and a comparison between lesion types. Image analysis is presented in this section, both manual and automatic analysis, and also Explainable AI methods, approaches, and evaluation.

2.1 Breast Cancer

As breast cells grow wildly they can create lumps. They can be classified as benign or malignant, a less invasive or a more invasive cancer respectively. If early discovered, the patient's treatment is less invasive while the chances of death caused by breast cancer are diminished. Early-stage cancer that is asymptomatic, may be detected by screening mammography, and the earlier any findings are found, the better. If breast cancer is diagnosed in its early stages, the chances of a patient's survival are bigger [53].

2.1.1 Breast Lesions

In mammogram analysis, a physician looks for specific lesions, such as calcification, lumps, or breast asymmetry. Those lesions can be malignant or benign, although most of them are benign [5]. Studies show that the incidence of breast malignancies was 1.1/100,000 for men and 128/100,000 for women [17].

Cossu et al [17] have studied the relationship between age and breast cancer incidence and mortality, the relation is shown in table 2.1

In a mammogram, breast density is a measure used to describe the composition of the breast, such as fibroglandular tissue and fat. These measure influence getting any findings in a mammogram. The denser the breast, the harder to find breast lesions, like calcifications or masses [5]. Even though breasts with higher density are not abnormal,

Table 2.1. Age incidence and mortality caused by breast cancer around the world, in women and men.- Source [17]

Age (years)	**Incidence/100.000**		**Mortality/100.000**	
	Males	**Females**	**Males**	**Females**
0-4	0	0	0	0
5-9	0	0	0	0
10-14	0	0	0	0
15-19	0	0	0	0
20-24	0	0.7	0	0
25-29	0	9.9	0	0
30-34	0	27.3	0	3.5
35-39	0	68.3	0	8.1
40-44	1.3	143.4	0	12.9
45-49	1.4	201.1	0.3	27.6
50-54	2.2	210.5	0.4	38.1
55-59	0.8	246,1	0.8	42
60-64	2.2	251.8	0.9	57.2
65-69	4.2	300.8	1	82.8
70-74	2.6	308.4	1.3	9.2
75-79	6.4	297.6	2.8	113.3
80-84	7.2	295.8	1.4	129.2
85+	2	235.9	2	182.9
Total	**1.1**	**128**	**0.4**	**31.6**

they have a 4-6 fold higher risk of having breast cancer when compared to fatty breasts [39].

Mammogram findings are described by radiologists, in a way established by the American College of Radiology (ACR). The lesion system that describes lesions is called Breast Imaging Reporting and Database System (BI-RADS), it contains seven categories and each one has a follow-up plan that is associated with each one to help radiologists in handling the patient's situation [58].

2.1.2 Calcifications

Calcifications are small deposits of calcium in the breast tissue, their color is white and may or may not be a sign of cancer. There are two possibilities to classify calcification, micro, and macro.

2.1.2.1 Macrocalcifications

Macrocalcifications are large deposits of calcium that are probably caused by aging of the breast arteries or old inflammation. Commonly they are related to noncancerous findings, and usually, it's not needed to do other examinations such as biopsy. This kind of calcification can become common among women older than 50 years old [5]. Figure 2.2 shows a mammogram that contains macrocalcifications as an example.

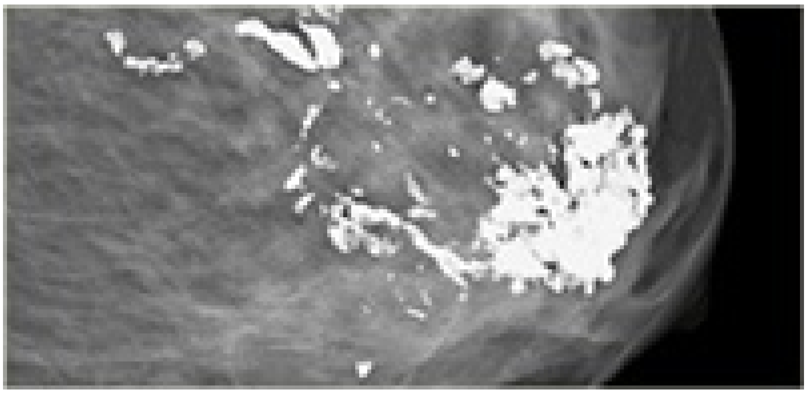


Figure 2.1. Image of a macrocalcification lesion on a mammogram. Note the large calcification region that appears as pixels with higher gray level values, when compared to the background tissues. Source: http://archive.is/9iYQU.

2.1.2.2 Microcalcifications

Microcalcifications are very tiny deposits of calcium in the breast tissue. It may or may not be a sign of cancer, but commonly its shape helps the radiologist to analyze if it is cancer or not, a biopsy is recommended to check for cancer depending on its look and shape. A mammogram that contains microcalcifications is shown in figure 2.2.

2.1.3 Masses

A mass can be a lump or a tumor [5]. A mass diagnosed in a breast can have many different meanings, including cysts and solid tumors, these findings can be a sign of cancer, but it does not necessarily mean it.

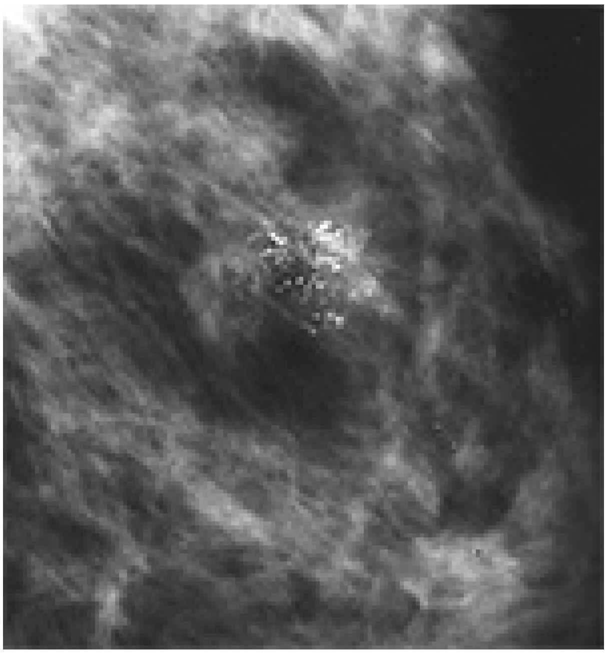


Figure 2.2. Image of a microcalcification lesion on a mammogram. Note the small calcification region that appears as pixels with higher gray level values, almost white points when compared to the background tissues. Source: http://archive.is/9iYQU.

2.1.3.1 Cysts

Cysts are fluid-filled sacs. A simple cyst is a fluid-filled sac with thin walls, this kind is not cancer and does not need to be investigated more, though if it's a solid cyst, a biopsy might be needed to check whether it is cancer or not [67]. Cysts can't be diagnosed by mammogram alone, it is necessary for breast ultrasound to confirm it, that's the reason the database used, deals with mass or calcification as mammogram findings.

2.1.3.2 Tumors

Tumors are solid masses, they are more concerning than cysts although not always indicate cancer. Figure 2.3 shows a tumor marked in a mammogram. Cysts and tumors are pretty much alike, in a mammogram they can look the same to the human eye.

2.1.4 Mass Vs calcification

Recognizing breast lesions in mammograms is a tough challenge, mainly for non-trained people. Figure 2.4 shows respectively a mammogram that contains mass and another one that contains calcifications. For a person who does not have training in reading mammograms, it is possible to say that they have no difference among lesion types, and also no difference from benign lesion to malignant lesion, however, for a skilled

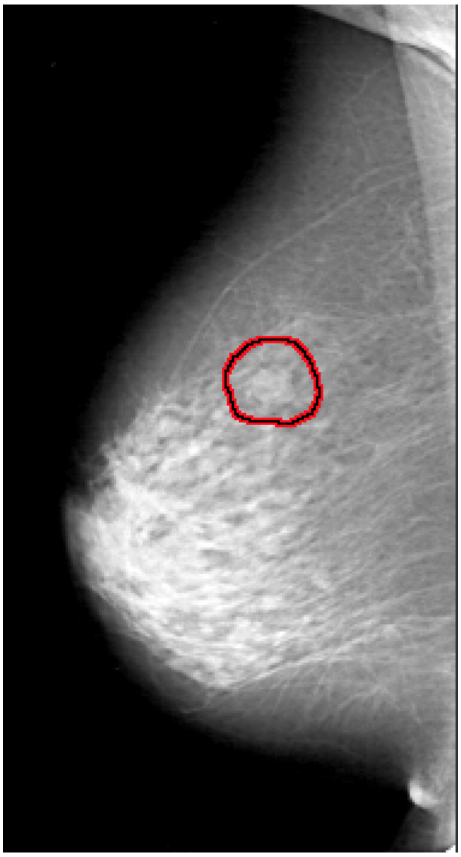


Figure 2.3. Medio Lateral Oblique mammogram View, that highlights a breast lesion surrounded in a red circle. The lesion is a benign breast tumor. Source: DDSM database

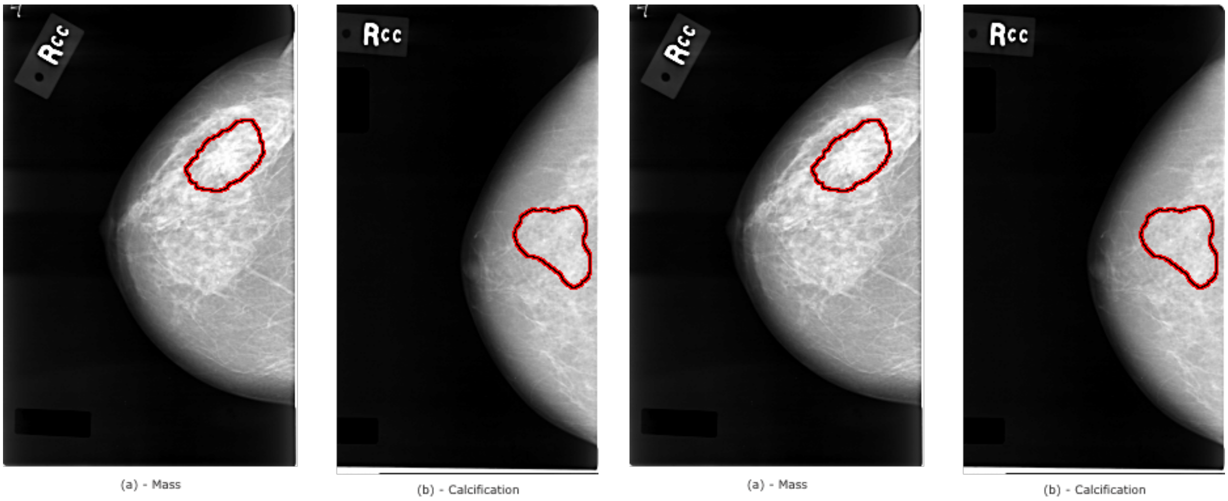
physician it is possible to discern between them and classify them correctly.

2.2 Image Analysis

According to Carneiro et al [11], in clinical settings, mammogram analysis is for the most part a manual process, which is susceptible to the subjective assessment of a radiologist, resulting in potentially large variability in the final estimation. The effectiveness of this manual process can be assessed by recent studies that show that this manual analysis has a sensitivity of 84% and a specificity of 91% [27]. Other studies show evidence that a second reading of the same mammogram either from radiologists or from computer-aided detection (CAD) systems can improve this performance [27].

2.3 Computer Aided Detection System

Computer Aided Detection Systems (CAD) are solutions designed to help physicians in analyzing the results of an examination. CAD solutions were developed to help radi-



(a) (b)

Figure 2.4. Example of Cranio Caudal and Medio Lateral Oblique mammogram views that show a mass lesion, and a calcification lesion in both views. Figure (a) - Mammogram that contains mass lesion. Figure (b) mammogram that contains calcification lesion - Source: DDSM database

ologists in reading mammograms[65]. According to Ribli et al [65], this kind of software usually analyses a mammogram and marks the suspicious regions to be reviewed by the radiologist. This kind of software was already in use by 2008 in the U.S, and about 74% of the screening mammograms were interpreted by them, however, its cost was over $400 million a year [50].

The current approaches used in CAD are based on describing an X-ray image, and machine learning for classification [15] [43]. However there is a controversial analysis about these approaches, [8],[10],[16] and [57], indicate that the results of CADs are promising, on the other hand, [71], [22], [21] have shown that CAD systems do not help radiologists' work in the U.S. This controversial result is a sign that these kinds of systems must be improved before using it [65].

Technology has developed at a great pace over the years, since the advances made by Alan Turing with his work asking if machines could think. The main difficulty in the early days of machine learning was the very low computational power, computers were able to execute a command and could not remember what was done before, so the concepts were done but the practice could not.

Technological development has led machine learning to develop as well. Neural Networks have developed over time as well, [3] details a brief history is in key events:

- 1943: McCulloch & Pitts show that neurons can be combined to construct a Turing machine [52].
- 1958: Rosenblatt shows that perceptrons will converge if what they are trying to

learn can be represented [66].

