An Explainable Multi-Task Neural Network Model for Breast Cancer Detection in Ultrasound Images

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Computer Science

in the

College of Graduate Studies

University of Idaho

by

Mohammad Karimzadeh

Major Professors: Min Xian, Ph.D.

Committee Members: Aleksandar Vakanski, Ph.D.; Boyu Zhang, Ph.D.

Department Administrator: Terry Soule, Ph.D.

December 2023

Abstract

Breast cancer is one of the most common types of cancer and despite the advancements in the medical field, it still remains one of leading causes of death among women. In recent years, Breast Ultrasound (BUS) has emerged as a promising imaging modality for breast cancer diagnosis, offering high sensitivity and specificity. However, the interpretation of BUS images can be challenging and subject to human error. Computer-aided diagnosis (CAD) systems based on Machine Learning approaches have demonstrated a potential to assist radiologists in the interpretation of medical images. Nevertheless, the black-box nature of best performing CAD systems has raised concerns about their interpretability and trustworthiness.

This thesis introduces MT-BI-RADS (Multi-Task BI-RADS), a novel explainable deep learning approach for breast cancer detection in BUS images. The proposed approach provides three levels of explanations for enabling radiologists to understand the decision-making process in predicting the presence of tumors in BUS images. First, the deep learning model outputs the categories of the BI-RADS descriptors that are used by radiologists for BUS image analysis and reporting. Second, the deep learning model outputs segmented regions in images that correspond to tumors. And third, the proposed approach outputs quantified contributions of each BI-RADS category toward the prediction of benign or malignant class, based on post-hoc explanation with Shapley values. By utilizing a combination of explanations at different levels of abstraction using ad-hoc and post-hoc methods, the aim is to increase the transparency of the internal mechanisms employed by the deep learning model for analysis of BUS images.

Experimental validation on a large BUS dataset demonstrates that the proposed model achieved high accuracy, sensitivity, and specificity for breast cancer detection. In addition, the

proposed approach achieved high-level of performance for prediction of the BI-RADS descriptors and lesion segmentation. The design of explainable AI methods for cancer detection can enhance the trustworthiness and acceptance of CAD systems by clinicians.

Acknowledgments

I would like to express my gratitude to my amazing major professors, Dr. Min Xian and Dr. Aleksandar Vakanski, who have consistently provided support and assistance in overcoming the various challenges I encountered during my master's studies. Additionally, I appreciate Dr. Zhang for their guidance and assistance throughout this project.

Dedication

I would like to dedicate this thesis to my parents whom I have not been able to meet during the last two years, and also to my wonderful family in U.S. who unreservedly supported me through this journey.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

1.1 Breast Cancer Diagnosis

Breast cancer is one of the most common forms of cancer worldwide [1], affecting women and, less frequently, men [2]. According to the World Health Organization (WHO), breast cancer is the most common cancer in women both in the developed and developing world, with a staggering 2.3 million new cases diagnosed in 2020 alone [3]. Unfortunately, despite advances in medical technology and increased awareness, breast cancer remains a significant cause of mortality, accounting for approximately 11.7% of emerging cancer cases, and 6.9% of cancer deaths globally [4]. Early detection and accurate diagnosis of breast cancer are essential for improving patient outcomes and reducing mortality rates.

1.2 Breast Ultrasound (BUS)

Medical imaging plays a crucial role in the early detection and diagnosis of breast cancer. Mammography, Magnetic Resonance Imaging (MRI), and Ultrasound imaging are among the most commonly used imaging modalities for breast cancer detection [5]. Breast Ultrasound (BUS) is a non-invasive imaging technique that uses high-frequency sound waves to create images of breast tissue [6]. The BI-RADS (Breast Imaging Reporting and Data System) lexicon is commonly used by clinicians to interpret BUS images and classify breast abnormalities [7].

In the past, mammography was considered the most efficient method for detecting breast cancer at an early stage. Nevertheless, due to several advantages such as absence of radiation,

quicker imaging, greater sensitivity and accuracy, and reduced cost, ultrasound (US) imaging has emerged as a significant alternative to mammography [8].

1.3 BUS Imaging

Breast ultrasound imaging is a widely used diagnostic tool in clinical settings for detecting and diagnosing breast diseases. CAD system based on the US image involves the following steps: image preprocessing, feature extraction and selection, classification and/or segmentation. The classification task refers to classifying the images with tumors into benign or malignant category. The segmentation task refers to separating the lesion region from the background tissue [8]. Segmentation plays a significant role in image analysis and includes detection, feature extraction, classification, and treatment. Segmentation helps physicians quantify the volume of tissue in the breast for treatment planning [9]. In recent years, there has been significant research on developing breast ultrasound image analysis methods to improve diagnostic accuracy and assist clinicians in making better decisions [8].

1.4 Computer-Aided Diagnosis (CAD)

Computer-aided diagnosis (CAD) refers to the use of computer technology to assist medical professionals in interpretation of medical images. CAD systems have been widely used in various medical imaging applications, such as mammography, ultrasound, chest radiography, and computed tomography (CT) scans. CAD systems are designed to analyze medical images and provide a second opinion to radiologists and clinicians, improving diagnostic accuracy and efficiency [10, 11, 12].

In recent years, machine learning techniques have been increasingly applied to CAD systems to improve their performance. Machine learning algorithms such as support vector machines

(SVMs), artificial neural networks (ANNs), and random forests (RFs) have been used to analyze medical images and make diagnostic decisions. These machine learning approaches have shown promising results in improving the accuracy of CAD systems and reducing the workload of radiologists and clinicians [13,14].

Despite the potential benefits of CAD systems, their adoption in clinical practice has been limited. One of the main barriers to their adoption is the lack of transparency and interpretability of these systems [15]. Radiologists and clinicians may be hesitant to rely on CAD systems that provide limited or no information about the decision-making process. To address this issue, researchers have been developing explainable AI (XAI) techniques to improve the transparency and interpretability of CAD systems, which can help increase their acceptance by medical professionals [16].

1.5 Explainable AI

To overcome challenges resulting from the black-box nature of machine learning algorithms designed for CAD systems, researchers have developed explainable AI (XAI) techniques to understand and interpret machine learning models. By providing insights into the most important features and variables that drive the predictions, XAI techniques can enable humans to trust and validate the decisions made by these models. Some of these techniques include model-agnostic methods, such as LIME and SHAP, which provide insights into the features that are most important for a given prediction, and model-specific techniques, such as decision trees and rule-based systems, which generate human-readable explanations of the decision-making process [17].

In the medical field, XAI is of particular importance as it can help doctors and clinicians to understand how AI systems make predictions, ultimately leading to better clinical decision-making [18]. XAI techniques can help to increase their acceptance by medical doctors and healthcare systems. With greater trust and understanding of AI systems, doctors may be more willing to adopt and utilize these systems in their clinical practice, leading to better patient outcomes and improved healthcare delivery.

1.6 Proposed Method

This thesis introduces an explainable deep learning model, MT-BI-RADS, for breast cancer detection using BUS images. The proposed model aims to provide transparent explanations for its predictions, enabling clinicians to understand how the model arrived at its decisions and increasing their confidence in the model's accuracy.

The proposed explainable deep learning (DL) model uses a hybrid ad-hoc and post-hoc explainability approach and provides both visual and quantitative explanations. Specifically, the model outputs the predicted category of each BI-RADS descriptor, as well as providing visual explanations via highlighting tumor regions in the images by employing image segmentation. In addition, SHAP technique is applied to provide post-hoc quantitative explanations of the significance of each BI-RADS descriptor toward the model predictions.

To evaluate the performance of the proposed model, we conducted experiments using a relatively large dataset of BUS images. The results demonstrate that the model achieves high accuracy and interpretability, with high sensitivity and specificity.

The proposed explainable deep learning model for breast cancer detection has the potential to significantly improve the accuracy and interpretability of BUS image analysis for breast cancer

detection. The transparent explanations provided by the model can help medical professionals to better understand the diagnostic process and increase their confidence in the use of machine learning models for cancer detection.

1.7 Thesis Contributions

The thesis offers several contributions, as follows.

- Development of an ad-hoc explainable multitask learning approach for breast cancer detection that concurrently outputs BI-RADS descriptors, tumor segmentation masks, and the tumor class.
- Improvement in the predictive abilities of the model by incorporating branches for BI-RADS descriptors classification and tumor segmentation into a multi-task learning framework.
- Providing post-hoc explanations using Shapley Values to quantify the contributions of the BI-RADS descriptors to the tumor classification into benign or malignant categories.

1.8 Thesis Organization

The thesis is organized as follows. Chapter 2 reviews related works, and Chapter 3 describes the architecture of the proposed deep learning model and used explainability approaches. Chapter 4 presents the results of experimental validation of the proposed approach and describes the explainability of the model. Chapter 5 provides a summary and concludes the thesis.

CHAPTER 2: RELATED WORKS

2.1 Breast Ultrasound Image Analysis

Breast Ultrasound is a non-invasive imaging technique that uses high-frequency sound waves to produce images of the internal structures of the breast. It is particularly useful in evaluating breast lumps or abnormalities that are detected on a mammogram or clinical breast exam. BUS can help determine whether a lump is solid or fluid-filled, and whether it is benign or malignant, or for evaluation of breast abnormalities. In recent years, there has been increasing interest in the use of BUS for breast cancer screening and diagnosis. The ease of use and real-time imaging capability make BUS a method of choice for guiding breast biopsies and other interventional procedures. The developments in BUS imaging tools, combined with the formulation of a standardized lexicon of solid mass features, have improved the diagnostic performance [19].

In the past few decades, numerous ML and DL approaches were developed for analysis of BUS images to categorize the images as normal or abnormal. In a study conducted by Liao et al. [20], a deep learning-based classification algorithm was utilized to differentiate between benign and malignant breast lesions in BUS images. The algorithm demonstrated high accuracy, achieving an area under the curve (AUC) of 0.98. Almajalid et al. [21] modified the U-Net architecture which is frequently for medical image segmentation, to create a robust and accurate tumor segmentation network for BUS images. Kalaf et al. [22] presented an architecture for breast cancer classification with an attention mechanism in an adapted VGG16 framework that distinguishes between the features of the background and targeted lesions. Tanaka et al. [23] designed a CAD scheme to classify benign and malignant tumors using an

ensemble of two CNN architectures. In another study [24], a Bayesian network based on Monte Carlo dropout was proposed to reduce accidental false positives and achieved a sensitivity of 0.88. In [25], radiomic features were extracted from breast lesion images to aid in the tumor segmentation. The model used a convolutional deep autoencoder to simultaneously perform segmentation of the breast lesions and extract radiomic features from the segmented regions. The authors in [26] proposed a mobile phone-based model that takes a photo of the ultrasound report as input and performs diagnosis on each image in three steps: reducing noise in the taken photo, classifying the image into malignant or benign, and detecting anomalies in the model performance to reduce false negative rates. Chen et al [27] utilized a convolutional neural network (CNN) to segment breast lesions in BUS images.

2.2 Explainability in ML

The recent progress in Machine Learning algorithms has facilitated the resolution of numerous problems that were once challenging for humans to solve [28, 29]. Decision tree, linear regression, SVM, and other ML algorithms can produce models that make highly accurate predictions. However, they fall short in scenarios where large amounts of training data are involved or when tackling complex issues, such as image classification and speech recognition. Deep learning models have demonstrated ability to process vast amounts of complex data and resolve intricate challenges. However, a major disadvantage of deep learning models, in contrast to ML algorithms like decision trees, is their lack of interpretability, thereby are referred to as "black-box" models. This is because the model's calculations and the basis of the predictions are not transparent to the model developer. In contrast, a prediction made by a decision tree, for example, can be tracked to examine the steps that led to the model's ultimate prediction. Although accuracy is paramount in developing an ML model, in certain sectors

such as healthcare, finance, and criminal justice, the transparency and interpretability of a model's predictions are extremely important to ensure that relevant experts and stakeholders have a clear understanding of the behavior of the model [30]. This will allow them to make informed decisions and avoid making costly errors that could impact people's health, company finances, and more [31, 32, 33].

The approaches to interpretability and explainability in ML are broadly categorized into ad-hoc (integrated) and post-hoc methods [28]. Integrated methods are based on using transparent ML approaches like linear regression, that are interpretable by the virtue of their design. Or, for more complex tasks using a hybrid approach, such as a combination of logistic regression and SVMs, can provide enhanced interpretability along with making up for the original model’s performance [28].