- Minsky & Papert show the limitations of perceptrons, killing research in neural networks for a decade [56].
- 1985: The backpropagation algorithm by GeoffreyHinton et al [1] revitalizes the field.
- 1988: Neocognitron: a hierarchical neural network capable of visual pattern recognition [25].
- 1998: CNNs with Backpropagation for document analysis by Yan LeCun [48].
- 2006: The Hinton lab solves the training problem for DNNs [33] [34].
- 2012: AlexNet by Alex Krizhevesky in 2012, which is a Neural Network that has won the ImageNet Visual Recognition Challenge and has shown that it is possible to use models in a large scale. [44]

Deep learning is one of the Machine Learning techniques that is very helpful in object recognition. Deep convolutional neural networks (CNN) have significantly outperformed the other methods for object recognition [44]. Deep CNN-s have reached the performance of a human in classifying images and in detecting objects [32].

2.4 Explainable Artificial Intelligence – XAI

Artificial Intelligence has reached an incredible performance compared with years before. Some approaches, such as Deep Learning models or Reinforcement Learning methods have exceeded human performance when it comes to Computer Vision. AI models are intrinsically black-box, so they are not explainable. It creates a barrier to implementing it in a clinical environment because of the characteristics of being a black-box model, such as lack of transparency, and so lack of trust [30]. Relying on critical systems such as classifying if a person has cancer or not, is difficult for experts without knowing how the classification was given. So it must be able to explain how it ended up with its decision.

Explainable AI (XAI) has been developing since the decade 1970, with attempts to explain the early AI models at that time, but it was coined in 2018 [45]. XAI dedicates itself to spreading trust in AI models, it is done by helping humans to understand how the AI models have made their predictions, in a way to mitigate the black box essence of AI algorithms. In the studies [54] and [29], the authors state that XAI is a technique that allows physicians to understand what happens behind automatic AI systems. For the image classification task, the pixels that were important to the prediction are highlighted to indicate what are the important features that have led the models to make their prediction. The highlighted features give information to the end-user on how the model came to its classification [62]. In the study [46] the authors state that end-users do not consider caring only about model accuracy, but they consider knowing how the model gives its output and how it might be influenced, so XAI is getting more and more important when deploying a model.

2.4.1 Explainability

In literature, the terms Explainable and interpretable are used with the same meaning in some parts, and in other parts, they are used with different meanings [79]. The lack of agreement on the meaning of explainability shows that there is a need for a common definition. The definition of explainability made by [55] is "explainability is the degree to which a human can understand the cause of a decision". Another definition of explainability is "explainability is the ability to justify an outcome in understandable terms for a human, and it is used interchangeably with the term interpretability" it was made by [20].

Explainability in AI models is needed for many reasons, legal and ethical reasons are reasons that are necessary for some environments. In the medical field for example, in order to implement a system in a clinical setting there are requirements and regulations that require these systems to be explainable. European Union's General Data Protection Regulation (GDPR) is an example of regulation, this regulation requires that organizations that use patient data for classifications and recommendations must provide on-demand explanations [36]. In case of non-compliance, the company that has implemented the system may receive penalties. Another important piece of information is that there are monetary incentives associated with explainable models, especially with Deep Learning models. Beyond these ethical and legal issues, explainability is important to make end-users like clinicians trust the result predicted by the model [35]. Methods used for explaining AI models, attempt to show reason in their classification, and so it will create trust between the end-user might be a physician, and the patient in the medical field. A consequence of using explainable models is that the number of wrong results may be reduced [30].

2.4.2 Ad-Hoc vs. Post-Hoc

Ad-Hoc approach is designed to make a model inherently explainable, this approach modifies the training procedure or the network architecture, so that it is possible to learn explainable features of the model. The advantages of this approach are: explanations are better than post-hoc techniques, because of the explanation is inherently designed while training the model. The second advantage is that this technique makes the explanation more trustworthy [30]. The disadvantages of ad-hoc methods are the accuracy and scalability, because when making a model inherently explainable there is a loss of accuracy, and the scalability also comes into question [30].

Post-hod techniques give explanations after the classification is made. This kind of technique is more common, because of its ease of implementation. The advantages of post-hoc techniques are the disadvantages of ad-Hoc methods, it does not have accuracy loss. The study [73] state that some people refer to post-hoc methods as diagnostic methods, because of their capability of diagnosing and their limitations for creating a complete explanation to the end user [30].

2.4.3 Model Agnostic vs. Model Specific

Model agnostic methods refer to explanation methods that can create an explanation for any AI model, with no restriction to the model's architecture. For model agnostic approaches, it is common to change the model's input and check what are the changes in the output. So the features can be analyzed what are the important features to the model prediction [30].

Model-specific methods refer to explanations that work only for a specific model. An example is a method used for CNNs is not able to be used in other architectures like LSTM or any other. This approach uses some aspects of the model architectures like feature maps that were produced from graph convolutions [30]. The choice of neural networks is limited, potentially excluding a neural network that could be a better fit.

2.4.4 Explanations: Global vs. Local

The global explanation is also called dataset-level explanation because it gives general relationships that were learned by the neural network. The global explanation is able to provide feature importance at the dataset level, and how much the image features contribute to the output prediction for the entire dataset [60]. The local explanation is the opposite of the Global explanation, it provides explanations for individual cases, only the single input for the model.

2.4.5 Attribution Methods

Most of the explanation methods are attribution based. This kind of method attempts to calculate the inputs of a model that are important to the model's prediction [30]. The classification of this method can be separated into two categories: backpropagation-based and perturbation based, both attempt to get the most important features of the input by removing one and checking the changes that happen with this modification. The attribution of each feature is calculated after that, and then the importance is ranked by attribution [30].

2.4.6 Backpropagation-based approaches

Visual explanation is also called saliency mapping, and it shows important parts of the image for a decision. Most techniques for saliency mapping use backpropagation-based approaches and, some use perturbation-based [75].

2.4.6.1 Class activation mapping (CAM)

This approach has been introduced by [80], The authors have replaced the fully connected layers of a CNN with a Global Average Pooling layer, on the last feature maps block. Medical images usually have multiple-scale information, so multi-scale CAMs are also proposed to be used.

2.4.6.2 Gradient-weighted class activation mapping (Grad-CAM)

Grad-Cam is a generalization of CAM, the authors of [68] have introduced this generalization. Grad-CAM can handle any type of CNN and it produces post-hoc local explanations. This technique is also used in medical image analysis.

2.4.6.3 Layer-wise relevance propagation (LRP)

Layer-wise relevance propagation (LRP) was introduced by [6], LRP uses the output of a network and backpropagates it throughout the network. For each iteration, LRP assigns a score to each input neuron [75].

2.4.6.4 Deep SHapley Additive exPlanations (Deep SHAP)

This unified approach for explaining results, using SHapley Additive exPlanations was introduced by [51]. This is a model-agnostic approach that uses Shapley Values, which was introduced by [69]. These values determine how the output features are influenced.

2.4.6.5 Trainable attention

The trainable attention method was proposed by [41], different from the previous techniques that highlighted regions of the image that the network has focused, this highlights in what proportion the network has paid attention to input images and then use this attention to amplify relevant areas and suppress the irrelevant ones.

2.4.7 Perturbation-based approaches

Perturbation-based methods perturb the model input so that it accesses the importance of the input-changed areas to the model purpose task.

2.4.7.1 Local interpretable model-agnostic explanations (LIME)

This method provides a Local explanation, and it is model agnostic. LIME replaces a complex model locally with a simpler model, by perturbing the input data the output will also change. The simpler model is used to learn how the changes in the input affect the output. The similarity of the input changed with the perturbations and the original input is used as a weight so that the explanations of the simpler model with the perturbed inputs have fewer effects on the last explanation. For a visual explanation in images, LIME uses super-pixels to show the import areas of the image [75]. The explanations are given by a set of intances X ($|X| = n$), an explanation matrix *W n x d'* is constructed. This matrix represents the importance of the interpretable components. When linear models are used as explanations, for an instance x_i and an explanation $g_i = \xi(x_i)$, Lime set W_{ij} = $|W_{gij}|$. Each j component (column) in W, it is denoted I_j for the global importance of that component the the explanation. LIME wants I such that features, that explain different instances have higher importance scores. Figure 2.5 shows an example problem W with $n = d' = 5$, and W is binary. I might score feature f2 higher than feature f1, $I_2 > I_1$, since the usage of f2 is more intense to explain more instances. In the case of image explanation, I measures something that is comparable across the super-pixels in different images.

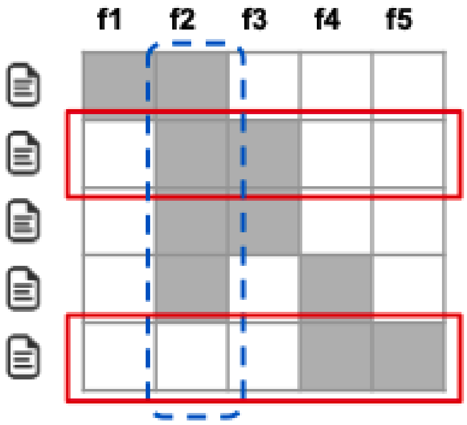


Figure 2.5. Example problem W, where Rows represent instances and columns represent features – source: [64]

2.4.7.2 Randomized Input Sampling for Explanations of Black Box Models (RISE)

This method is similar to LIME, it generates random masks of an image and feeds them into the original model. The visual explanation is generated by combining masks, and the weights are the output. The highlighted area is the pixels that are most important in the image, so it turns RISE into a very interpretable model. RISE is also model agnostic, local explainable and it considers individual pixels, different from LIME which gets a group of pixels, called super-pixels. In the study [62], the authors have concluded that RISE is a technique that works better on mammograms than LIME.

2.4.7.3 Meaningful perturbation

The input image is perturbed to check the changes in the predictions of the trained model. Instead of using other perturbations like occlusion sensitivity that blocks pieces of the image, in the study [24], the authors have introduced this method and they suggest simulating natural or plausible effects, so it can create meaningful perturbations. So the explanations are more meaningful.

2.4.8 Pros and cons

Among explainability methods and IA models, there is a trade-off between model performance and model explainability. Usually, the more complex the model, the more performance it can achieve like CNNs, and the less explainable the model is. Figure 2.6 shows the relation between interpretability and accuracy

The ease of use of XAI techniques is a pro of using XAI, most of them are plug-

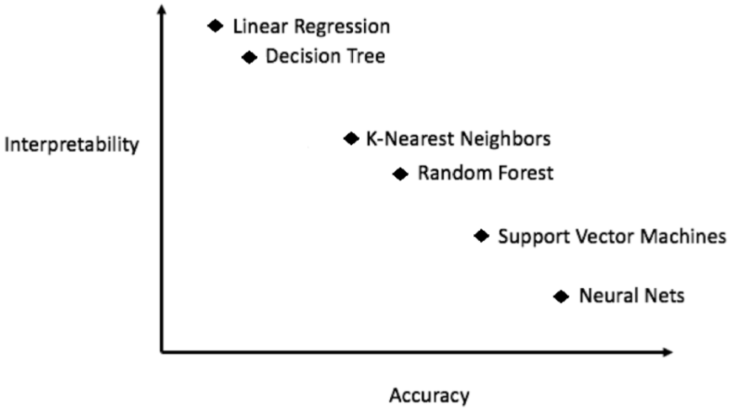


Figure 2.6. Relation between interpretability and model accuracy, the more accurate the model is the harder to make it interpretable – source: [30]

and-play. Post-hoc model agnostic techniques are the easiest to be used, they mostly use perturbation methods to map the important input features to the model output prediction. The model-based and Ad-hoc techniques are more difficult to implement compared to post-hoc model agnostic, but they also used plug-and-play techniques [75].

The XAI validation can be checked if the model explanation is correct or not with end-user experts. In the medical field, it can be done by asking radiologists to check the explanations generated [75].

The robustness of the explanations is checked by changing input aspects of the neural network and measuring what these changes cause in the explanation. Usually, it is checked with visual explanation randomizing test parameters on data tests [75].

The computational cost of XAI techniques for visual explanation cases, using backpropagation-based methods and perturbation-based methods can be analyzed as follows. The backpropagation-based methods pass through the model network once, this is considerably fast. While the perturbation-based methods use extensive perturbations in the input method, many times repeating the process to check how the input changes influence the output result. So when comparing the two methods, the perturbation-based methods are more costly [75].

2.4.9 XAI in Medicine

Explanability for AI models within the medical field is crucial to create trustful systems that can be safely and responsibly suitable for clinical implementation, so XAI attempts to build trust in the models. In the study [61], the authors state that there are challenges that physicians face with XAI such as not all visualizations are interpretable, and [4] say that for physicians, XAI techniques are not satisfactorily robust. In the study [62], the authors state that the focus of medical XAI is diagnosing rare diseases, health trends, and tumor classification. In the study [42], the authors state that the explanations that some methods like LIME and SHAP do not improve human decision when an expert checks an image with and without explanation. In the study [75], the authors have studied XAI explanations with two radiologists to check how good the explanations are. The radiologists evaluated that the relevant areas were highlighted but with irrelevant areas as well, in cancer cases for malignant and benign cases, the radiologists said that clinical features were not considered, like shapes, margins, density of the lesion, and structural distortion. Among the different explanation methods evaluated, the radiologists concluded that RISE has produced the most correct explanations. After analyzing all the methods, the radiologists concluded that the features that they would want in the real-world context do not match the explanations generated.