Post-hoc explainable approaches extract information from a trained model to provide explanations. Thus, this class of approaches does not compromise a black-box model’s performance [28]. In these approaches the trained model is inspected to find a set of parameters that illuminate how the model calculates the predictions. This category usually includes explanations using visualizations of internal features [34, 35, 36], visualizations of local regions [37, 38, 39, 40], or textual explanations [41, 42, 43]. Many post-hoc methods are model agnostic and can be applied to provide explanations regardless of the trained model [44]. Among the many introduced post-hoc methods, SHAP [37] has been commonly used for interpreting model predictions by assigning an importance value to the contribution of each feature to a particular prediction.

2.3 Explainable ML in Medical Imaging

Despite numerous efforts to utilize ML models for medical image analysis, these methods have yet to be widely embraced by healthcare professionals, primarily due to a lack of transparency in those models' predictions. As healthcare is a high-risk domain, it is imperative that experts have a full understanding of the models to make informed decisions that mitigate potential consequences, as there are concerns about the black-box nature of the developed ML methods in medical image analysis [44].

In the published literature, prior works have used saliency maps, such as CAM [45], GradCAM [46], and other visualization methods, to provide visual explanations of the models predictions [47, 48]. In [49] a CNN was developed that visualized the last convolution layer in order to explain the model developed for the classification of liver lesions. In [50, 51] the authors developed a network for breast cancer detection and used Grad-CAM to highlight the important regions of the input image. In [52] a convolution feature extractor along with a SVM classifier were implemented to classify images with signs of prostate cancer and generated CAM heatmaps to display the important regions of the lesions. The goal in these works is to explain the predictions by visualizing the regions of the image that lead to the final classification. Even though these visualizations can provide some level of explanation, highlighting a specific part of image does not provide full explanation, because there might be more than one pattern present in that specific part of image, and highlighting a region cannot explain if the model has been able to capture all those patterns or not.

In recent years, there has been growing interest in developing explainable deep learning models for medical image analysis, particularly for cancer detection [53, 54, 55, 56, 57]. Explainable AI (XAI) has emerged as a promising approach for addressing the interpretability and

explainability issues of machine learning models, making them more accessible and trustworthy to clinicians and several works have focused on making explainable models for CAD [58, 59, 60].

Several related studies introduced XAI methods for breast cancer detection based on the BI-RADS lexicon. For example, Shen et al. [61] introduced an explainable ML classifier for locating suspected lesions in mammograms. Wu et al. [62] proposed DeepMiner, a DL architecture for tumor classification that uses BI-RADS descriptors for generating text explanations in mammography. Also, Kim et al. [63] developed DL models that utilized the tumor shape and margin in mammograms to predict the class label and BI-RADS category. One major drawback of these approaches in mammography is that they rely on only two or three BI-RADS descriptors, which may not provide enough information to fully elucidate the complex process of tumor classification.

Among the XAI methods for BUS images, Zhang et al. [64] employed only the shape and margin descriptors to predict the tumor class in BUS images. In [65], an ensemble model was proposed with explanations based on statistical texture features of BUS images, which are less useful for radiologists. Also, approaches that concentrated on generating textual reports for explaining NN models for BUS [66], were proposed in the literature. BI-RADS-Net [67] proposed an ad-hoc explainable DL model for tumor classification and an extension of that work, BI-RADS-NET V2 [68], incorporates a quantitative explainer that employs a linear model to provide an interpretable alternative to the "black-box" classifier. Despite the advancements by the efforts in related works, the explainability of CAD systems for breast cancer diagnosis is still an open research problem that requires further attention.

CHAPTER 3: METHODS

3.1 BI-RADS Lexicon

The Breast Imaging Reporting and Data System (BI-RADS) lexicon is a standardized reporting system used by radiologists to communicate the findings of mammography, ultrasound, and MRI breast imaging studies. It was developed by the American College of Radiology (ACR) to improve communication between radiologists and referring physicians and to ensure consistent and accurate interpretation of breast imaging studies.

The BI-RADS lexicon uses standardized terminology to describe the features of breast lesions, including size, shape, margins, density, and other characteristics. Each lesion is assigned a category based on the likelihood of malignancy, ranging from 0 (incomplete evaluation) to 6 (known biopsy-proven malignancy).

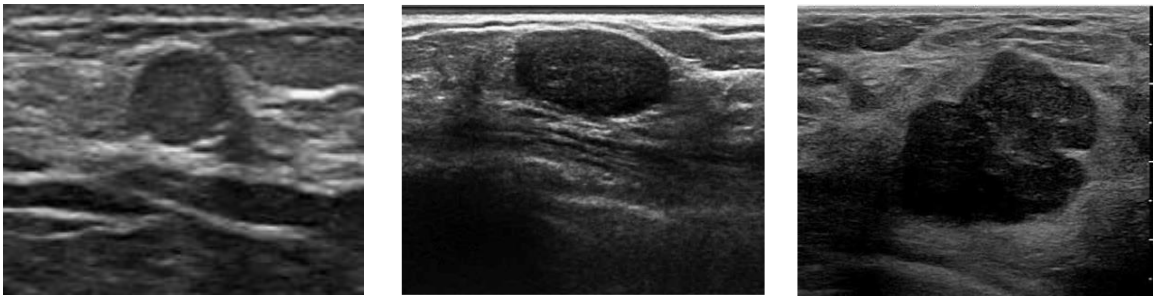
In simpler terms, the BI-RADS lexicon is a standardized language that radiologists use to describe breast imaging studies to other healthcare professionals. It helps ensure that everyone is using the same terminology and that the findings are accurately communicated. It also helps guide clinical decision-making by providing information about the likelihood of malignancy for each lesion.

For BUS images, the BI-RADS descriptors are listed in Table 3-1. They include shape, orientation, margin, echo pattern, and posterior features. Examples of BUS images for the BI-RADS descriptors are presented in the following figures.

Table 3-1 - Descriptors for BUS images

BI-RADS Descriptors	**Descriptors Class**
Shape	Oval, Round, Irregular
Orientation	Parallel, Not parallel
Margin	Circumscribed, Not circumscribed (Indistinct, Angular, Micro-lobulated, Spiculated)
Echo Pattern	Anechoic, Hypoechoic, Isoechoic, Hyperechoic, Complex cystic and solid, Heterogeneous
Posterior Features	No posterior features, Enhancement, Shadowing, Combined pattern

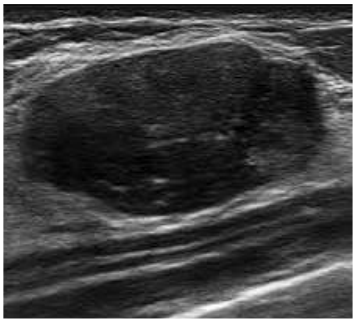
Shape: This BI-RADS descriptor describes the shape of a mass seen on breast imaging, such as mammography or ultrasound. The possible shapes include round, oval, and irregular. Figure 3-1 shows different shapes in breast tumors.



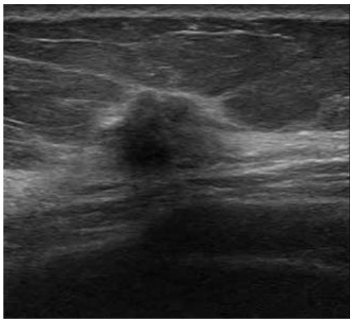
a) Round shaped tumor b) Oval shaped tumor c) Irregular shaped tumor

Figure 3-1 Examples of breast tumors categorized by the defined shape patterns in the BI-RADS lexicon. The shape patterns include round, oval, and irregular, which describe the shape of a mass seen in a breast image.

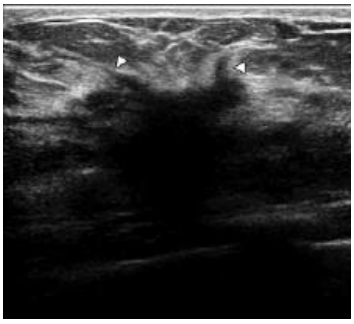
Margin: This BI-RADS descriptor describes the edges of a mass seen in breast imaging. The possible margin types include circumscribed and not-circumscribed, which itself divides into four subcategories of micro-lobulated, angular, indistinct, and spiculated margin (Figure 3-2).



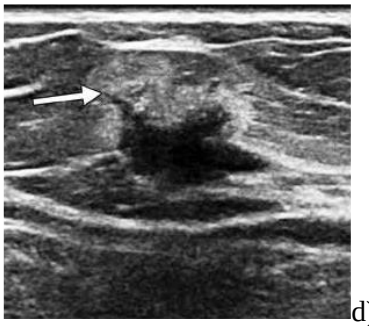
a) Circumscribed margin



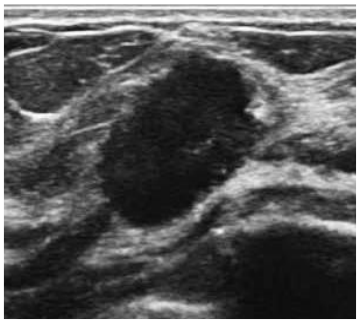
b) Indistinct margin



c) Spiculated margin



d) Angular margin



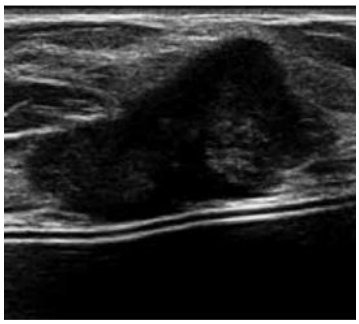
e) Micro-lobulated margin

Figure 3-2 Examples of breast tumors categorized by the defined margin patterns in the BI-RADS lexicon. The margin patterns include circumscribed and not-circumscribed, which divides into indistinct, spiculated, angular, and micro-lobulated categories, describing the edge of a mass seen on a breast image.

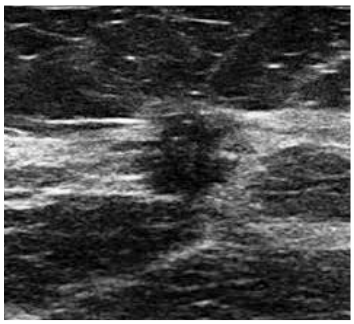
Orientation: This BI-RADS descriptor describes the direction of a mass in relation to the surrounding breast tissue. The possible orientations include parallel to the skin and not parallel to the skin (Figure 3-3).

Echo pattern: This BI-RADS descriptor refers to the appearance of echoes on a breast ultrasound image (Figure 3-4). The ultrasound machine sends high-frequency sound waves through the breast tissue, and the echoes that bounce back are then translated into an image on the screen. The echo pattern is a description of the density and distribution of the echoes within the breast tissue.

Posterior features: This BI-RADS descriptor describes the appearance of the tissue behind a mass or lesion seen on ultrasound. The possible features include posterior acoustic enhancement (increased brightness behind the mass), posterior acoustic shadowing (decreased brightness behind the mass), no posterior features, or a combination of both enhancement and acoustic shadowing (Figure 3-5).