2.4.10 Measuring Model Explanations

XAI model explanation can be evaluated in some ways, it can be Application-grounded, Human-grounded, or functionally grounded. To be considered a good explanation, it must give an insight into how the neural network has achieved its decision or make it understandable for humans [75].

The application-grounded evaluation uses experts, they use the application to test the explanations that are generated. In medical images, a radiologist might be invited to test the application, so the advantage is that an expert inspects the explanations to check how good it is, the disadvantage is that it may be costly to get an expert scheduled to review the results [75].

The human-grounded evaluation uses human experiments, but different from application-grounded, the experiments are simpler. Instead of experts, laypersons test the explanation and judge the quality of the visual explanations. This experiment is less costly because it uses laypersons and it still has a general notion of the quality of explanations. The disadvantage is that the quality of the evaluation is a proxy of the actual quality [75].

Functionally grounded evaluation does not have humans in the evaluation process, it uses other proxies to evaluate the quality of the explanation. These proxies include measures that have already been taken and validated by a human user. In cancer cases, it may use the grounded truth annotation made by an expert with the explanation of the model [75].

2.5 Performance Metrics for models' analysis

Dealing with lesions that can be cancerous is tough work, and it must be handled carefully. For a person to be sure that he or she has been diagnosed as having a cancerous nodule, there must be other exams to confirm, not just the mammogram. When it is about Dealing with a diagnosis of having a cancerous nodule or not, it is less harmful to someone to get a positive diagnosis for having cancer and in the end, it was a false result, rather than getting a diagnosis of not having cancer and in the end, it turns that the patient did have cancer.

The person who does not know if he or she has cancer does not get treatment, so the number of false negative cases must be as minimum as possible. A false negative measure is a case that the diagnosis is negative but was positive, table 2.2 shows a matrix that relates the predicted result from the model and the actual result, this matrix is called the confusion matrix and it is the base for extracting machine learning measures from trained models.

Table 2.2. Confusion Matrix of predicting diagnosis model

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

The confusion matrix is used as a source to calculate many metrics, such as accuracy, recall, specificity, and f1-score. The metrics show a specific rate, for example, specificity shows the rate of false positives, while recall shows the rate of true positives. The metrics are used to check the performance of the trained model, it is needed to verify how good the model is. The effort of this work will be to reduce the amount of false negatives cases, and the number of false positives as well. The aim of reducing the false positive cases is to reduce the unnecessary cases of the patient who must undergo a biopsy. The metrics to fulfill the needs of this work are F1-Score, recall, specificity, and precision.

2.5.1 Precision

Precision measures the ratio of true positives (TP) over the total cases of predicted positives, which corresponds to the sum of true positives and false positives (FP). It is given by

$$P_r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{2.1}$$

The precision metric is used to check how well the model is doing while considering the cases that which breast lesions are known as cancer and the number of cases that the model predicted as cancer.

2.5.2 Recall or Sensitivity

Note that sensitivity measures the ratio of true positive over the total amount of known positive data True Positive (TP) plus False Negative (FN), and it is given by

$$R = \frac{TP}{TP + FN}. \tag{2.2}$$

Sensitivity considers the total amount of known true cases and the cases in the model made the mistake of assuming that it was not cancer when it actually was.

2.5.3 F1-Score

F1-Score means absolute error, it aims to balance between two other metrics: recall and precision. That's the reason this work needs the other two metrics, to be used in the F1-Score. It can only be high if both recall and precision are also high. The F1-Score is given by

$$F_1 = 2 \cdot \frac{1}{\frac{1}{precision} + \frac{1}{recall}}. \tag{2.3}$$

2.5.4 Specificity

Note that specificity measures the ratio of the true negative cases predicted over the total negative in the data. It can be given by

$$Spec = \frac{TN}{TN + FP}. \tag{2.4}$$

2.5.5 Accuracy

Accuracy reports how well the model is going, it describes the number of correct predictions over the total number of input samples. It can be given by

$$acc = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.5}$$

2.6 Research's challenges

Small publicly available datasets are a big challenge faced in this work while training the lesion classifier. A model is as good as its dataset in quality and the number of samples in the dataset for training a Deep Learning model to achieve better results must be as big as possible. The image size of the mammograms and the image resolution are huge. On average a full mammogram achieves 4000x5000 pixels, and the ROI images, which are the lesions themselves are also much bigger than the input size of the VGG16 network which is 244x244. The ROI dimensions can be 500x500 or even bigger. Resizing the images to fit the model input may lose some image information and distort in some cases. Computational power to train a Deep Learning model needs to be high, also analyzing breast lesions for a nonexpert like a radiologist is a very difficult task. The model classification works with the ROI image instead of the full mammogram, so the images used during the model training were all with the ROI images.

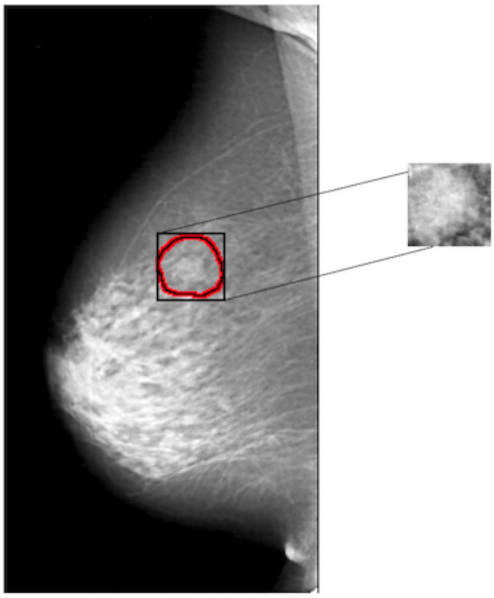


Figure 2.7. Comparison between Full mammogram and ROI (lesion) image size. A mammogram can have over 5000x4000 and the ROI itself can measure 500x600

3 Implemented Methods For Mammogram Classification

3.1 Used Mammogram Datasets and Data Augmentation and Image Preprocessing

In the medical field, public datasets are scarce, many studies use datasets from their institution. Once the datasets were not public it was necessary to discover what resources were available, and 3 datasets were found: CBIS-DDSM, INBreast, and MIAS. All these datasets found are public and their use is free. The datasets will be used to train and test the Convolutional Neural Network that will be responsible for classifying breast lesions in mammograms as cancerous or non-cancerous.

3.1.1 Curated Breast Imaging Subset of DDSM (CBIS-DDSM)

CBIS-DDSM (Curated Breast Imaging Subset of DDSM) is a standardized and updated version of another dataset, the Digital Database for Screening Mammography (DDSM), this is the biggest public dataset found with 2620 images. The images are presented as DICOM files, this format is the standard for medical images and they are divided into 2 big groups, the abnormalities of calcification and mass. Each group has the training and the test set.

One mammogram can contain calcification and also mass, so the same mammogram can be in both subgroups: calcification and mass. The mammogram's views are craniocaudal (CC) and mediolateral oblique (MLO), these views are the standard ones for screening mammography. The CC is a top view of the breast and the MLO is a side view that is taken from a certain angle [2]. Figure 3.1 shows how the mammograms are taken in each view and figure 3.2 shows how they look in a mammogram.

The distribution of cases among the classification is shown in tab.3.1 and the metadata are presented in four CSV files, each one for a group, the information contained in the files of the mammogram are:

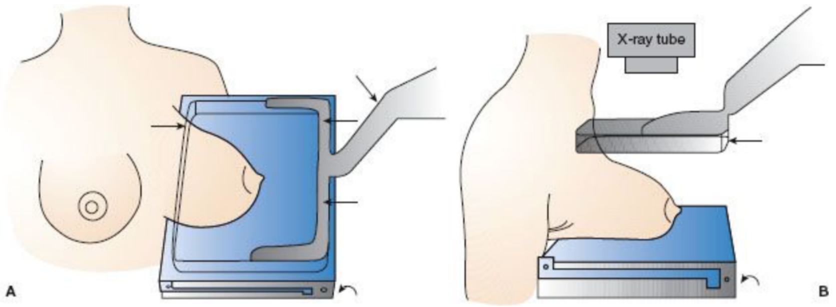


Figure 3.1. Breast positions for craniocaudal mammogram view and mediolateral oblique mammogram view. Figure A - Illustration of the positioning of the beast to take a craniocaudal mammogram view. Figure B - Illustration of the breast being positioned to take a mediolateral oblique mammogram view — Source: http://archive.is/CcL5Y

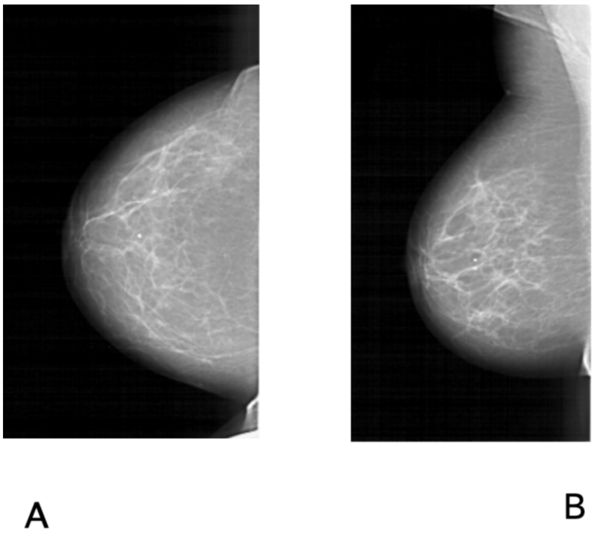


Figure 3.2. Two distinct mammogram views for the same patient with no lesion highlight. Figure A craniocaudal mammogram view; B mediolateral oblique mammogram view. Source: http://archive.is/CsL5Y.

- Patient ID: the first 7 characters of images in the case file
- Density category
- Breast: Left or Right
- View: CC or MLO
- Number of abnormality (there can be multiple abnormalities)
- Mass shape when there is one.
- Mass margin when there is one.
- Calcification type if there is one.

- Calcification distribution when applicable
- BI-RADS classification
- Pathology: Benign, Benign without call-back, or Malignant
- Subtlety rating: Radiologists' rating of difficulty in viewing the abnormality in the image
- Path to image files

3.1.2 Mammographic Image Analysis Society (MIAS)

This dataset is from the Mammographic Image Analysis Society (MIAS) produced in the United Kingdom, it has 322 digitized films. This dataset was produced from mammograms selected from a major center participating in the United Kingdom National Breast Screening Program. [72] The mammogram views are mediolateral oblique (MLO), it has normal, benign, and malignant cases, and the images are distributed in PGM format.

Abnormalities in the data set are, calcifications, circumscribed masses, spiculated masses, architectural distortions, asymmetries, miscellaneous and normal. This data set is distributed as 66 benign, 52 malignant, and 204 normal The data was mixed and split into train, test, and validation. Samples of malignant and benign images are shown in figure 3.3.

3.1.3 INBreast

The INBreast dataset is made of images acquired at a Breast Center, which is located in a University Hospital in Porto, Portugal, it contains 115 cases and 410 images. 90 cases are from women with both breasts, which leads to 4 images per case, the other cases are

Table 3.1. CBIS-DDSM dataset amount of images contained

	Benign	**Malignant**	**Benign without callback**	**Total**
Mass				
Test	194	147	37	**378**
Train	577	637	104	**1318**
Calcification				
Test	130	129	67	**326**
Train	528	544	474	**1546**

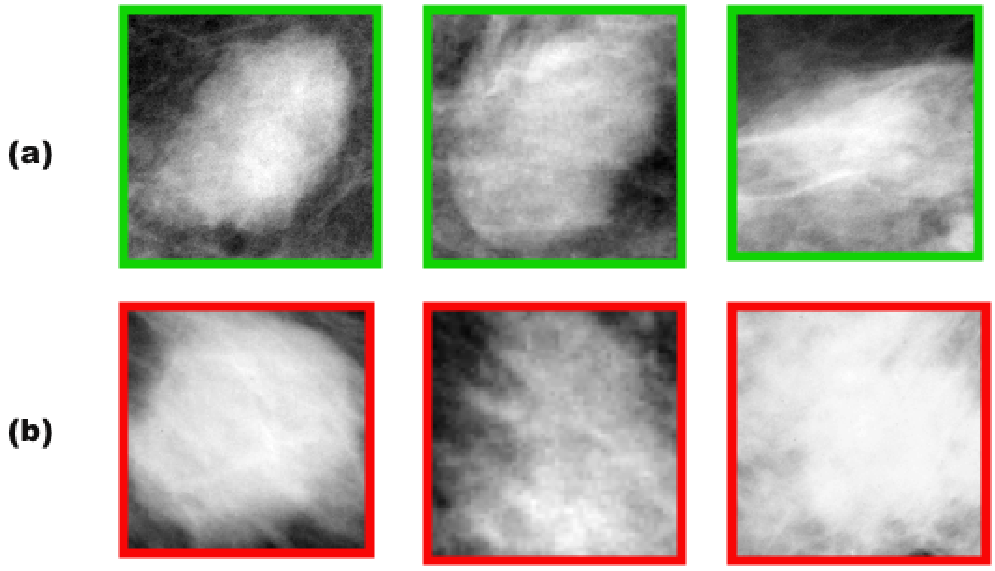


Figure 3.3. Difference between malignant breast lesions and benign breast lesions. Figure (a) benign lesion samples Figure (b) malignant lesions samples

from mastectomy patients and there are 2 images per case [9]. The lesions on this data set are masses, calcifications, asymmetries, and distortions.