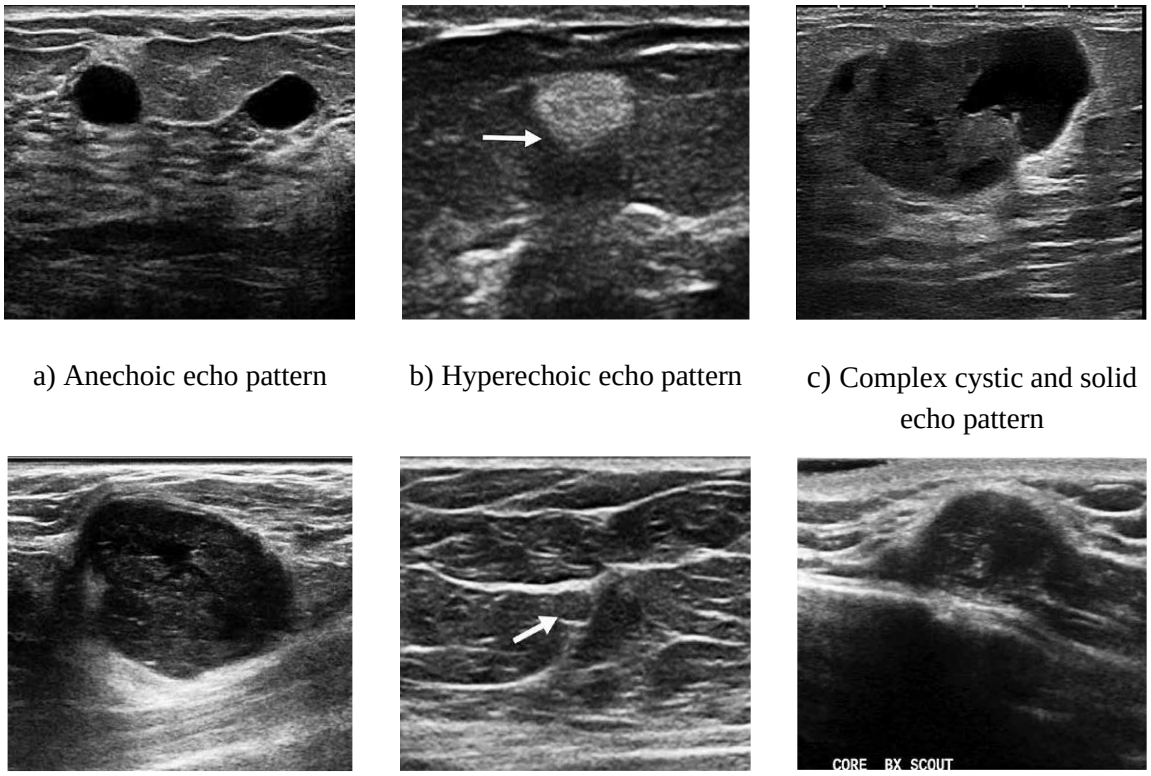


a) Tumor parallel to the skin



b) Tumor not-parallel to the skin

Figure 3-3 Examples of breast tumors categorized by the defined orientation patterns in the BI-RADS lexicon. The orientation patterns include parallel and not-parallel, which describe the alignment of a mass seen on a breast image.



a) Anechoic echo pattern

b) Hyperechoic echo pattern

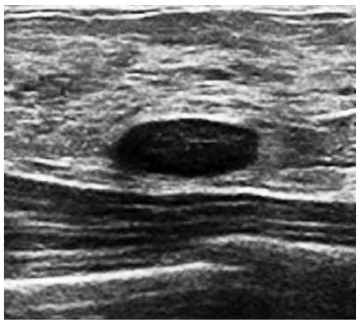
c) Complex cystic and solid echo pattern

d) Hypoechoic echo pattern

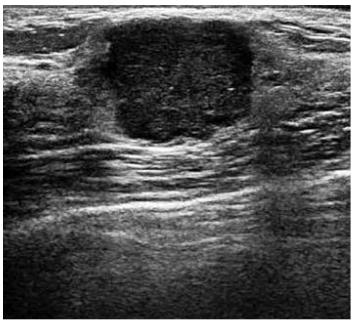
e) Isoechoic echo pattern

f) Heterogeneous echo pattern

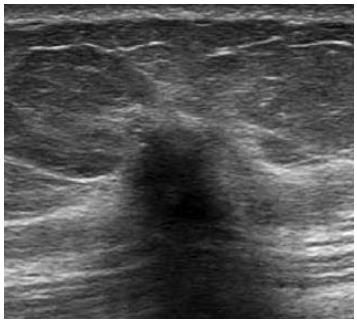
Figure 3-4 Examples of breast tumors categorized by the defined echo patterns in the BI-RADS lexicon. The echo patterns include anechoic, hyperechoic, complex, hypoechoic, isoechoic, and heterogeneous, which describe the internal echo texture of a mass seen on a breast ultrasound.



a) No posterior features



b) Enhancement posterior feature



c) Shadowing posterior feature

d) Combined posterior features

Figure 3-5 Examples of breast tumors categorized by the defined posterior features in the BI-RADS lexicon. The posterior feature patterns include no posterior feature, posterior enhancement, shadowing, and combined features, which describe the characteristics of the posterior aspect of a mass seen on a breast ultrasound.

The BI-RADS descriptors are crucial in assisting medical experts to identify the type of tumor and decide on the appropriate course of action to ensure the patient's optimal outcome. These standardized reporting tools allow physicians to precisely convey the tumor's properties and cooperate more efficiently with their peers to deliver complete care. As a result, the lexicon promotes a streamlined and productive decision-making procedure that ultimately leads to enhanced patient results.

The BI-RADS assessment categories are used to standardize the interpretation and reporting of breast imaging studies, including mammography, ultrasound, and MRI. There are six assessment categories in the BI-RADS system, ranging from 0 to 5, with additional sub-categories that can be used to provide more specific information about the findings. Category 0 is an incomplete assessment, indicating that additional imaging evaluation or comparison with prior studies is needed to make a final assessment. Category 1 is a negative assessment, indicating that there are no suspicious findings. Category 2 is a benign assessment, indicating the presence of benign findings such as cysts or lymph nodes. Category 3 is a probably benign assessment, indicating that there is a low likelihood (less than or equal to 2%) of malignancy, but short-term follow-up is recommended to ensure stability. Category 4 is a suspicious assessment, indicating that there are findings suspicious for malignancy, and biopsy or other further evaluation is recommended. Category 5 is a highly suggestive of malignancy assessment, indicating a high likelihood of malignancy, and biopsy or other appropriate action is recommended [69]. A summary of the 6 categories is provided found in Table 3-2.

It is worth highlighting that the BI-RADS assessment categories do not provide a definitive diagnosis but serve as a framework to assist radiologists in evaluating breast abnormalities. The ultimate diagnosis and management strategy should be determined based on various factors, including imaging results, clinical data, and input from other healthcare professionals, and should involve discussions with the patient. The use of the BI-RADS system helps guarantee that an adequate level of assessment and follow-up is performed for every patient according to their imaging outcomes.

3.2 Data

The proposed method was evaluated using 2,186 BUS images which were obtained by combining three different data sets: BUSI [70], BUSIS [71] and HMSS [72] dataset. The BUSIS dataset consists of 562 images, of which 306 images contain benign masses in 256 contain malignant tumors. For the BUSI dataset, we used a subset of 630 images that contain

Table 3-2 - BI-RADS assessment categories

BI-RADS Category	Assessment	Malignancy Risk
0	Incomplete/Need further imaging evaluation	N/A
1	Negative	<1%
2	Benign	<1%
3	Probably benign	2-10%
4A	Low suspicion of malignancy	2-10%
4B	Moderate suspicion of malignancy	10-50%
4C	High suspicion of malignancy	50-95%
5	Highly suggestive of malignancy	>95%
6	Known biopsy-proven malignancy	N/A

mass findings, of which 421 have benign and 209 have malignant tumors. HMSS consists of 1,700 images, out of which 720 are benign samples and the other 994 are malignant.

The patterns of tumors in BUS images of benign samples are typically characterized by being parallel with the skin surface, oval-shaped, and possessing a circumscribed margin. In contrast, malignant samples exhibit a more diverse range of appearance and can vary significantly from image to image. Accordingly, learning features representations of malignant images is more challenging for the ML models in comparison to benign images. Given the limited number of

breast ultrasound images available from the three datasets, we used solely the malignant samples from the HMSS dataset with the other two datasets. Consequently, the models were evaluated on all BUSIS 562 and BUSI 630 images, as well as the 994 malignant images from the HMSS dataset.

For more information on the BUSI, BUSIS, and HMSS datasets, please refer to their respective sources [70, 71, 72].

3.3 Network Architecture

The proposed model MT-BI-RADS is an extension of BI-RADS-NET V1 and BI-RADS-NET V2 [68], which were also developed by our team. To provide a historical perspective in the development of the proposed method, we will first briefly introduce BI-RADS-NET V1 and BI-RADS-NET V2.

3.3.1 BI-RADS-NET V1

The structure of the BI-RADS-NET V1 is shown in Fig 3-6. It consists of two main parts: a shared backbone network and specific networks for each task, which include branches for predicting the BI-RADS descriptors, BI-RADS assessment category (i.e., likelihood of malignancy), and tumor category. The backbone network utilizes convolutional and max-polling layers to extract important features from the input BUS images. The learned feature maps are used by the BI-RADS descriptors branch to predict the five descriptors from Table 3-1. The outputs of the BI-RADS descriptors are combined with the feature maps from the base network and passed on to a regression branch to determine the likelihood of malignancy, which outputs a continuous value ranging from 0% to 100%. The tumor classification branch

combines the features maps from the backbone network and the other two branches to output a binary benign or malignant class label.

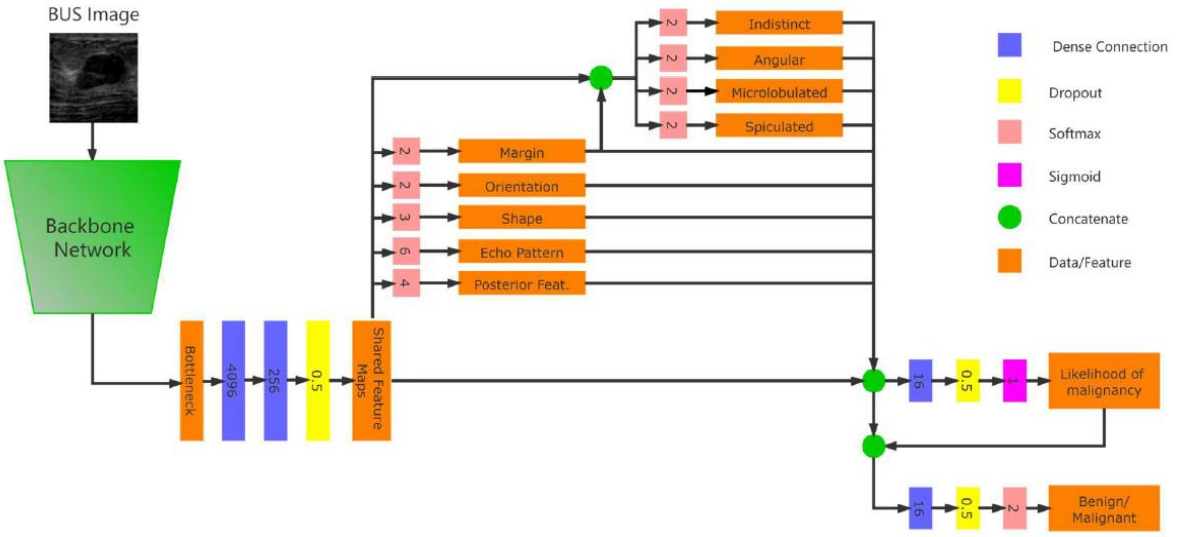


Figure 3-6 Architecture of BI-RADS-NET V1

The BI-RADS descriptors have specific ground-truth labels listed in Table 3-1, which include two classes for shape, three classes for orientation, six classes for echo pattern, and four classes for posterior features. Since the margin descriptor may have multiple annotations, the first branch for the margin in BI-RADS-NET V1 has only two classes (circumscribed and not-circumscribed), and another sub-branch is introduced afterward that outputs binary values for the indistinct, angular, micro-lobulated, and spiculated margin sub-classes.

The likelihood of malignancy branch outputs continuous values corresponding to the BI-RADS assessment categories shown in Table 3-2, with the median likelihood of malignancy adopted as follows: Category 3 – 1%, Category 4A – 6%, Category 4B – 30%, Category 4C – 72.5%, and Category 5 – 97.5%. Using a regression branch to predict continuous values for the likelihood of malignancy rather than categorical variables helps the network deal with inter-

observer variability in the BI-RADS category labels. It is also worth noting that the used datasets do not contain images with BI-RADS 0, 1, 2, or 6 categories.

In the multitask model, Task 1 to 5 are the BI-RADS descriptors, Task 6 to 9 are the sub-classes for the margin BI-RADS descriptor, Task 10 is the BI-RADS likelihood of malignancy, and Task 11 is the tumor classification branch. For each task k, the network loss function is denoted by $\mathcal{L}_k(X_k, Y_k)$, where X_k is the predicted value and Y_k is the ground-truth label (for classification) or value (for regression). Since the outputs of the likelihood of malignancy branch (Task 10) and the tumor classification branch (Task 11) both reflect the level of risk that the present tumor in the image is malignant, we added an additional loss term $\mathcal{L}_a$ to enforce an agreement between the two branches. The total loss of the model is calculated as the weighted sum of all tasks, that is, $\sum_{i=1}^{K} \lambda_i \mathcal{L}_i(X_i, Y_i) + \lambda_a \mathcal{L}_a\left(|Y_{11} - X_{10}|, |Y_{11} - Y_{10}|\right)$ [67]. The symbol λ_i denotes the weight coefficient of task i, K = 11 is the number of tasks, and λ_a is the weight coefficient for the $\mathcal{L}_a$ term

3.3.2 BI-RADS-NET V2

The second version of BI-RADS-NET introduced a quantitative explainer, that approximates the classifier, considered a 'black-box', with an explainable linear model. Fig 3-7 shows the architecture of BI-RADS-NET V2. The quantitative explainer predicts two weight vectors for benign and malignant descriptors, respectively, and calculates the dot products between the feature groups and the predicted weights. The outputs of the explainer, corresponding to benign and malignant, are expected to be equal to the classifier output before the final SoftMax layer. The residual is defined as the average differences between the explainer output and the classifier output on benign and malignant, reflecting the similarity between them. The explainer's contribution of each descriptor is evaluated by the corresponding weight. The

different BI-RADS descriptors are not equally important in the diagnostic process, and the explainer provides a way to evaluate the contribution of each descriptor.

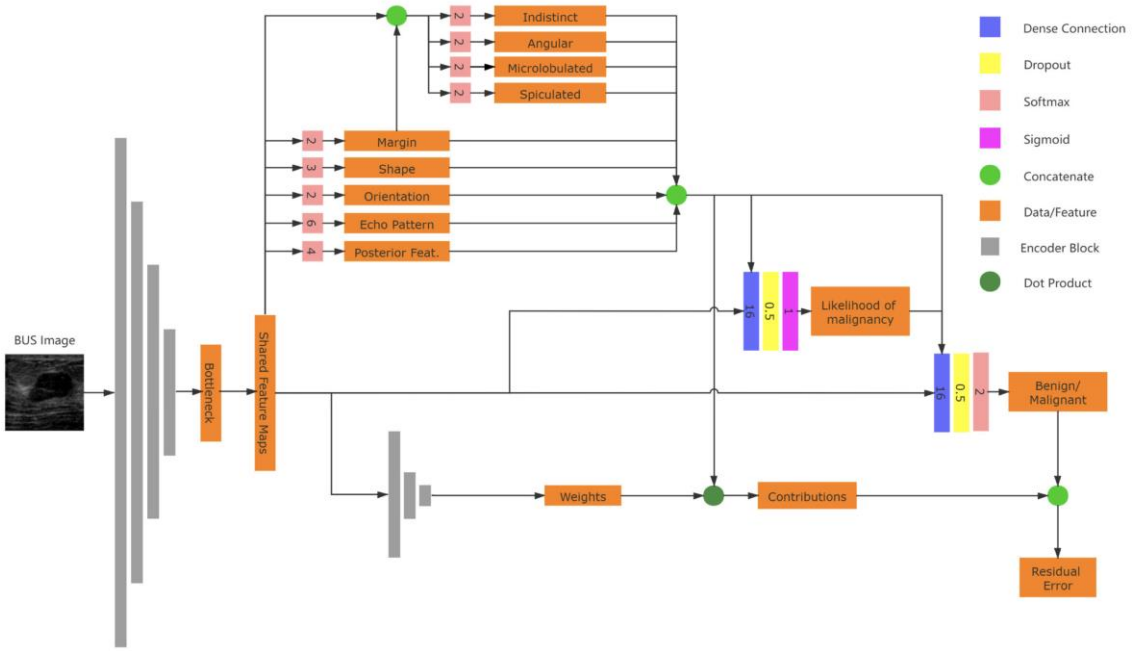


Figure 3-7 Architecture of BI-RADS-NET V2

3.3.3 MT-BI-RADS

The explainable network proposed in this thesis is MT-BI-RADS, which is a revised version of BI-RADS-NET V1, with the aim to provide a greater level of explainability to the users for understanding the predictions made by the model. The architecture of MT-BI-RADS is depicted in Fig 3-8. One important addition in the new architecture is a segmentation branch to predict the tumor region in BUS images. The outputs of the proposed model provide information to radiologists regarding the model's ability to differentiate the tumor region from the surrounding tissue in an ultrasound image and the area being considered as tumor region. Also, since the primary goal of this work is to explain the role of different BI-RADS descriptors

in the final prediction of the model, the BI-RADS category branch was removed from the architecture.

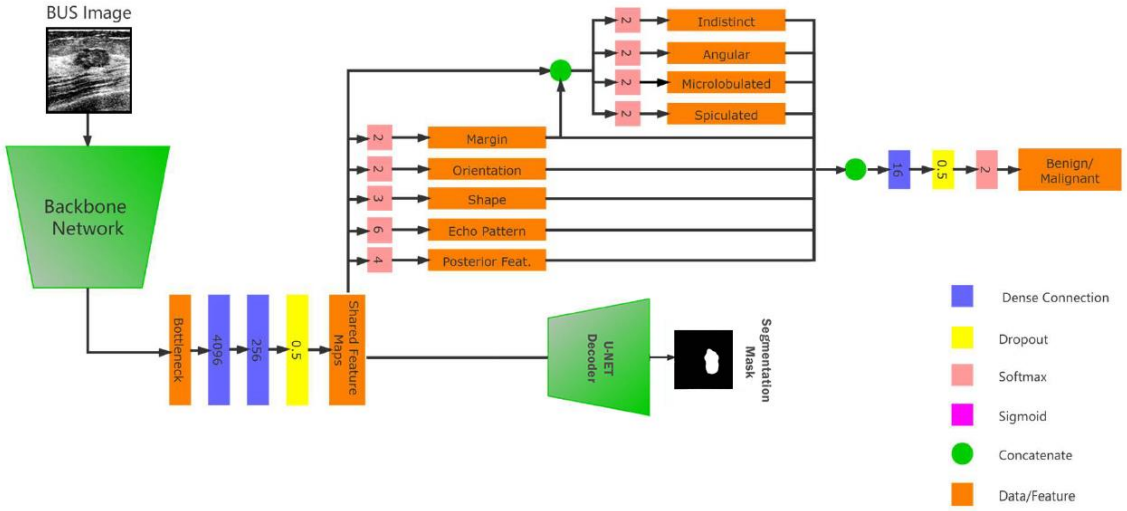


Figure 3-8 Architecture of the proposed model

The proposed architecture offers explainability in the context of tumor detection by providing a comprehensive set of information about the input image. The system classifies each BI-RADS descriptor into one of several categories and reports the corresponding probability of each category. Additionally, it produces a segmentation mask that highlights the region of the input image corresponding to the detected tumor.

In this model architecture, Task 1 to 5 are the BI-RADS descriptors, Task 6 to 9 are the sub-classes for the margin BI-RADS descriptor, Task 10 is the tumor classification branch, and last but not least, Task 11 is the image segmentation. The loss function for each task is denoted by $\mathscr{L}$, where X is the predicted value and Y is the ground-truth label (for classification) or value (for regression). The total loss of the model is calculated as the weighted sum of all tasks, with

λ denoting the weight coefficient of each task. Cross-entropy is used as loss function for the classification branches and Dice Loss with smoothness of 1 is used for the decoder (semantic segmentation) branch.

$$\sum_{i=1}^{K} \lambda_i \mathcal{L}_i(X_i, Y_i) + \lambda_{11} \mathcal{L}_{Dice}$$

The symbol λ_i denotes the weight coefficient of task i, K is the number of tasks except image segmentation, and λ_{11} is the weight coefficient for the dice loss term. For the loss weight coefficients λ_1 to λ_{11}, we adopted the following values: (0.2, 0.2, 0.2, 0.2, 0.2, 0.1, 0.1, 0.1, 0.1, 0.5, 0.6). That is, the largest weight was assigned to the tumor class branch.

To enhance the explainability this framework utilized Shapley Values to calculate the contribution of the different BI-RADS descriptors to the final classification of the input samples. This explanation can provide the medical staff with a clearer idea of how much emphasis the model exerts on the different BI-RADS descriptors when predicting the malignancy or benignancy of a tumor. In the next subsection we will elaborate on the implementation of this feature.

3.4 Shapley Values

Shapley values is a technique used in cooperative game theory to distribute the collective reward earned by a group of players among themselves in a just and equitable manner. The idea was initially proposed by economist Lloyd Shapley back in 1953 [74].

In the domain of data science and machine learning, Shapley values are utilized to explain the results of a model by assessing the contribution of each feature towards the final outcome.

Shapley values permit the estimation of the significance of individual features in a prediction, taking into consideration the interactions between them.

To compute the Shapley value for a specific feature, all feasible alliances of features must be taken into account, and the marginal contribution of the feature to the prediction for each group must be calculated. The Shapley value for the feature is the average of these marginal contributions across all conceivable alliances. The Shapley value provides an average measurement of its contribution to the prediction compared to all possible feature combinations. This approach helps evaluate the relative significance of various features in a model and pinpoint the features that have the most influence on the predictions.

Shapley values have gained widespread acceptance in machine learning to explain complex models, including neural networks and decision trees. By offering a method to comprehend the individual features' contributions to the prediction, Shapley values can help build confidence in the model and improve decision-making transparency.

One popular tool for calculating the Shapley Values is SHAP (SHapley Additive exPlanations) [37] which is a popular open-source library that build upon Shapley Values and externs them to the problem of interpreting and explaining the output of machine learning models.

The formula for calculating the Shapley value of a feature *i* in a given instance *x* is given below [74]. Intuitively, the Shapley value of a feature *i* represents the average contribution of that feature to all possible coalitions in a game. It considers all possible orderings of the features and computes the marginal contribution of feature *i* when it is added to each possible subset of features. The Shapley value is then the weighted average of these marginal contributions, where the weights are given by the combinatorial factor [74]:

$$\phi_i(f) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [\, f(S \cup \{i\}) - f(S)]$$

where $\phi_i(f)$ is the Shapley value of feature i in instance f. The sum is taken over all subsets S of features that do not include i, and $f(S \cup \{i\}) - f(S)$ is the marginal contribution of the feature i when it is added to the subset S.

In the proposed approach, we used SHAP as a post-hoc approach to calculate the contribution of the BI-RADS descriptors to the final output for the tumor class. Thus, a submodel was extracted from a trained BI-RADS NET V3, whereas the outputs of the BI-RADS predictors in the MT-BI-RADS are the inputs and the output of the tumor class branch is the output. The extracted submodel is used by SHAP to determine what is the marginal contribution of each input of the submodel (each BI-RADS descriptor) to the output of submodel (tumor class).

CHAPTER 4: EXPERIMENTAL RESULTS

4.1 Data Preprocessing

The original BUS images in the dataset were of various sizes and needed to be resized before used for modeling. Unlike general object recognition tasks, resizing BUS images can affect the tumor's shape and orientation labels, as the morphological features of the tumor can be distorted. To prevent this distortion, the original BUS images were first cropped to the largest square segment that encompassed the tumor and then resized to 256x256 pixels. Additionally, for the single-channel grayscale BUS images, two more channels were added. One channel was created by applying histogram equalization, and another by smoothing. This simple pre-processing step proved to improve the model's performance reducing variations across the images in the BUSIS, BUSI, and HMSS datasets and resulting in a more uniformly distributed set of images.

4.2 Cross Validation

In our experiments, we employed a five-fold cross-validation approach, which involved randomly dividing the entire dataset into five equally sized subsets. In each round of experiments, we designated four of these subsets as the training set and the remaining subset as the test set, with 80% of the samples used for training and 20% for testing. During each round of experiments, 15% of the training samples were set aside as the validation dataset, to monitor the model's performance on this dataset and determine the appropriate hyperparameters and to overfitting. We evaluated the performance of the system based on the average of the five-fold cross validation.

4.3 Parameter initialization

The backbone selection has a notable effect on the performance of the model, hence, we evaluated the performance differences of different backbones. , were updated to learn specific features of BUS images from the training data.

4.4 Data Augmentation

We enhanced the accuracy of our BUS image model by implementing data augmentation techniques, which including zoom (20%), width shift (10%), rotation (5 degrees), shear (20%), and horizontal flip. We did not use up-down flip as it alters the relative position of tissues and could not preserve the morphological features of tissues and positional relationships between organs in the image. Our experiments showed that incorporating these data augmentation methods enhanced the performance of the system.

4.5 Hyperparameter Tuning

The hyperparameters for the training process were chosen through empirical evaluation. A batch size of 6 was used, and the models were optimized with the adaptive moment estimator (Adam) algorithm. The initial learning rate was 10^-5, and if the validation loss did not decrease for 15 epochs, it was reduced to 10^-6. Training was stopped if the validation loss did not decrease for 30 epochs to avoid overfitting. The weight coefficients for the loss terms were assigned values of (0.2, 0.2, 0.2, 0.2, 0.2, 0.1, 0.1, 0.1, 0.1, 0.5, 0.6) to the orientation, shape, margin, echo pattern, posterior features, indistinct margin, angular margin, spiculated margin, microlobulated margin, tumor classification, and the image segmentation branches, respectively, with the highest weight assigned to the segmentation branch and the tumor class branch was assigned the second largest weight. In the loss function presented in the Chapter 3, these weight coefficients were noted as λ_i where i denotes the index of each task.