3.2 Image Preprocessing

The images were split to get the cropped images where only the ROI of the breast lesion was taken to make the experiments. A script to convert the images from .DICOM to .png was applied, and another one to split images into training, test, and validating. It was also needed to resize the images into a pattern of 224x224.

In order to improve the image's contrast, a script to apply Global contrast normalization was used on the split dataset, and Among the BIRADS classification, only two classes were taken to the training dataset. BIRADS 2 and BIRADS 5. The edge classes were chosen in an attempt to take more different features between the lesions that are cancerous or not.

3.3 Data Augmentation

The more data to train a Deep Learning model, the better. The performance of the model is influenced by the amount of data [31]. A very popular technique to help increase the amount of data in the datasets is called Data Augmentation, it can increase the dataset 10 times its original size. It's common knowledge that the amount of data helps to prevent overfitting when training a network with a small amount of data. A way of performing data augmentation is by adding noise or transformation to the data. The images were augmented using random transformations, so the model would not see the exact image twice. The parameters of the augmentation were done according to the work done by [14], and the transformations were height and width shifts with a fraction of 0.25 of the total original image, a random rotation range of 0-40 degrees, a shear range of 0.5 and a zoom range of [0.5 - 1.5]. The images were also flipped horizontally and the fill mode strategy for filling newly created pixels was used, after rotating or using a width/height shift these pixels may appear. To handle the data augmentation it was used the Keras ImageDataGenerator, which generates batches of data to perform real-time data augmentation.

The number of benign lesions was updated to 814 and the malign 1185 images in the training set, they are all ROIs cropped of the full mammogram. The test and validation datasets were not augmented, and it kept being 9 and 38 for benign and malignant images respectively, for the test set and the validation set the number of images was 9 and 36 respectively as well.

3.4 Convolution Neural Network

Convolutional Neural Network (CNN) was inspired by animal cortex observation and its studies. The start point was in 1968 with [38], where the cat's visual cortex has been studied. It was discovered that the visual cortex contains arrangements of simple and complex cells, and it makes the animal visual cortex a very powerful visual processing system, so many attempts to emulate its behavior have been done since this study was published.

In 1988 the network's structure was proposed by [25], but due to the hardware limitation, it was not possible to be widely used, however, in 1998 [48] succeeded in using CNN to the problem of handwritten digit classification. Although the development in the 90s, CNNs have become widely used just by 2012 with AlexNet, when the computational power increased considerably, the cost of training a CNN decreased, and the GPU use helped to make the training of the network faster.[44]

According to [3] the advantages of CNNs over other techniques such as Deep Neural Networks (DNNs) are described as:

- It's more similar to the human visual processing system

- It is more optimized in structure for processing 2D and 3D images
- Fewer parameters and connections than a fully connected network of similar size.

Although the CNNs advantages, when applying them on a large scale to high-resolution images it becomes very expensive computationally, due to the current use of GPUs paired with the highly-optimized implementation of convolution it is possible to make the large CNNs training easier [44].

The arrangement of a CNN layer is handled in a three-dimensional way: width x height x depth. The depth dimension is about the image's channel, if it is RGB, the depth is 3, if it is in grayscale it is 1. The neurons in one layer do not connect to all the neurons of the next layer differently from a regular Neural Network, but just into a restricted region. Figure 3.4 shows the comparison between the two networks.

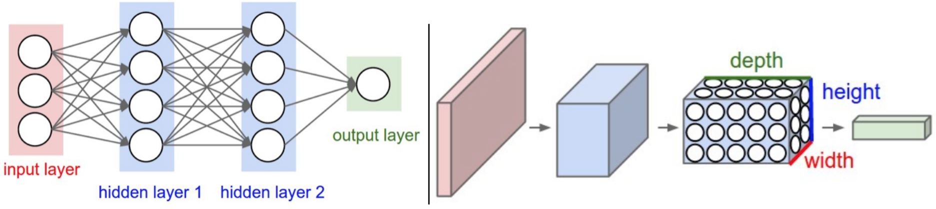


Figure 3.4. Comparison of Recurrent Neural Network Structure and Convolutional Neural Network Structure. - Source: http://cs231n.github.io/convolutional-networks/

Convolution network uses a linear operation, the convolution, and that's why it has this name[49]. Instead of using regular matrix multiplication which is used in fully connected networks [28], it uses the convolution operation.

The architecture of a CNN can have many different variations, however, in general, it has convolutional and pooling layers, which are grouped in modules, and one or more are fully connected layers [63]. An example is shown in figure 3.5, the problem of classifying an image. A car image is used as input, afterward passing through several modules (convolution and pooling layers) the input is used in the fully connected layer and then the output is given.

3.4.1 Convolutional Layers

Convolutional layers have a fundamental importance as a CNN architecture component, they are responsible to perform feature extraction. It consists in combining linear and nonlinear operations, the convolution operation, and the activation function. A small array of numbers called a kernel is applied across the input image which is a matrix of numbers as figure 3.6 shows.

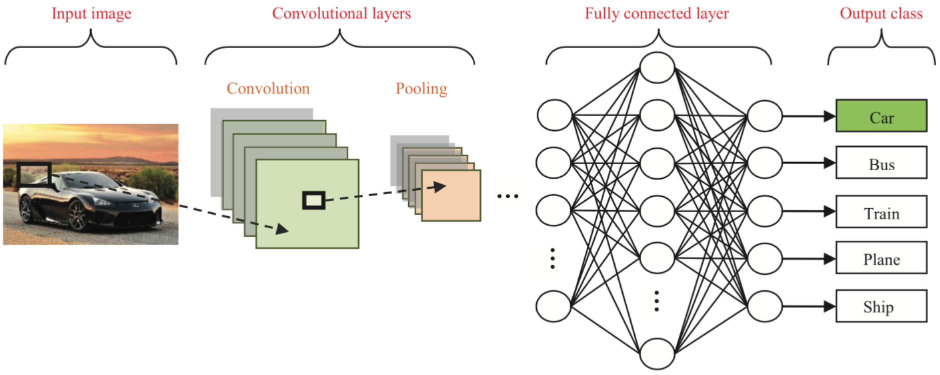
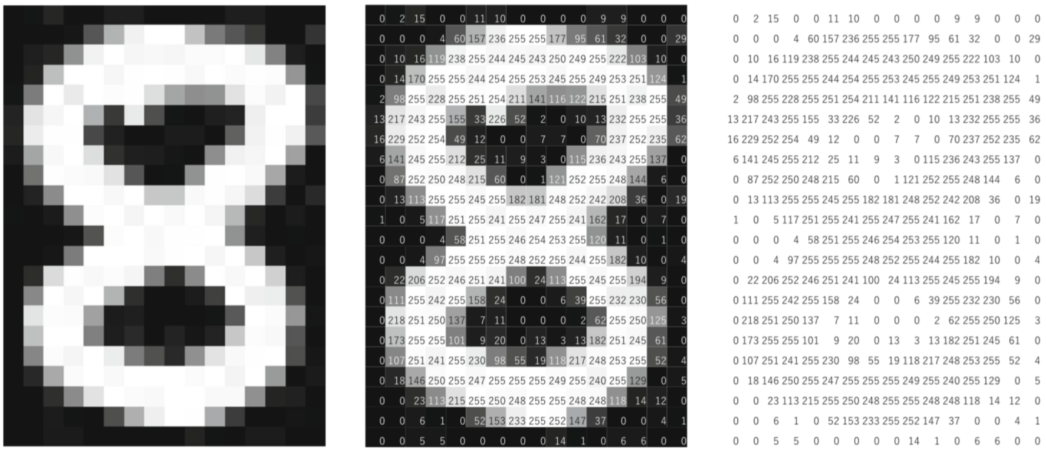


Figure 3.5. Example of CNN classification structure, starting from an input image, passing through convolutional layers of the neural network, fully connected layer, and the final output- Source: [63]



0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

Figure 3.6. How an image is handled by computers, a matrix of numbers between 0 and 255, these numbers correspond to the pixel brightness. - Source: http://yann.lecun.com/exdb/mnist/

The product between the kernel and the input is calculated for each location of the matrix and summed to obtain the output array that calls feature map [76]. The study of [63] says that the kth output feature map Y$_k$ can be computed as

$$Y_k = f(W_k * x) \tag{3.1}$$

In the equation 3.1, x is the input image, the feature map is W$_k$, and f is the nonlinear activation function that allows the extraction of nonlinear features [78]. Some traditional activation functions such as sigmoid and hyperbolic tangent were widely used previously, these functions are mathematical representations of a biological neuron behavior. Although the most common function used nowadays is the rectified linear Unity (ReLU), this work also uses it. The ReLu function simply computes f(x) = max(0,x), figure 3.7 shows the behaviors of the activation functions mentioned [63][76].

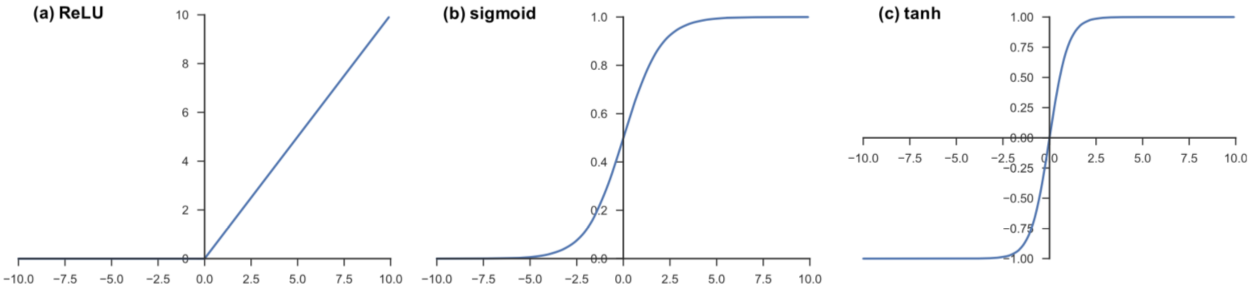


Figure 3.7. Common activation functions used in neural networks. a) rectified linear unit (ReLU) b) sigmoid c) hyperbolic tangent. - Source: [76]

3.4.1.1 Pooling Layers

The pooling layers aim to reduce the spatial resolution of the feature maps and thus achieve spatial invariance to input distortions and translations. It was a common practice to use the average pooling aggregation layer to propagate the average of all input values to the next layer [49][48], however, current studies show that max-pooling aggregation layer [44] propagates the maximum value with a receptive field to the next layer [37], it selects the largest element and to do so it uses the following equation 3.2

$$Y_{kij} = maxX_{kpq},_{(p,q)\in\Re_{ij}} \tag{3.2}$$

The output of this pooling operation, associated with the kth feature map is represented in the equation as Y_{kij}, X_{kpq} represents the element at the location *(p,q)* in the pooling region $\Re_{ij}$(Yu et al,.2014)

3.4.1.2 Fully Connected Layers

On the previous layers, several layers of convolutional and Pooling layers were stacked to get more abstract features. The Fully Connected layers also known as Dense layers, receive the output of the previous layers usually in a 1 Dimension scalar array. The transformation from a 2D image array to a 1D array that is the input to the Fully Connected Layers, is done by the Flatten Layer, this layer is responsible to convert multi-dimensional arrays to a one-dimensional array, and direct the max-pooling layer output to the Fully Connected Layer input, figure 3.8 shows how this process happens.