4.6 Evaluation Metrics

The performance of the classifier is evaluated using accuracy, sensitivity, and specificity. Denoting the number of true positive, true negative, false positive, and false negative as TP, TN, FP, and FN, we can calculate the accuracy sensitivity and specificity.

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{4-1}$$

$$Sensitivity=\frac{TP}{TP+FN} \tag{4-2}$$

$$Specificity=\frac{TN}{TN+FP} \tag{4-3}$$

4.7 Ablation Studies

The results from the experimental validation of MT-BI-RADS are shown in Table 4-3. The results indicate that the network achieved over 80% accuracy for all five BI-RADS descriptors, whereas the tumor class accuracy reached 91.3%. Importantly, the model achieved 94% sensitivity. Due to space limitation, the results for the margin sub-classes are not presented in the table, however, for all 4 sub-classes the accuracy exceeded 80%. The table also presents the results of an ablation study performed to evaluate the impact of the different components in the design of MT-BI-RADS. The ablation study assesses the contributions by the data augmentation, pretrained network parameters, additional image channels with histogram equalization and smoothing, and cropping of the original images to square-size segments. Table 4-3 also presents a comparison between the proposed model with a VGG-16 backbone to ResNet50 and DenseNet121 backbones. The proposed approach furnishes ad-hoc explainability concurrently with the training/testing phases, by offering explanations regarding the tumor class, the segmented tumor region, and the predicted BI-RADS categories. Sharing

the encoder by all branches in MT-BI-RADS ensures that the feature maps used for tumor classification are also used for obtaining the explanations for the BI-RADS descriptors and the segmentation mask.

Table 4-3 Ablation study of the impact of different components in the network on the performance of the classifier

Method	**Tumor Class**			**BI-RADS Descriptors**					**Segmentation**
	Accuracy	Sensitivity	Specificity	Orientation	Shape	Margin	Echo. Pat	Pos. Feat	Dice Score
MT-BI-RADS	**0.913**	**0.940**	0.858	0.845	0.884	0.886	0.806	0.839	0.827
Without Augmentation*	0.892	0.897	0.880	0.850	**0.887**	0.875	0.808	0.835	0.813
Without Pretraining*	0.876	0.895	0.837	0.778	0.841	0.838	0.696	0.718	0.790
Single Channel Images*	0.828	0.917	0.650	0.765	0.823	0.810	0.664	0.689	0.781
Without Image Cropping*	0.790	0.931	0.508	0.734	0.788	0.777	0.646	0.654	0.742
DenseNet Backbone	0.908	0.912	**0.898**	0.850	0.874	0.880	0.815	0.805	0.837
ResNet Backbone	0.903	0.925	0.856	**0.864**	0.882	0.891	**0.814**	0.820	**0.842**
MobileNet Backbone	0.909	0.930	0.869	0.848	0.885	0.807	0.796	**0.867**	0.841
BI-RADS-NET V1	0.900	0.923	0.885	0.848	0.877	**0.887**	**0.814**	0.834	N/A

4.8 Image Segmentation

Image segmentation is a powerful technique in medical image analysis, particularly in cancer tumor detection models, as it can extract specific regions of interest from medical images, such as tumors, organs, or blood vessels. Accurate segmentation can provide detailed and precise information about the location and extent of the tumor within the image, enabling better diagnosis, treatment planning, and monitoring of disease progression. In addition, segmentation can help explain the tumor region in input images by providing a segmentation

mask that highlights the specific region of the image containing the tumor. As stated before, the VGG-16 [75] model had proved to contribute the most to the overall accuracy of network's tumor classification.

VGG-16 is a deep convolutional neural network architecture that was introduced in 2014 by a team of researchers at the University of Oxford. It is named after the Visual Geometry Group, which is the research group where the network was developed [75]. VGG-16 is composed of 16 layers, which include 13 convolutional layers and 3 fully connected layers. The convolutional layers have small 3x3 filters and are stacked on top of each other, which leads to a deep representation of the input image. Fig 4-9 shows the architecture of VGG-16.

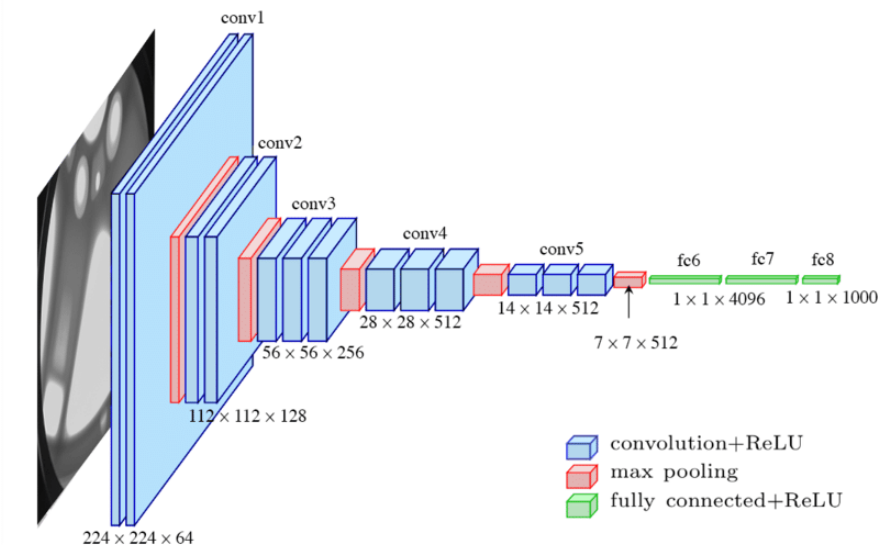


Figure 4-9 VGG-16 architecture [76]

On the other hand, U-Net [77] is considered one of the most effective deep neural network architectures for image segmentation, especially in biomedical applications. It has been shown to achieve state-of-the-art performance on several benchmark datasets and has been widely

adopted in the medical imaging community [77, 78, 79]. Fig 4-10 shows the architecture of U-Net.

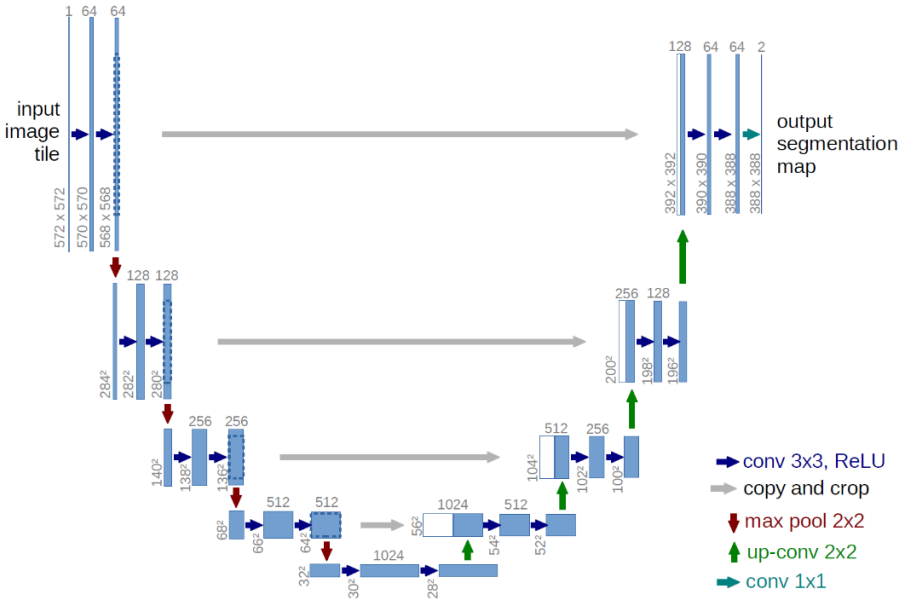
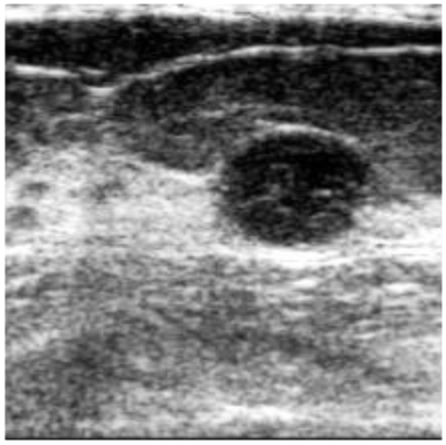


Figure 4-10 U-Net architecture [77]

As it can be inferred from the image, the U-Net architecture consists of an encoder, which progressively down-samples the input image. The down-sampling helps to extract higher-level features and abstract representations of the input image. The decoder in U-Net is responsible for up-sampling the feature maps generated by the encoder back to the original input size while also combining the features from the encoder's corresponding layers. The decoder's role is to reconstruct the original image while preserving the spatial information of the input image.

Our proposed network uses VGG-16 as the backbone Encoder, whereas the second part of the U-Net was used as the Decoder for the image segmentation task. The convolution layers of the VGG-16 were mapped to the up-sampling blocks of U-Net. The combination of VGG-16 and U-Net in medical image analysis offers multiple advantages. VGG-16 can extract high-level features from the input image, which U-Net can use to produce a precise segmentation map

that can identify the tumor's location and size. This approach has demonstrated state-of-the-art results in various medical image analysis tasks, including tumor detection, segmentation, and classification [77, 80]. Additionally, VGG-16's simple architecture allows it to be readily integrated with other neural networks, such as U-Net, to improve their performance. The use of transfer learning, where a pre-trained VGG-16 model is fine-tuned on medical image datasets, has been shown to enhance the accuracy and robustness of U-Net models for tumor segmentation [81]. combining the two architectures provided us with a highly accurate image segmentation branch with a dice score of 82.74%. Fig 4-11 shows examples of two tumor images and their respected segmentation mask.



a) A BUS image

b) The corresponding segmentation mask

c) A BUS image

d) The corresponding segmentation mask

Figure 4-11 Presentation of two input BUS images and the constructed segmentation mask for them

4.9 Quantitative Explanations

As stated before, the radiologist, the medical doctors and the data scientist examining this framework not only need to know the accuracy of the model in differentiating between the malignant and benign tumors, but also they need to be able to verify the predictions and ensure that the basis of the classification is valid. We implemented three different explanations

including image segmentation that reveals predicted tumor regions, probability of the different categories of BI-RADS descriptors and also Shapley Values that quantify the contribution of the BI-RADS descriptors to the tumor classification.

SHAP provides visualization of the calculated Shapley Values in which the descriptors that are pictured with red color explain how they led to benignancy of a sample and blue values explain how they led to the malignancy of sample. The value in front of each descriptor indicates the predicted probability of the category. Since the inputs to the SHAP explainer are one-hot vector outputs of the descriptor classifiers, the model predicts a probability for each of the one-hot vectors. For instance, for orientation feature of a sample the model might predict two values of [0.15, 0.85] which means the probability of being parallel is 15% and the probability of being not-Parallel is 85%.

In this subsection we present the SHAP explanations of several benign and malignant tumor samples. Each figure presents the BUS ultrasound image, ground truth labels of the tumor, ground truth tumor class and the predicted tumor class.

a)

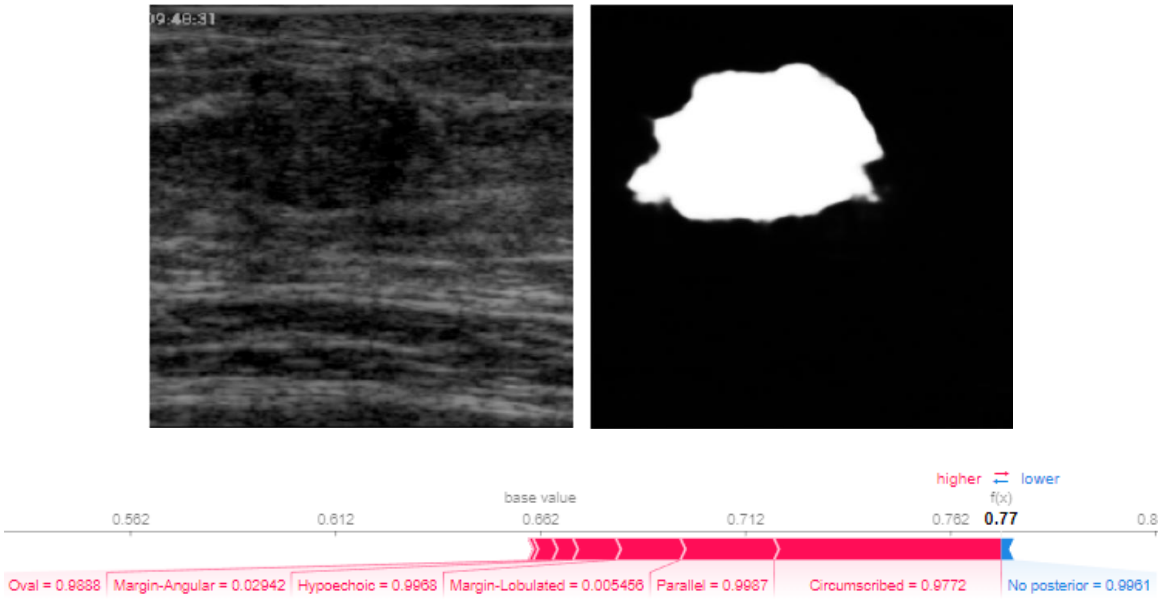


Figure 4-12 Image and the post-hoc explanations of a benign sample with parallel orientation, circumscribed margin, hypoechoic echo pattern and oval shape

Fig 4-12 displays a benign sample. The model predicted that the sample has a non-lobulated Margin pattern with a probability of 99.45%, echo pattern of type Hypoechoic with a probability of 99.58%, is Oval shaped with a probability of 98.88%, and the probability of having lobulated margin is less than 1%. On the other side of the visualization we can see that model predicted that this sample has no posterior feature with a probability of 99.61%. Being in the blue section means that it contributed to the malignancy of the sample, however, in general, having no posterior features does not necessarily contribute to malignancy of a sample. However, since the model is assigning a slight probability to acoustic shadowing of the sample, it is assigning a slight contribution to the malignancy of the sample.

By observing the red labels and their respected values and comparing them with the ground truth label of the sample we can tell that being circumscribed, parallel, micro-lobulated,

hypoechoic, oval shaped and having no posterior features are all correctly predicted by the model. Also, the longer the bars in the SHAP graph, the greater is the contribution to the malignancy or benignancy of a sample. In this case the margin (circumscribed), orientation (parallel) and non-lobulation are the main contributions to the classification decision of the model. Thus, regarding the fact that the labels are predicted correctly and the SHAP explanations comply with the BI-RADS lexicon, it can be concluded that the explanation for this sample is accurate.

b)

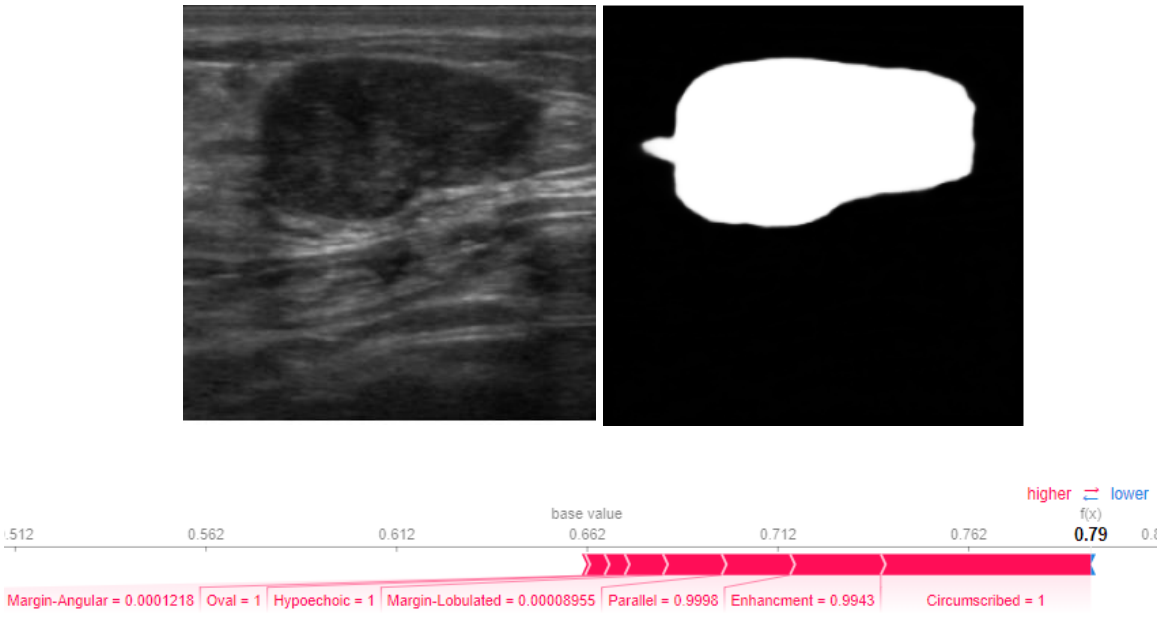


Figure 4-13 Image and the post-hoc explanations of a benign sample with parallel orientation, circumscribed margin, hypoechoic echo pattern, enhancement and oval shape

For the image in Figure 4-13, the ground truth tumor class is benign, and the model also has classified it as benign. By comparing the labels of the BI-RADS descriptors with the ground truth labels, we can see that the model has been able to accurately predict the type of margin, shape, orientation, posterior feature and echo pattern of the sample. According to the

explanation, having all these features and having a close-to-zero probability for angular and lobulated margin all lead to the benignancy of the sample, which complies with the BI-RADS lexicon.

c)

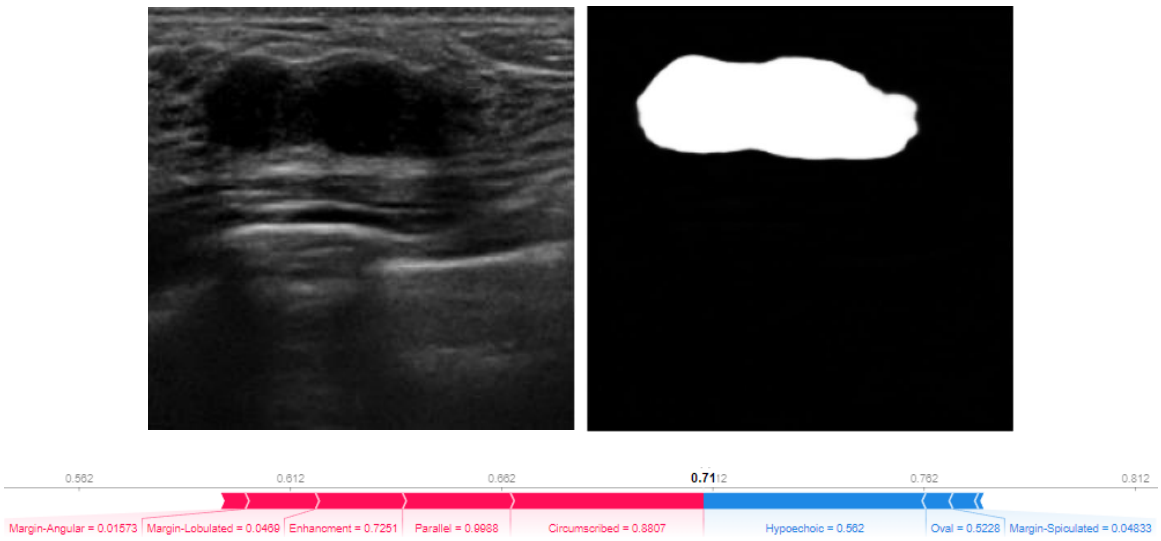


Figure 4-14 Image and the post-hoc explanations of a benign case with a circumscribed margin, being parallel to the skin, enhancement, a hypoechoic echo pattern, spiculation and oval shape, having with both malignancy and benignancy signs

Fig 4-14 displays the output mask of the model for a benign tumor sample. The SHAP explanations reveal that although the attributes of having a hypoechoic echo pattern and low probability of being oval shaped (which we assume is because the model assigns a probability to the irregular shape of the tumor), and low probability of spiculated margin contributed towards the tumor's malignancy (indicated by the blue bar), still the posterior enhancement, being parallel to the skin, and having a circumscribed margin were stronger predictors of benignancy, ultimately leading to being classified as benign.

d)

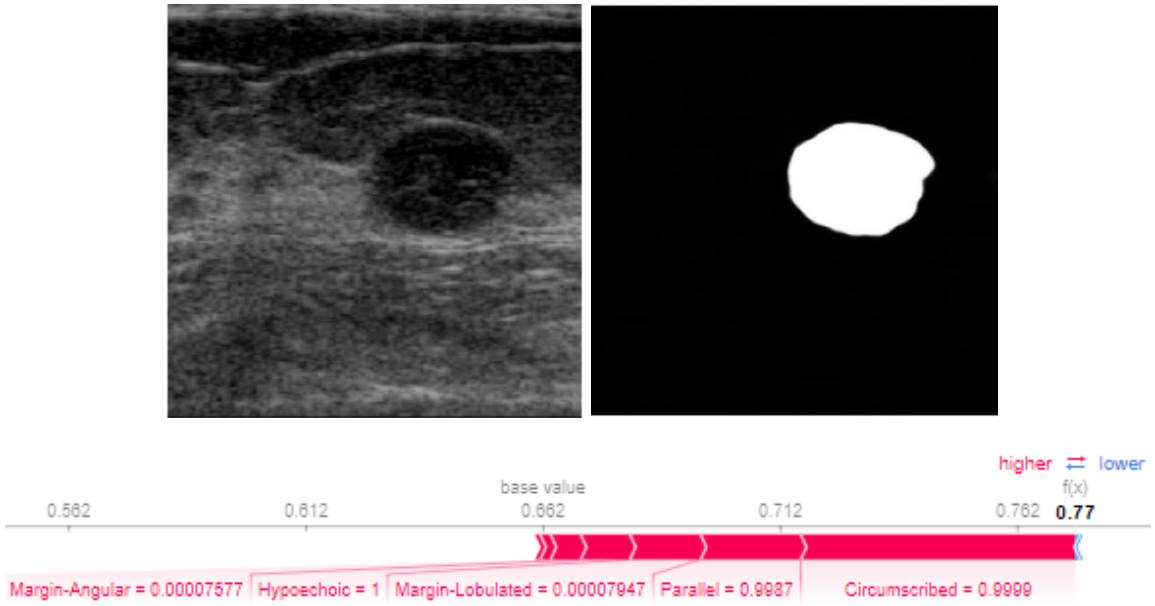


Figure 4-15 A benign case with parallel orientation, circumscribed margin, and hypoechoic echo pattern

Fig 4-15 displays the predicted segmentation mask and the SHAP values for a benign sample. The values indicate that the probability of the sample being parallel to the skin is 99.87%, having a circumscribed margin is 99.99%, having a hypoechoic echo pattern is 100%, as well as there is a fairly low probability of the sample having lobulated and angular margins. These explanations imply that the model correctly interpreted that all these descriptors led to the benignancy of the prediction, as expected.

e)

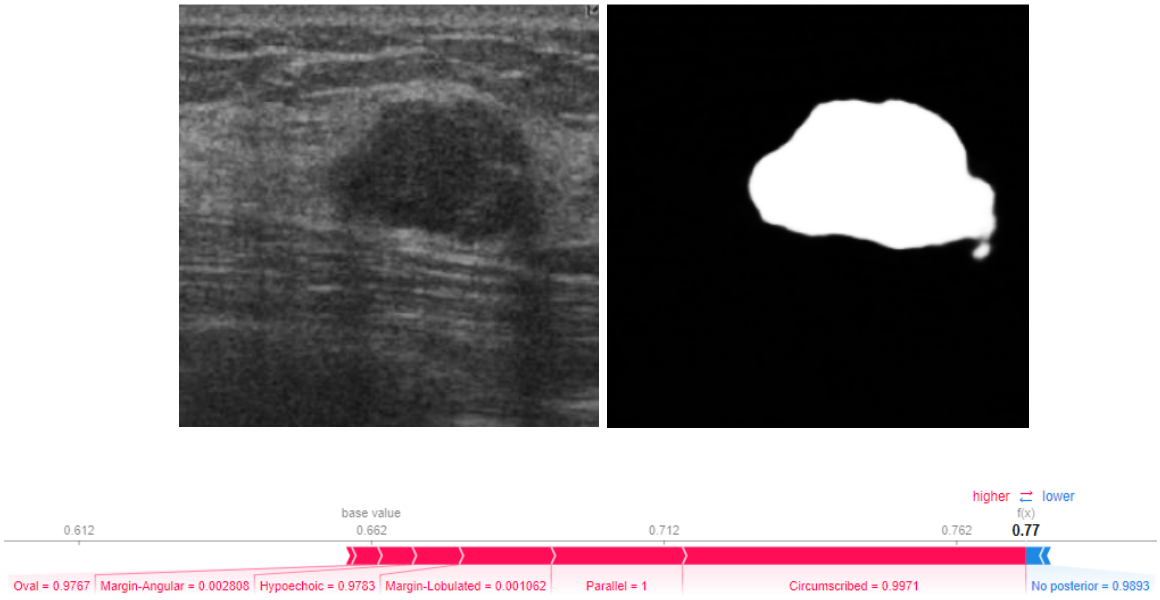


Figure 4-16 A benign sample with a parallel orientation, hypoechoic echo pattern, circumscribed margin, no posterior features shadowing and oval shaped