Once the features are extracted they are mapped by a subset of the fully connected layers to the final output that contains the probabilities of the output be the mapped classes. The final layer has the same amount of output as the classes, each layer has an activation function. Although there is no number of layers predefined, in most cases this number varies from two to four layers, it has been observed in architectures as

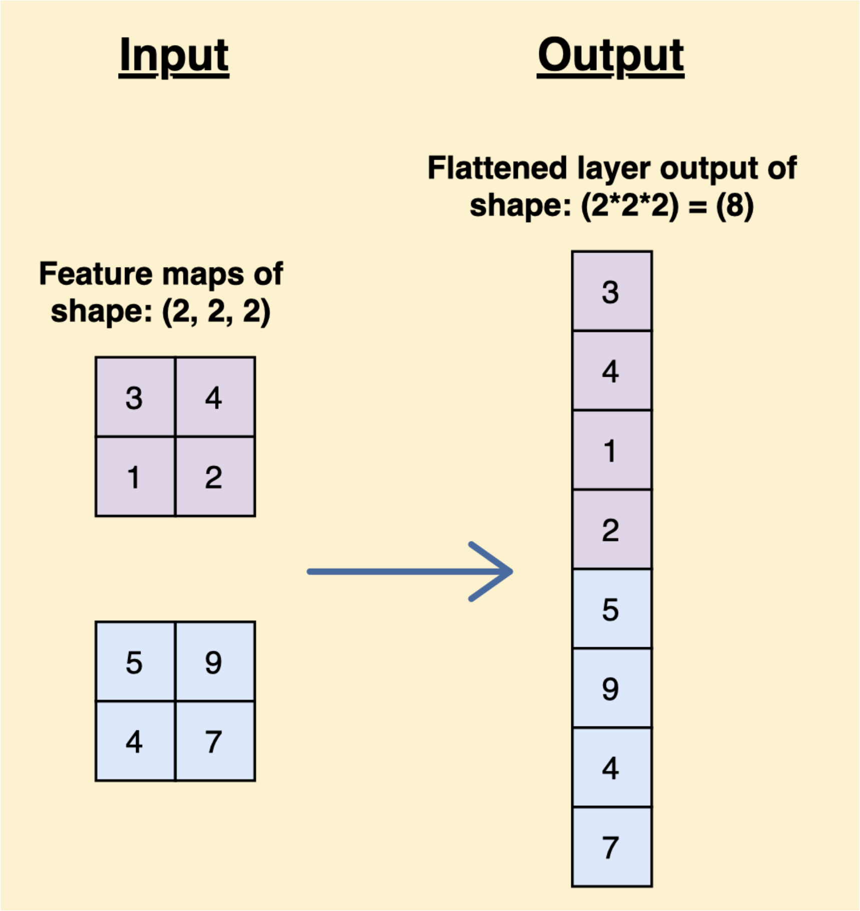

Input
Output
Feature maps of shape: (2, 2, 2)
3
4
1
2
5
9
4
7
Flattened layer output of shape: (2*2*2) = (8)
3
4
1
2
5
9
4
7

Figure 3.8. Flatten layer converting multi-dimensional arrays to a one-dimensional array. - Source: [59]

[48] and [44], once these layers have a considerable computational cost. When dealing with classification problems, the last Fully Connect Layer, as a standard uses a different activation function, the choice depends on the problem that is being dealt with, to the problem of image classification it is normally used softmax classification layer [63][3][76]

4 Proposed breast cancer classification algorithms and explanations of the results

Artificial Neural Networks (ANNs) in general including CNNs, use learning algorithms to adjust their parameters. Bias and weights are the parameters adjusted, to get the expected output, and the most common algorithm used is the Backpropagation [48][49]. This algorithm uses the Cost function and the gradient descent to optimize itself. The cost function indicates the difference between what is expected, the predictions, and what was gotten, the gradient descent updates the learnable parameters of the network, in an attempt to minimize the loss.

A common challenge that is faced while training a CNN is overfitting, it fits the data used to train the network very well, but when used in real problems, the network is unable to classify the input correctly. Therefore the ability to generalize the correct recognition of unseen data is critically affected. Overfitting can be mitigated for example by using regularization.

4.1 CNN Architectures

In the past years, CNNs have shown greater performance as they go deeper. This work was done using one of the state-of-the-art networks, that was pre-trained on ImageNet. The transfer of learning from natural images to breast cancer images was done, as well as fine-tuning.

4.1.1 Building a Model With VGG16

VGG16 is a published network from Oxford University, and it is one of their best models [14], and it was chosen to be used in this work because of its simplicity and robustness. It is a deep CNN and very simple. This network has the input in a fixed size of 224x224, the image goes through 5 convolutional blocks, figure 4.1 shows more details about the structure of this network. The total of convolutional layers is 13, the amount of max pooling layers is 5, and 2 fully connected layers. The size of its filters is 3x3 and the stride of the convolution is fixed to 1, the pooling layer is the max-pooling.

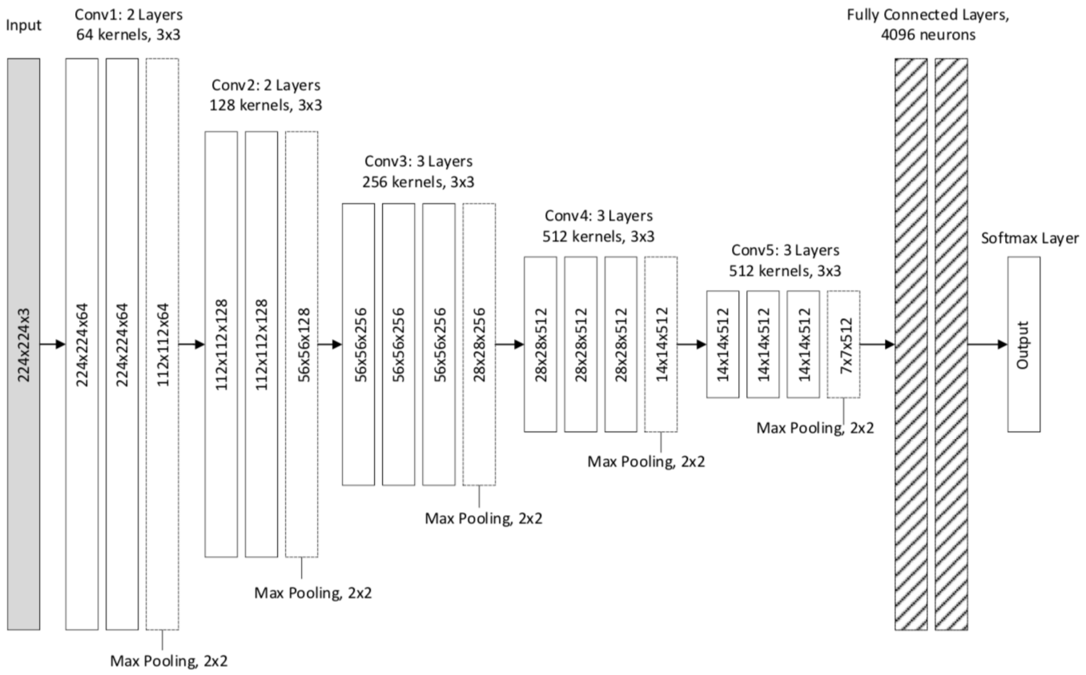


Figure 4.1. VGG16 achitecture detailed in layers, convolutional blocks, fully connected block, activation function layer. - Source: [74]

4.2 Transfer Learning

Training deep learning models with a very low amount of data is a tough challenge, and one of the strategies to overcome this difficulty is the Transfer Learning approach. It consists in training a network in a specific domain that has a large amount of data, and after that, retraining part of the network into another domain of images. The authors [13] and [70] have already shown that transfer learning is beneficial even if the domains are unrelated to one another, which is exactly the case of this work, natural and medical images. The use of pre-trained models is very helpful in the case of scarce data[14], and it can extend beyond it, for example when it comes to the effective initialization technique for complex models [47][40].

The CNN model VGG16 used in this work was used with the weights from the ImageNet. The whole architecture of the model was kept, but the original Fully-connected layers were discarded. The original fully-connected layers would not fit the proposes of this work, because it has been built for the ImageNet dataset that has 1000 outputs, each one for a different class. A new fully-connected layer that fits the breast lesion purposes was built, and the VGG16 base model got it appended to make a completed model. This new model has just 2 distinct classes as its output, 0 which means benign output, and 1

- which means malignant.

Initially, the VGG16 model was used as a feature extractor, and a Softmax layer was used as a classifier to train the new fully-connected layers with the pre-trained weights from the ImageNet model. The training occurred for 10 epochs, and after that, a fine-tuning strategy was done to attempt to achieve the best result with the VGG16 model.

4.3 Fine-tuning

It's known that the first convolutional layers in a CNN learn more generic features and perform tasks like edge detection, it can be useful for many different tasks. As the layers go deeper, the features get more specific to the classes of the dataset [77]. Since lesions on mammograms are very different from the ImageNet images, this work attempts to fine-tune the CNN model that was chosen, VGG16 to adjust the features of the convolutional blocks and then, turn them into more specific to the mammogram's lesion data. The weights of the pre-trained model were fine-tuned using the breast lesion data, so the backpropagation on the unfrozen layers adapted their weights to the new data.

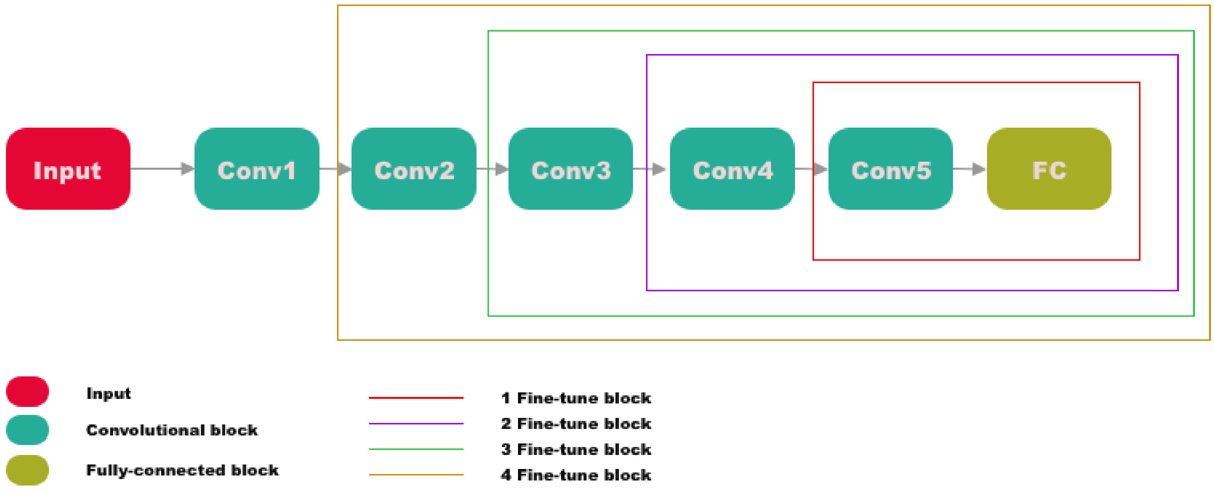


Figure 4.2. Fine-tuning strategy that can be applied to the VGG16 CNN model. - Source: [74]

Some approaches to fine-tuning the CNN model were taken into consideration to test which one would have the best results with the model. The amount of frozen convolutional blocks was 1 block, 2 blocks, 3 blocks, and 4 blocks and Figure 4.2 has the details about it. Table 4.1 shows the number of layers of the model and the number of layers to each block that we intended to fine-tune.

Table 4.1. Amount of layers in each convolutional block

Model	Total number of layers	Last 1 Convolutional Block	Last 2 Convolutional Block	Last 3 Convolutional Block	Last 4 Convolutional Block
VGG16	23	4	8	12	15

4.4 Choice of hyper-parameters

Choosing the hyper-parameters is a very important step because the results are influenced by them. For the first step of training just the new fully-connected layer the ADAM (Adaptive Moment Estimation) optimizer was used, this method is made to combine the advantages of two other popular methods, AdaGrad and RMSProp [14]. When fine-tuning the model the Stochastic Gradient Descent (SGD) was used with many different attempts of initial Learning rates, the attempts ranged from 1e-2 until 1e-7. It used the approach of early stopping the training, the parameter to be monitored was the validation loss with patience of 30 epochs. Other important steps to avoid overfitting besides data augmentation, are L2 regularization and dropout, both were used in this work. The L2 regularization is used to penalize large weights and prefers the smaller ones. The L2 regularization operates on the weight matrix W and can be written like this: R(W) = $\sum_i \sum_j \mathrm{W}^2_{i\,j}$, and the loss function turns into L = $\frac{1}{N}\sum_{i}^{N}{}_{=1}$ L$_i$ + λ R(W), and λ is a hyper-parameter that is responsible for controlling the amount of regularization that is being applied. In this work, λ is used with the value of 0.1, which has shown the best result. In the fully-connected model it was added a dropout layer, this layer randomly turns off the activation while training the model, and the probability used was 0.5.

During the training, at each epoch, the results of the model were checked for the one that had the highest validation accuracy and it was saved as the validation accuracy was better.

4.5 Making the model explainable

After training the breast lesion classifier, it is time to make its predictions explainable, in order to do that, explainability techniques must be applied. This work uses post-hoc, model-agnostic techniques. These techniques are useful because it fits CNN networks, ad-hoc and model-specific techniques depend on less powerful and simpler models. This work uses Local Interpretable Model-agnostic Explanations(LIME), and Randomized Input Sampling for Explanations of Black Box Models (RISE). Both methods are based on perturbing the model and getting the features that are more important to the output result of the lesion classification. Once the model was trained, the next step is to apply

LIME and RISE to the model and make predictions. The model is used to predict the result of a lesion, Lime and RISE analyze how the model does its prediction and gets the features that were important to the model to achieve the presented result, it is possible to get the images highlighted with features that made the model returns the prediction presented to the user.

The parameters for both XAI techniques, used in this work were 1000 samples, 5000 samples, and 10000 samples for LIME. These parameters are the amount of perturbation for fine-tuning the model's feature analyzer. Features that are important and the ones that are not that important to the final prediction are shown in both techniques. It is presented in different colors, red for important features and blue for not-so-important features.