This sample is identified as benign. By comparing the predicted values we can see that all the descriptors (parallel orientation, hypoechoic echo pattern, circumscribed margin, no posterior features shadowing and oval shaped) have been predicted correctly, and according to BI-RADS lexicon all these contribute to benignancy of a tumor. The reason that 98.9% probability of having no posterior feature is identified as a malignancy factor by SHAP is that in fact the model has identified some very slight indications of acoustic shadowing (1.1%) in the tumor image and thus considers posterior feature as a factor that can slightly lead to malignancy.

f)

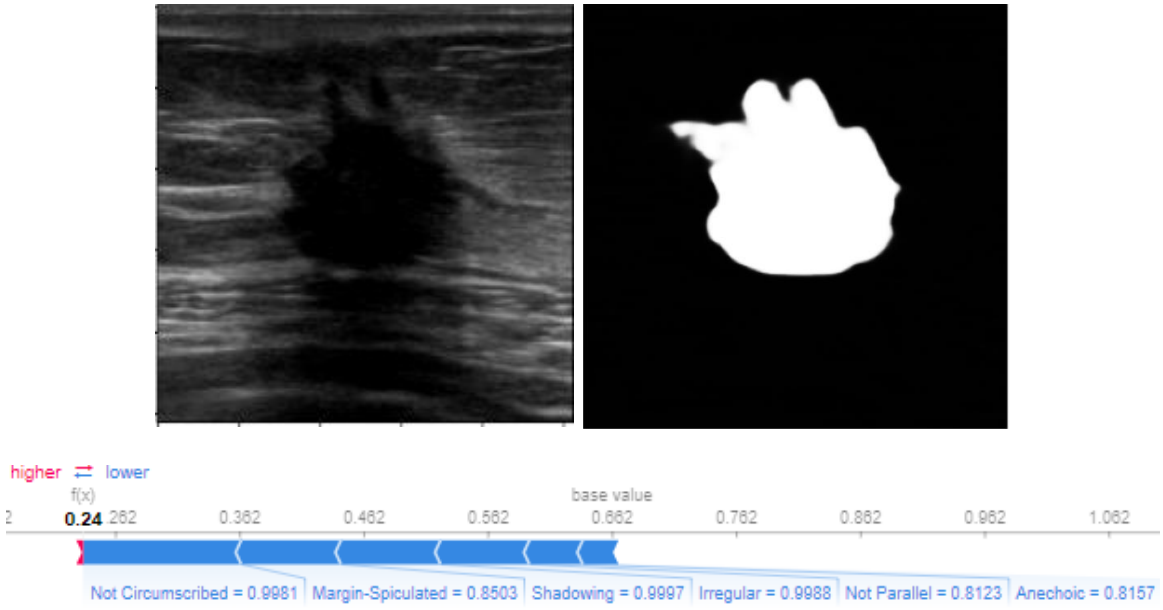


Figure 4-17 A malignant sample with not-parallel orientation, not-circumscribed margin, anechoic echo pattern, spiculated margin and irregular shape

According to both the ground truth and the model's prediction this sample is malignant, and as can be seen the not-circumscribed margin, acoustic shadowing, spiculated margin, irregular shape, not-parallel orientation and anechoic echo pattern have been predicted correctly by the model, all of which, according to the explainer, contribute to malignancy of the sample. According to BI-RADS lexicon all these features are considered contributors to tumor's malignancy and the explainer is justifying the predictions accurately.

g)

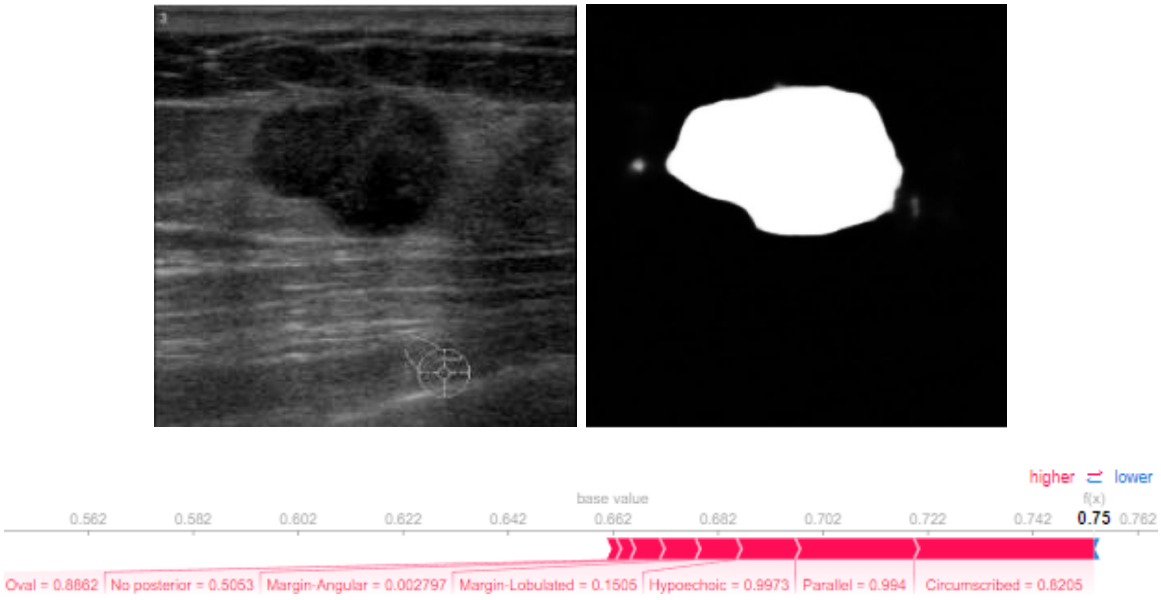


Figure 0-18 A malignant sample with a parallel orientation, hypoechoic echo pattern, circumscribed margin, and oval shape that the model has misclassified as benign

According to the ground truth label this tumor is malignant, however, the model has identified the sample as benign. Considering the model's predictions for the sample having a circumscribed margin, parallel orientation, hypoechoic echo pattern and not having posterior features or angular and micro-lobulated margins and classifying it as an oval shaped sample, all the predicted labels match with the ground truth and also according to both BI-RADS lexicon and the explainer, all of those features are contributors to tumor's benignancy. In other words, all the descriptors are counting towards the sample's benignancy, and provided that information the model's classification can be considered correct. Still, the biopsy has indicated that this sample is malignant, although all the descriptors of the image are benign signs.

h)

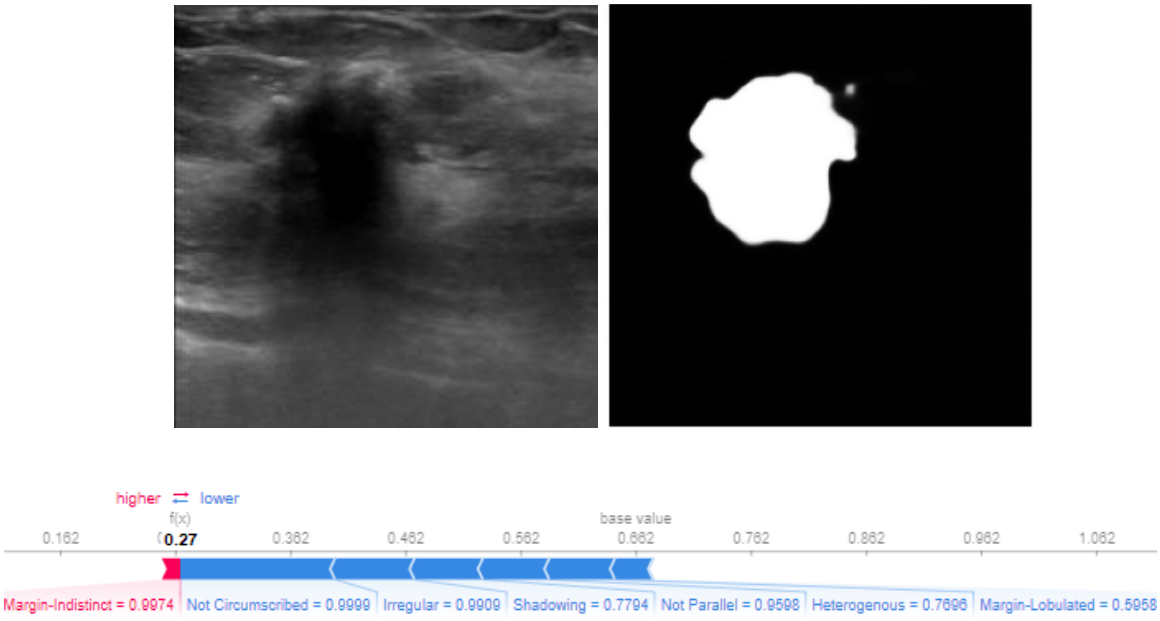


Figure 0-19 A malignant sample with irregular shape, not-circumscribed margin, not parallel orientation, heterogenous echo pattern, and micro-lobulated margin

This sample is malignant and the model has classified it correctly as malignant. All the descriptor labels, such as not circumscribed margin, irregular shape, not-parallel orientation, heterogeneous echo pattern, lobulated margin are predicted correctly. The model has identified shadowing in the image, however, the ground truth assigns no posterior features to this sample. It is believed that this happened because of the similarity of the right side of the image to acoustic shadowing.

j)

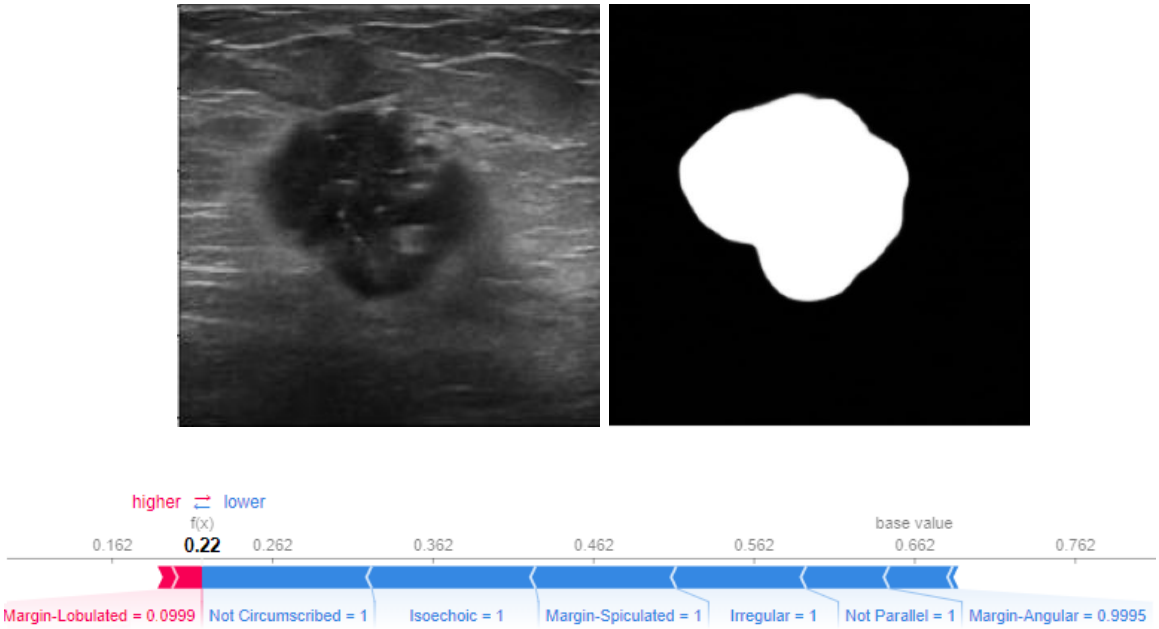


Figure 0-20 A malignant case with a non-circumscribed margin, isoechoic echo pattern, spiculated and angular margin, and irregular shape, and an orientation not parallel to the skin