RISE is close to LIME, but it differs in how it analyzes the images. RISE uses pixel analysis, while LIME uses superpixels, which is a group of pixels. RISE experiments were done using different parameters like LIME. The result of RISE explanation is a heat map in the image analyzed, the red areas are the most important parts of the image that have influenced the output prediction.

As an example, the MNIST dataset is used here to illustrate the use of LIME and RISE on a toy problem before being applied in a breast cancer context. Figure 4.3 shows a handwritten number, it is the digit 7, this is a sample of a dataset called MNIST. This dataset is well known in Machine Learning World, a classification model has been trained for digits recognition, its accuracy is 98%, and LIME and RISE will be applied to this model in order to make its output become explainable.

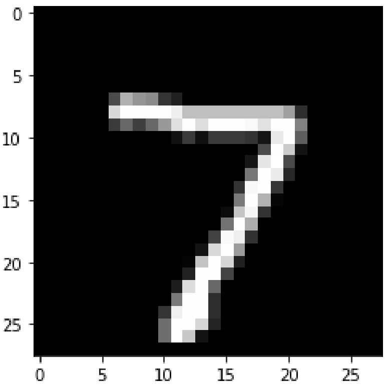


Figure 4.3. Input image as an example of MNIST dataset, a dataset of handwritten digits - Source: MNIST dataset loaded from Keras

Figure 4.4 shows the features that have been influenced the most when predicting the digit. It was a correct prediction, a number 7 and the red area highlights the area of the most important part that has guided the model to infer that it was actually a number 7 digit.

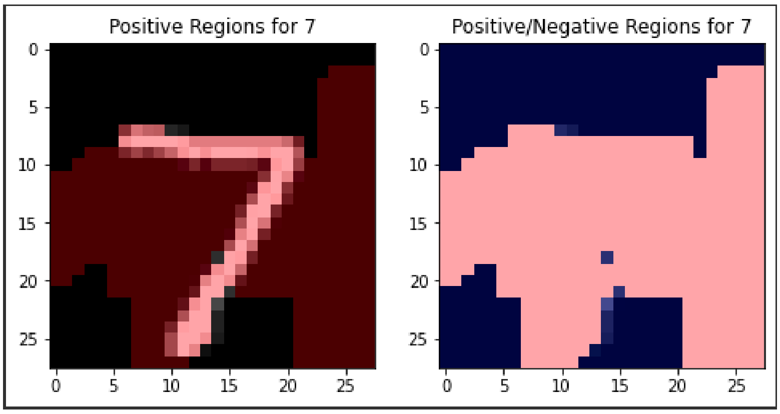


Figure 4.4. Features highlighted with the most important ones to the model's prediction using LIME, what areas of the image the model took into consideration to the classification 7

Figure 4.5 shows the explainability of RISE for the same image, it is a heat map, where the red area is the most important feature for getting the result that the model has predicted for classifying the input image. There are other colors like red and blue, but the most important parts are highlighted as red.

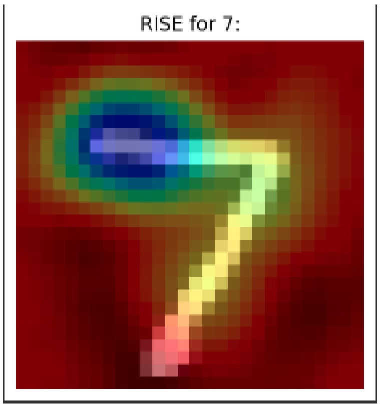


Figure 4.5. Highlighted important areas using RISE to the prediction of the classification of number 7 to the input image of MNIST

5 Results and Discussion

The breast lesion classifier has been trained and the results using the VGG16, pre-trained weights trained in the ImageNet challenge, and fine-tuned breast lesions are presented in this chapter. The training curve for loss, accuracy, validation loss and validation accuracy is shown by figure 5.3. The metrics used to evaluate the model's performance were accuracy, sensitivity, and specificity. The model's accuracy is 68%, sensitivity 77%, and specificity 65%. These metrics all were taken with the test set that was split from the training set, the confusion matrix of the test sessions is illustrated below in Figure 5.1, and the ROC curve is shown in figure 5.4.

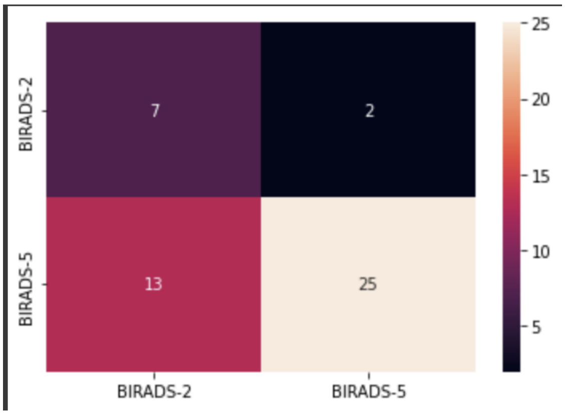


Figure 5.1. Confusion matrix with the results of the VGG16 breast lesion classifier trained model

A master's degree thesis from 2020 made by [12] in the University of St. Andrews has achieved a close performance to a mammogram classifier. The authors have achieved 67% accuracy, the approach was for the whole mammogram classification instead of only the ROI lesion. In this work, the result achieved is 68% accuracy. The type of mammogram image, that is hard to analyze, and the lack of publicly available data, makes it a hard task to get good performance. The tough challenge is proven because of other studies than this work, like [12] have also attempted but the final result is very close to the result of this work.

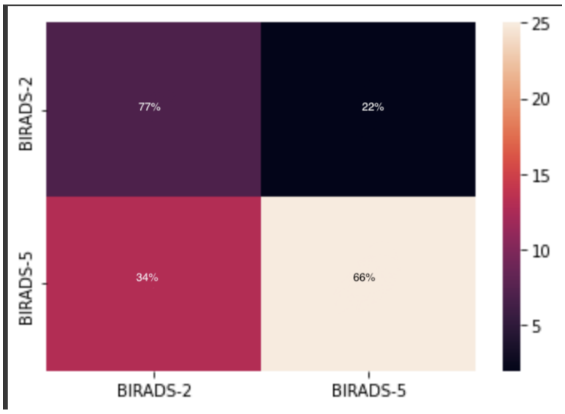


Figure 5.2. Confusion matrix with the results in percentage of the VGG16 breast lesion classifier trained model

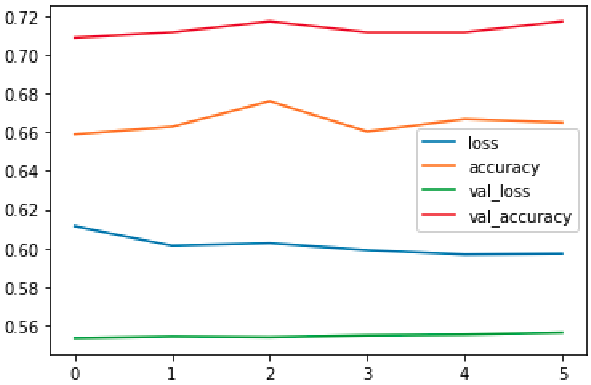


Figure 5.3. Training curve for loss, accuracy, validation loss and validation accuracy

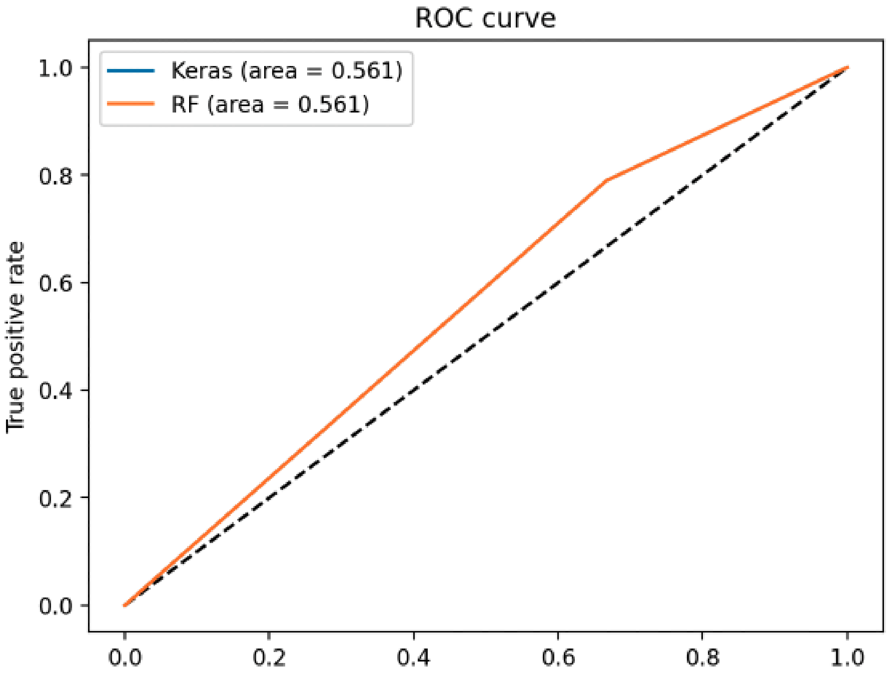


Figure 5.4. Roc curve for VGG16 classifier

5.1 XAI EXPERIMENTS

The experiments were made using LIME and RISE, two explainability techniques capable of making the model's prediction explainable. The experiments were executed with 3 images, 2 benign and 1 malignant. The best way of analyzing the lesions' explanations is using an expert to check if the highlighted regions were correct and helpful. However, it was not possible to use an expert physician in mammogram analysis, so, the second option was the analysis made by a non-expert person with little knowledge of lesions. The parameters for each technique were changed to check the final result of the prediction's explanation.

5.1.1 LIME

Lime uses superpixels for highlighting important areas of the image in the model's prediction, it can be calculated using the sample parameter while using the perturbation in the model to create the explanation mask. This parameter has been trained in three different approaches, 1000, 5000, and 10000 samples.

In the case of 1000 samples running, in a malignant lesion that the trained model correctly classified, figure 5.5 shows the explanation mask where the important parts of the image are highlighted in red. It shows in this case that the most important part of the image to make the prediction was the top. To check the explanation viability, it seems to be correct, since the highlighted area is about the area that the lesion has undefined borders, differently than the bottom of the image, that has well defined borders.

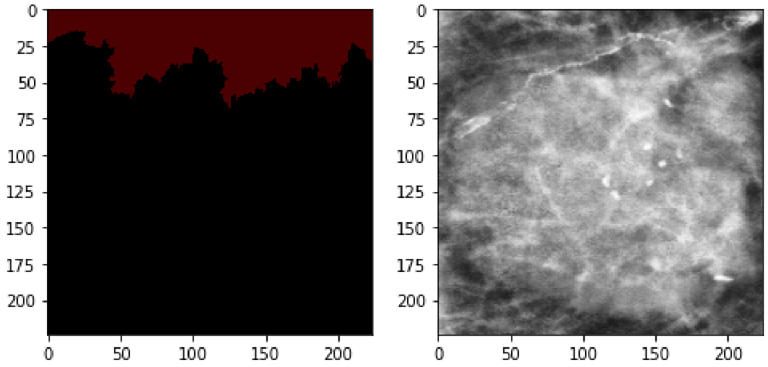


Figure 5.5. Explainability mask, generated with 1000 samples on LIME, that shows important features for the lesion classification, and the raw lesion image

Figure 5.6 shows the explanation results when LIME uses 5000 samples for perturbing the prediction, the mask has changed from the first explanation. The highlighted area has also changed from the experiment of 1000 samples. The areas got mainly in the borders of the lesion, so it seems that the most important area for the model was the borders.

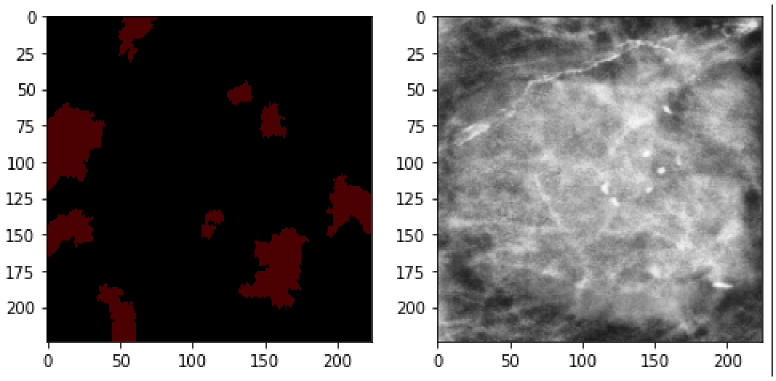


Figure 5.6. Explainability mask, generated with 5000 samples on LIME, that shows important features for the lesion classification, and the raw lesion image

In figure 5.7, the experiment uses 10000 samples for explaining the predicted result, it differed from the 5000 samples experiment, and it matched the result from 1000 samples. Highlighting the top of the lesion as the most important, it indicates that the usage of 1000 or 10000 samples highlights the real important areas

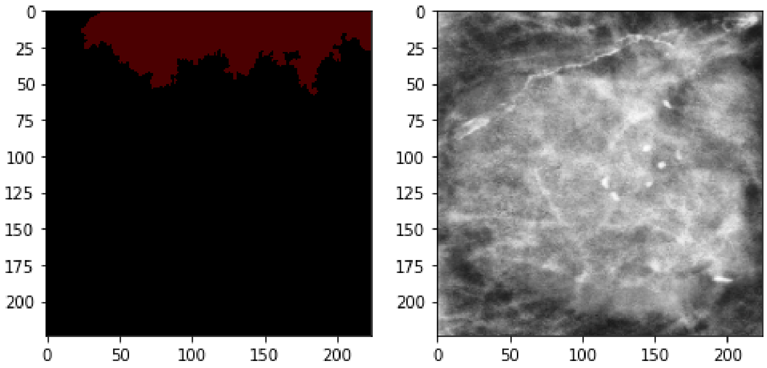
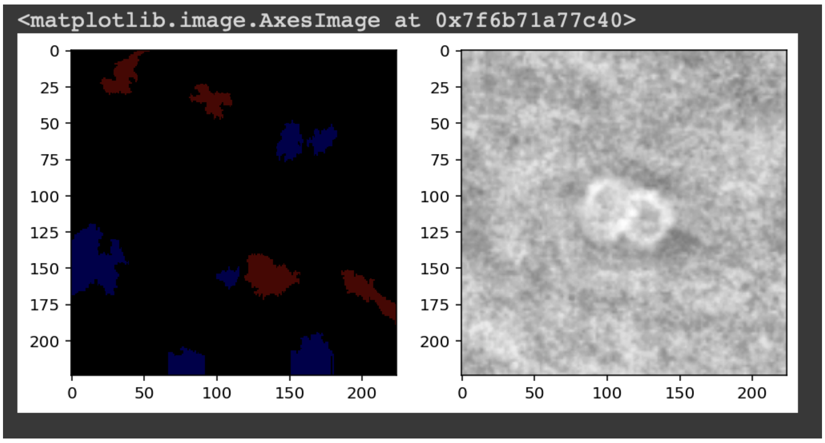


Figure 5.7. Explainability mask, generated with 10000 samples on LIME, that shows important features for the lesion classification, and the raw lesion image

Now the second breast lesion is to be analyzed and used as an experiment, it is a benign case, and the feature explanation mask and the raw lesion are shown in figure 5.8. The features painted in red are the important ones to the prediction, and the blue regions are the opposite of the red ones.

The explanation set using 5000 samples on LIME for a benign correct lesion classification has been shown in Figure 5.8, it does not show the blue areas that Figure 5.7 shows, representing the areas that have not influenced the correct prediction. More red areas have appeared in this explanation set.



```
<matplotlib.image.AxesImage at 0x7f6b71a77c40>
```

Figure 5.8. Explainability mask, generated with 1000 samples on LIME, that shows important features for the lesion correct benign classification, and the raw lesion image

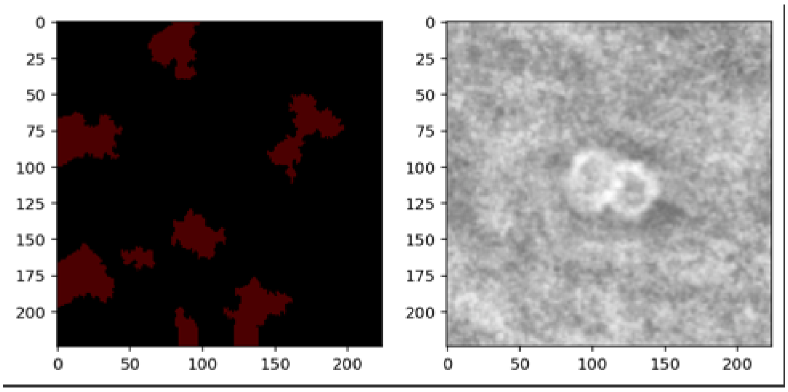


Figure 5.9. Explainability mask, generated with 5000 samples on LIME, that shows important features for the lesion correct benign classification, and the raw lesion image

Using the sample parameter as 10000, the areas of the previous experiment have grown. There has been some consistency with the explainability masks for 1000 samples, 5000 samples, and 10000 samples. The explainability mask and the lesion are shown in Figure 5.10 This

The next case is a benign lesion that the model has predicted as malignant, this case is interesting because the explanation is about to show why the model made the wrong prediction and what part of the image has influenced this result. The explanation mask and the lesion are presented in Figure 5.11

Using 5000 samples with LIME, it has generated one red feature as important for the prediction and many blue areas in the picture, showing the features that let the model not predict the lesion as benign, that was the correct result.

Using the set of 10000 samples for explaining the prediction, LIME has generated an explainability mask with the features at the top of the image, this time no blue areas are

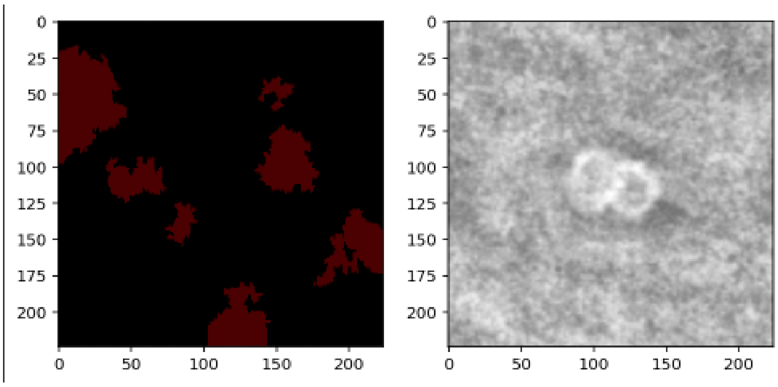


Figure 5.10. Explainability mask, generated with 10000 samples on LIME, that shows important features for the lesion correct benign classification, and the raw lesion image

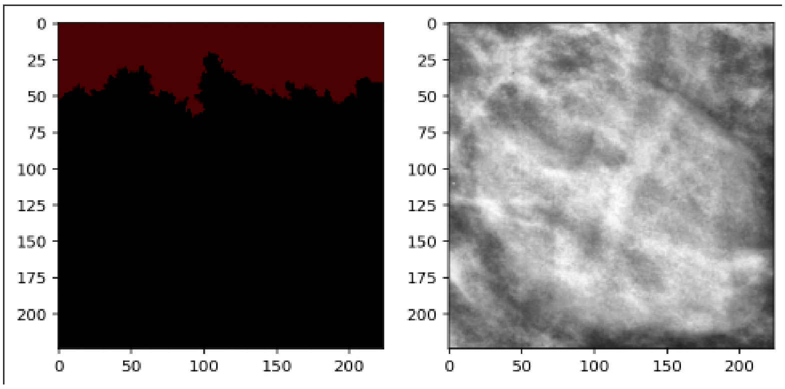


Figure 5.11. Explainability mask, generated with 1000 samples on LIME, that shows important features for the lesion wrong malignant classification, and the raw lesion image

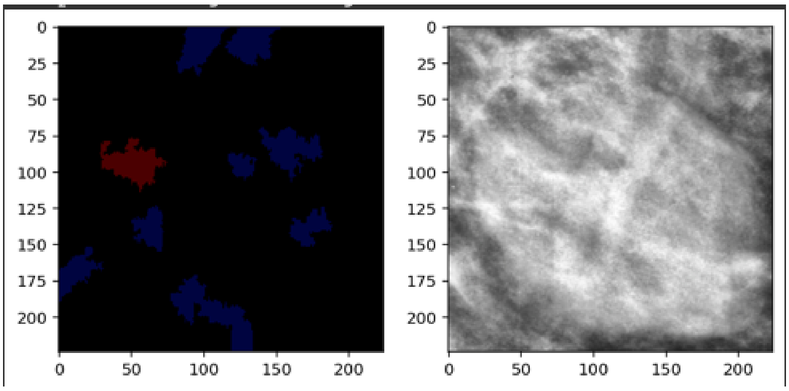


Figure 5.12. Explainability mask, generated with 5000 samples on LIME, that shows important features for the lesion wrong malignant classification, and the raw lesion image

highlighted. Only the important features for the final prediction have appeared, figure 5.13 shows the explanation for this case.

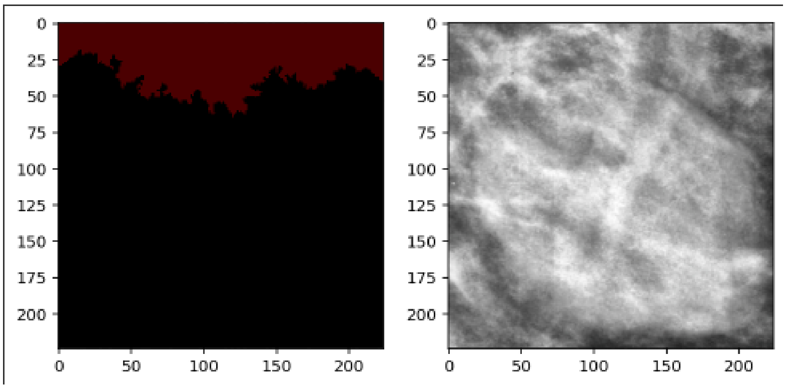


Figure 5.13. Explainability mask, generated with 10000 samples on LIME, that shows important features for the lesion wrong malignant classification, and the raw lesion image

5.1.2 RISE

The same images used for LIME explanation are used in this section for RISE, so it is possible to compare the two explanation cases for the same model classification in the three images. A correct benign image, a correct malignant, and a wrong malignant prediction, that actually was benign. The difference between the techniques can be highlighted while explaining the important features of the images. RISE explainability mask is a heat map, where the red areas are the most important part of the image, there are also other colors like green and blue, that are region less important.

The first image, figure 5.11 is malignant that was correctly predicted by the classifier, RISE has created an explanation mask mainly blue with points red in the lesion area and a strong red point in the lower left corner.

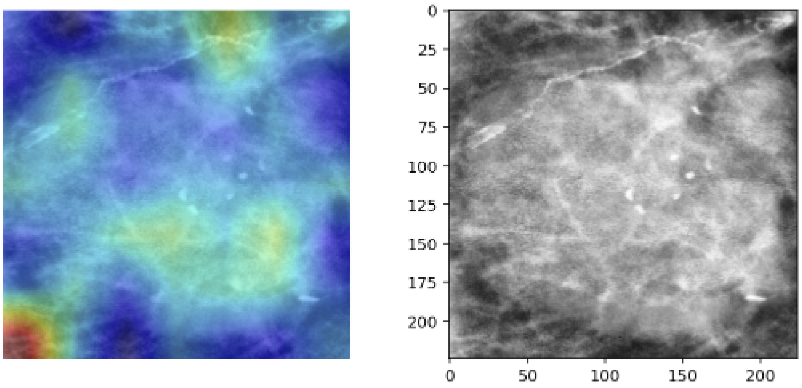


Figure 5.14. RISE Explainability mask, generated with 5000 samples, that shows important features for the lesion classification, and the raw lesion image

The second image, figure 5.12 is a benign lesion, that was also correctly classified by the classifier. The explanation area is red inside the lesion area and also around it. There are also green and blue areas, this can guide the most important part of the image for the classifier model.

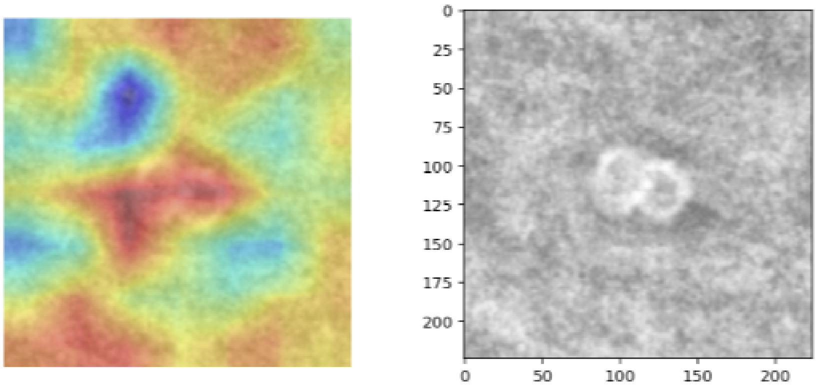


Figure 5.15. RISE Explainability mask, generated with 4000 samples, that shows important features for the lesion wrong malignant classification, and the raw lesion image

The third image, figure 5.13 is benign, and the model has predicted it as malignant. For this wrong classification, the explanation mask generated three red areas, and the lesion was highlighted in green and a little blue.

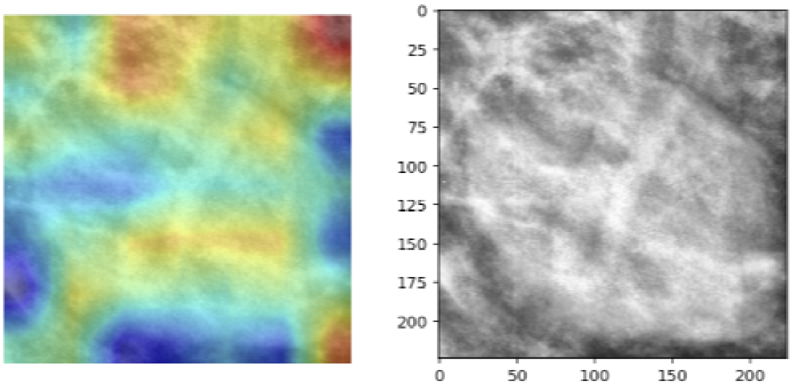


Figure 5.16. RISE Explainability mask, generated with 4000 samples, that shows important features for the lesion wrong malignant classification, and the raw lesion image

5.2 RISE vs LIME

After checking both explanation masks generated by RISE and LIME, it is possible to agree with the study of the authors [62] that have concluded that the explanation generated by RISE seems to be more meaningful. So it is possible to say, after a non-expert evaluation that RISE performs better on highlighting import areas for the VGG16 breast cancer classification model.

5.3 Future work

Considering next steps and future work, the results of highlighted areas can be analysed and be checked if they are really relevant and meaningful for experts in mammogram analysis. A classification model might be trained again with a bigger dataset in order to improve the metrics performance such as accuracy, sensitivity and precision. A lesion patter can be studied and be verified with the explanation quality with mammogram analysis expert, to check if the model's explanation are getting good results considering experts.

XAI can be used as a model feedback to the machine learning engineer who has trained it about how good the model is performing with highlighted features, so that it can guide de engineers on building better models.

6 Conclusion

Breast cancer is growing around the world, and having ways to help physicians get breast lesions classified correctly and make their daily work better, is a valid contribution. A CAD system may be used as a second reading on a mammogram, and so might reduce the need of having two radiologists to do a double mammogram reading in some regions that are difficult to have even one, having two in this case would be much harder.

Artificial Intelligence models in the medical field are very hard to build, because of the lack of publicly available datasets, especially with deep learning models. Deep Neural Networks need a lot of data to achieve a good performance, having a small dataset is a hard challenge when training a CNN. Mammogram analysis is a tough task to do, compared to other image analyses for cancer cases, like brain cancer. Even a non-expert person can check a brain MRI and see something different that might be cancer, different from breast lesions, which are difficult even for experts. The results of the breast lesion classifier are 68% of accuracy, and it could be improved. In order to do that, the number of images to train the model must be increased.

Explainable Artificial Intelligence is an outstanding way of clarifying AI models, intrinsically they are black-box and so it presents only the final output. Knowing features that are important and that have influenced the output, is more trustful when an explanation mask is shown with the predicted result. Among the XAI techniques, the ones used in this work were LIME and RISE, two explainable methods that used a perturbation-based approach, model agnostic, that produces local explanation and post-hot. Between both XAI techniques, RISE and LIME, RISE uses pixel explanation masks while LIME uses superpixels, that is a group of pixels to create important regions for highlights. The masks created by RISE seem to be more meaningful than the ones created by LIME, which confirms the study of [62].

The use of Artificial Intelligence in the medical field is promising, it might help the daily work of physicians and also help patients in diagnosing early-stage cancer, the early treatment will be less invasive, and increase chances of surviving when it is an early detected cancer.

References

[1] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[2] Hajar Mohammedsaleh H Alharbi, Paul Kwan, Ashoka Jayawardena, and ASM Sajeev. Fuzzy image segmentation for mass detection in digital mammography: Recent advances and techniques. In *Image Processing: Concepts, Methodologies, Tools, and Applications*, pages 769–792. IGI Global, 2013.

[3] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Mahmudul Hasan, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

[4] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[5] American Cancer Society. What is breast cancer? https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html, 2017. Accessed: 2018-10-20.

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[7] Silvia Bessa, Inês Domingues, Jaime S Cardosos, Pedro Passarinho, Pedro Cardoso, Vitor Rodrigues, and Fernando Lage. Normal breast identification in screening mammography: a study on 18 000 images. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 325–330. IEEE, 2014.

[8] Robyn L Birdwell, Debra M Ikeda, Kathryn F O'Shaughnessy, and Edward A Sickles. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. *Radiology*, 219(1):192–202, 2001.

[9] Breast Research Group. Get inbreast database. http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database. Accessed: 2018-11-02.

[10] Rachel F Brem, Janet Baum, Mary Lechner, Stuart Kaplan, Stuart Souders, L Gill Naul, and Jeff Hoffmeister. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. *American Journal of Roentgenology*, 181(3):687–693, 2003.

[11] Gustavo Carneiro, Yefeng Zheng, Fuyong Xing, and Lin Yang. Review of deep learning methods in mammography, cardiovascular, and microscopy image analysis. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, pages 11–32. Springer, 2017.

[12] Adam Jaamour; Ashay Patel; Shuen-Jen Chen. *Breast Cancer Detection in Mammograms using Deep Learning Techniques*. Master thesis, University of St Andrews, 2020.

[13] Hao Chen, Dong Ni, Jing Qin, Shengli Li, Xin Yang, Tianfu Wang, and Pheng Ann Heng. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE journal of biomedical and health informatics*, 19(5):1627–1636, 2015.

[14] Hiba Chougrad, Hamid Zouaki, and Omar Alheyane. Deep convolutional neural networks for breast cancer screening. *Computer methods and programs in biomedicine*, 157:19–30, 2018.

[15] I Christoyianni, A Koutras, E Dermatas, and G Kokkinakis. Computer aided diagnosis of breast cancer in digitized mammograms. *Computerized medical imaging and graphics*, 26(5):309–319, 2002.

[16] Stefano Ciatto, Marco Rosselli Del Turco, Gabriella Risso, Sandra Catarzi, Rita Bonardi, Valeria Viterbo, Pierangela Gnutti, Barbara Guglielmoni, Lelio Pinelli, Anna Pandiscia, et al. Comparison of standard reading and computer aided detection (cad) on a national proficiency test of screening mammography. *European journal of radiology*, 45(2):135–138, 2003.

[17] Antonio Cossu, Panagiotis Paliogiannis, Federico Attene, Giuseppe Palmieri, Mario Budroni, Ornelia Sechi, Carlo Torre, Francesco Tanda, and Fabrizio Scognamillo. Breast cancer incidence and mortality in north sardinia in the period 1992–2010. *Acta Medica Mediterranea*, 29(2):235–239, 2013.

[18] Inês Domingues, Pedro H Abreu, and Joäo Santos. Bi-rads classification of breast cancer: a new pre-processing pipeline for deep models training. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1378–1382. IEEE, 2018.

[19] Carl J D'Orsi. The american college of radiology mammography lexicon: an initial attempt to standardize terminology. *AJR. American journal of roentgenology*, 166(4):779–780, 1996.

[20] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[21] Joshua J Fenton, Linn Abraham, Stephen H Taplin, Berta M Geller, Patricia A Carney, Carl D'orsi, Joann G Elmore, William E Barlow, and Breast Cancer Surveillance Consortium. Effectiveness of computer-aided detection in community mammography practice. *Journal of the National Cancer institute*, 103(15):1152–1161, 2011.

[22] Joshua J Fenton, Stephen H Taplin, Patricia A Carney, Linn Abraham, Edward A Sickles, Carl D'Orsi, Eric A Berns, Gary Cutter, R Edward Hendrick, William E Barlow, et al. Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine*, 356(14):1399–1409, 2007.

[23] Jacques Ferlay, Clarisse Héry, Philippe Autier, and Rengaswamy Sankaranarayanan. Global burden of breast cancer. In *Breast cancer epidemiology*, pages 1–19. Springer, 2010.

[24] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017.

[25] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[26] Karthikeyan Ganesan, U Rajendra Acharya, Chua Kuang Chua, Lim Choo Min, K Thomas Abraham, and Kwan-Hoong Ng. Computer-aided breast cancer detection using mammograms: a review. *IEEE Reviews in biomedical engineering*, 6:77–98, 2012.

[27] Maryellen L Giger. Medical imaging and computers in the diagnosis of breast cancer. In *Photonic Innovations and Solutions for Complex Environments and Systems (PISCES) II*, volume 9189, page 918908. International Society for Optics and Photonics, 2014.

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[29] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[30] Mehmet A Gulum, Christopher M Trombley, and Mehmed Kantardzic. A review of explainable deep learning cancer detection models in medical imaging. *Applied Sciences*, 11(10):4573, 2021.

[31] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. 2009.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[33] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[34] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[35] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[36] Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019.

[37] Gary B Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[38] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[39] CW Huo, GL Chew, KL Britt, WV Ingman, MA Henderson, JL Hopper, and EW Thompson. Mammographic density—a review on the current understanding of its association with breast cancer. *Breast cancer research and treatment*, 144(3):479–502, 2014.

[40] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[41] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[42] Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, 3(3):740–770, 2021.

[43] Thijs Kooi, Geert Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*, 35:303–312, 2017.

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[45] Michele La Ferla, Matthew Montebello, and Dylan Seychell. An xai approach to deep learning models in the detection of ductal carcinoma in situ. *arXiv preprint arXiv:2106.14186*, 2021.

[46] Colton Ladbury, Reza Zarinshenas, Hemal Semwal, Andrew Tam, Nagarajan Vaidehi, Andrei S Rodin, An Liu, Scott Glaser, Ravi Salgia, and Arya Amini. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. *Translational Cancer Research*, 11(10):3853, 2022.

[47] Himabindu Lakkaraju, Richard Socher, and Chris Manning. Aspect specific sentiment analysis using hierarchical deep learning. In *NIPS Workshop on deep learning and representation learning*, 2014.

[48] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[49] Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, pages 143–155, 1989.

[50] Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, and Diana L Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11):1828–1837, 2015.

[51] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[52] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[53] Robert J McKenna Sr. The abnormal mammogram radiographic findings, diagnostic options, pathology, and stage of cancer diagnosis. *Cancer*, 74(S1):244–255, 1994.

[54] Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, 54(10s):1–40, 2022.

[55] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[56] Marvin Minsky and Seymour A Papert. *Perceptrons: An introduction to computational geometry.* MIT press, 2017.

[57] Marilyn J Morton, Dana H Whaley, Kathleen R Brandt, and Kimberly K Amrami. Screening mammograms: interpretation with computer-aided detection—prospective evaluation. *Radiology*, 239(2):375–383, 2006.

[58] National Cancer Institute. Mammograms. https://www.cancer.gov/types/breast/mammograms-fact-sheet, 2016. Accessed: 2018-07-24.

[59] National Cancer Institute. What is neural network flatten layer. https://www.educative.io/answers/what-is-a-neural-network-flatten-layer#, 2023. Accessed: 2023-12-07.

[60] Julian D Olden, Michael K Joy, and Russell G Death. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological modelling*, 178(3-4):389–397, 2004.

[61] Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiuwen Liu, and Zhe He. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27(7):1173–1185, 2020.

[62] Amy Rafferty, Rudolf Nenutil, and Ajitha Rajan. Explainable artificial intelligence for breast tumour classification: Helpful or harmful. In *Interpretability of Machine Intelligence in Medical Image Computing: 5th International Workshop, iMIMIC 2022, Held in Conjunction with MICCAI 2022, Singapore, Singapore, September 22, 2022, Proceedings*, pages 104–123. Springer, 2022.

[63] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[64] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[65] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):4165, 2018.

[66] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[67] Debbie Saslow, Carla Boetes, Wylie Burke, Steven Harms, Martin O Leach, Constance D Lehman, Elizabeth Morris, Etta Pisano, Mitchell Schnall, Stephen Sener, et al. American cancer society guidelines for breast screening with mri as an adjunct to mammography. *CA: a cancer journal for clinicians*, 57(2):75–89, 2007.

[68] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[69] LS Shapley. 20. quota solutions of n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 343–360. Princeton University Press, 2016.

[70] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald 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.

[71] Rebecca Smith-Bindman, Philip Chu, Diana L Miglioretti, Chris Quale, Robert D Rosenberg, Gary Cutter, Berta Geller, Peter Bacchetti, Edward A Sickles, and Karla Kerlikowske. Physician predictors of mammographic accuracy. *Journal of the National Cancer Institute*, 97(5):358–367, 2005.

[72] John Suckling, J Parker, D Dance, S Astley, I Hutt, C Boggis, I Ricketts, E Stamatakis, N Cerneaz, S Kok, et al. Mammographic image analysis society (mias) database v1. 21. 2015.

[73] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR, 2019.

[74] Lazaros Tsochatzidis, Lena Costaridou, and Ioannis Pratikakis. Deep learning for breast cancer diagnosis from mammograms—a comparative study. *Journal of Imaging*, 5(3):37, 2019.

[75] Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, page 102470, 2022.

[76] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, pages 1–19, 2018.

[77] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[78] Dingjun Yu, Hanli Wang, Peiqiu Chen, and Zhihua Wei. Mixed pooling for convolutional neural networks. In *International Conference on Rough Sets and Knowledge Technology*, pages 364–375. Springer, 2014.

[79] Chase Lipton Zachary. The mythos of model interpretability. *Communications of the ACM*, pages 1–6, 2016.

[80] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.