Fig 4-20 displays the output mask of the model for a malignant case. According to the SHAP explanations, the spiculated margin, angular margin, not being parallel to the skin, not having a circumscribed margin, and having an irregular shape contributed the most to the malignancy. This is expected, as these are well-known malignancy characteristics of breast tumors. Although the isoechoic echo pattern contributed to the prediction, this is not a typical malignancy sign, and requires further analysis.

k)

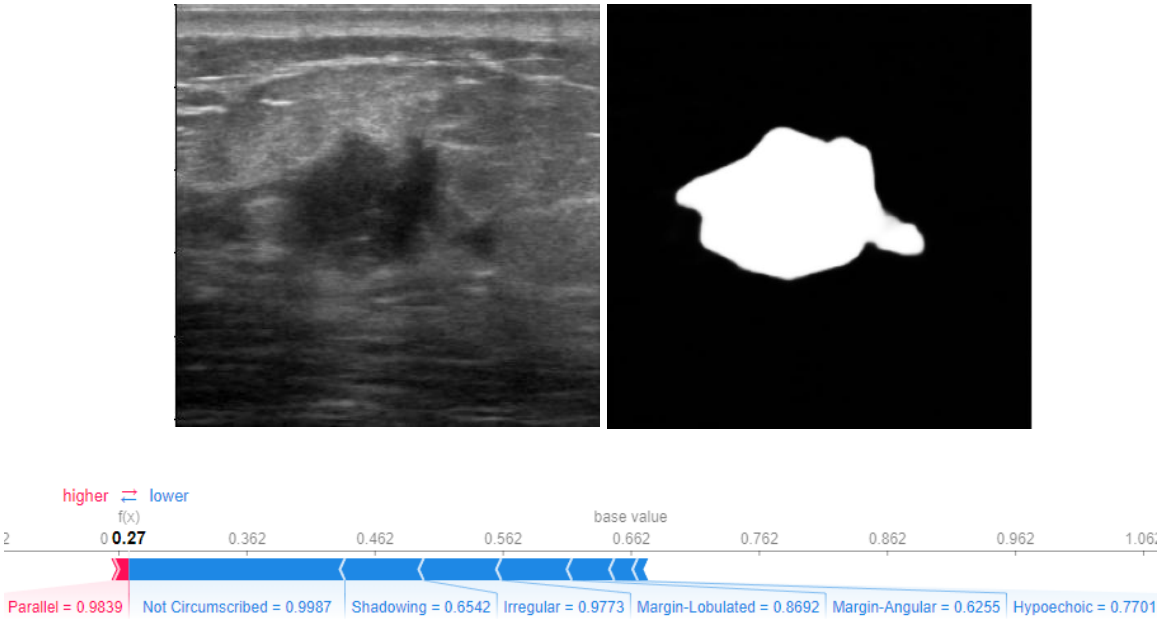


Figure 0-21 A malignant case with a non-circumscribed margin, hypoechoic echo pattern, lobulated and angular margin, an irregular shape, and an orientation parallel to the skin having acoustic shadowing.

Fig 4-21 displays the output mask of the model for a malignant case with a non-circumscribed margin, irregular shape, acoustic shadowing, micro-lobulated margin, angular margin, and hypoechoic echo pattern. SHAP explanations show that these characteristics have significantly contributed to the malignant prediction, whereas being parallel to the skin had small benign contributions. These explanations match well the malignancy characteristics of breast tumors.

CHAPTER 5: CONCLUSION

In conclusion, this thesis proposes an explainable deep learning model for breast cancer detection using Breast Ultrasound Images and BI-RADS descriptors. The model incorporates image segmentation and Shapley Values to provide transparent explanations for its predictions, enabling clinicians to understand how the model arrived at its decision and increasing their confidence in the model's accuracy.

The proposed model achieves high accuracy and interpretability, with high sensitivity and specificity. The use of image segmentation provides transparent visual explanations for the tumor regions in the input images, and the application of Shapley Values as a post-hoc method provides an additional level of explainability without compromising the performance. By quantifying the contribution of each BI-RADS descriptor to the final prediction, we provide transparent and verifiable explanations for the model's decision-making process.

Our proposed model takes advantage of the BI-RADS lexicon and has the potential to improve the accuracy and interpretability of CAD systems for breast cancer detection. The transparent explanations provided by our model can help medical professionals to better understand the diagnostic process and increase their confidence in the use of machine learning models for cancer detection.

References

[1] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, Vignat J, Gralow JR, Cardoso F, Siesling S, Soerjomataram I. Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast. 2022 Dec;66:15-23. doi: 10.1016/j.breast.2022.08.010. Epub 2022 Sep 2. PMID: 36084384; PMCID: PMC9465273.

[2] Giordano, Sharon H. "Breast cancer in men." New England Journal of Medicine 378.24 (2018): 2311-2320.

[3] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, Vignat J, Gralow JR, Cardoso F, Siesling S, Soerjomataram I. Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast. 2022 Dec;66:15-23. doi: 10.1016/j.breast.2022.08.010. Epub 2022 Sep 2. PMID: 36084384; PMCID: PMC9465273

[4] https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf

[5] Davis PL, Staiger MJ, Harris KB, Ganott MA, Klementaviciene J, McCarty KS Jr, Tobon H. Breast cancer measurements with magnetic resonance imaging, ultrasonography, and mammography. Breast Cancer Res Treat. 1996;37(1):1-9. doi: 10.1007/BF01806626. PMID: 8750522.

[6] K. Priya, V. Senthilkumar, J. Samson Isaac, S. Kottu, V. S. Ramakrishna and M. Jogendra Kumar, "Breast Cancer Segmentation by K-Means and Classification by Machine Learning," 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2022, pp. 651-656, doi: 10.1109/ICACRS55517.2022.10029301.

[7] https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads

[8] Huang, Q., Luo, Y. & Zhang, Q. Breast ultrasound image segmentation: a survey. Int J CARS 12, 493–507 (2017). https://doi.org/10.1007/s11548-016-1513-1.

[9] Michael E, Ma H, Li H, Kulwa F, Li J. Breast Cancer Segmentation Methods: Current Status and Future Potentials. Biomed Res Int. 2021 Jul 20;2021:9962109. doi: 10.1155/2021/9962109. PMID: 34337066; PMCID: PMC8321730.

[10] Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized Medical Imaging and Graphics, 31(4-5), 198-211.

[11] Sutton GC. Computer-aided diagnosis: a review. Br J Surg. 1989 Jan;76(1):82-5. doi: 10.1002/bjs.1800760126. PMID: 2645015.

[12] Debelee, T.G., Schwenker, F., Ibenthal, A. et al. Survey of deep learning in breast cancer image analysis. Evolving Systems 11, 143–163 (2020). https://doi.org/10.1007/s12530-019-09297-2.

[13] Giger ML. Machine Learning in Medical Imaging. J Am Coll Radiol. 2018 Mar;15(3 Pt B):512-520. doi: 10.1016/j.jacr.2017.12.028. Epub 2018 Feb 2. PMID: 29398494.

[14] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017 Dec;42:60-88. doi: 10.1016/j.media.2017.07.005. Epub 2017 Jul 26. PMID: 28778026.

[15] Aniek F. Markus, Jan A. Kors, Peter R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies, Journal of Biomedical

Informatics, Volume 113, 2021, 103655, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103655.

[16] Waddah Saeed, Christian Omlin, Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowledge-Based Systems, Volume 263, 2023, 110273, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2023.110273.

[17] Guidotti, Riccardo & Monreale, Anna & Turini, Franco & Pedreschi, Dino & Giannotti, Fosca. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys. 51. 10.1145/3236009.

[18] Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in explainable AI. arXiv preprint arXiv:1810.00184.

[19] Sehgal CM, Weinstein SP, Arger PH, Conant EF. A review of breast ultrasound. J Mammary Gland Biol Neoplasia. 2006 Apr;11(2):113-23. doi: 10.1007/s10911-006-9018-0. PMID: 17082996.

[20] W. -X. Liao et al., "Automatic Identification of Breast Ultrasound Image Based on Supervised Block-Based Region Segmentation Algorithm and Features Combination Migration Deep Learning Model," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 4, pp. 984-993, April 2020, doi: 10.1109/JBHI.2019.2960821.

[21] Almajalid, R., Shan, J., Du, Y., & Zhang, M. (2018). Development of a deep-learning-based method for breast ultrasound image segmentation. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1103-1108). IEEE.

[22] Kalafi, E. Y., Jodeiri, A., Setarehdan, S. K., Lin, N. W., Rahmat, K., Taib, N. A., ... & Dhillon, S. K. (2021). Classification of breast cancer lesions in ultrasound images by

using attention layer and loss ensemble in deep convolutional neural networks. Diagnostics, 11(10), 1859. https://doi.org/10.3390/diagnostics11101859.

[23] Tanaka, H., Chiu, S. W., Watanabe, T., Kaoku, S., & Yamaguchi, T. (2019). Computer-aided diagnosis system for breast ultrasound images using deep learning. Physics in Medicine & Biology, 64(23), 235013. https://doi.org/10.1088/1361-6560/ab5eb5.

[24] Zhang, Z., Li, Y., Wu, W., Chen, H., Cheng, L., & Wang, S. (2021). Tumor detection using deep learning method in automated breast ultrasound. Biomedical Signal Processing and Control, 68, 102677. https://doi.org/10.1016/j.bspc.2020.102677.

[25] Vigil, N., Barry, M., Amini, A., Akhloufi, M., Maldague, X. P. V., Ma, L., Ren, L., & Yousefi, B. (2022). Dual-intended deep learning model for breast cancer diagnosis in ultrasound imaging. Cancers, 14(11), 2663. https://doi.org/10.3390/cancers14112663

[26] X. Qi, F. Yi, L. Zhang, Y. Chen, Y. Pi, Y. Chen, J. Guo, J. Wang, Q. Guo, J. Li, et al., "Computer-aided diagnosis of breast cancer in ultrasonography images by deep learning," Neurocomputing, vol. 472, pp. 152–165, 2022.

[27] Chen G, Dai Y, Zhang J. C-Net: Cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation. Comput Methods Programs Biomed. 2022 Oct;225:107086. doi: 10.1016/j.cmpb.2022.107086. Epub 2022 Aug 24. PMID: 36044802.

[28] F. K. Došilović, M. Brčić and N. Hlupić, "Explainable artificial intelligence: A survey," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2018, pp. 0210-0215, doi: 10.23919/MIPRO.2018.8400040.

[29] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539

[30] Slack, Dylan, et al. "Reliable post hoc explanations: Modeling uncertainty in explainability." Advances in neural information processing systems 34 (2021): 9391-9404.

[31] Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI-Explainable artificial intelligence. Sci Robot. 2019 Dec 18;4(37):eaay7120. doi: 10.1126/scirobotics.aay7120. PMID: 33137719.

[32] Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160.

[33] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " 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 (pp. 1135-1144).

[34] Zeiler MD, Taylor GW, Fergus R. Adaptive Deconvolutional Networks for Mid and High Level Feature Learning. In: IEEE International Conference on Computer Vision, ICCV 2011. Barcelona, Spain; 2011, 2018-2025.

[35] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: IEEE International Conference on Computer Vision, ICCV 2017. Venice, Italy; 2017, 618-626.

[36] Olah C, Mordvintsev A, Schubert L. Feature Visualization. Distill 2017;2:e7. doi:10.23915/distill.00007.

[37] Lundberg, S. M. & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (ed.), Advances in Neural Information Processing Systems 30 (pp. 4765--4774) Curran Associates, Inc.

[38] Henelius A, Puolamäki K, Ukkonen A. Interpreting Classifiers through Attribute Interactions in Datasets. CoRR 2017;abs/1707.07576:

[39] Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017. Sydney, Australia; 2017, 3319-3328.

[40] Smilkov D, Thorat N, Kim B, Viégas FB, Wattenberg M. SmoothGrad: removing noise by adding noise. CoRR 2017;abs/1706.03825:

[41] Ribeiro MT, Singh S, Guestrin C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIKDD International Conference on Knowledge Discovery and Data Mining. New York, USA; 2016, 1135-1144.

[42] Tan S, Caruana R, Hooker G, Lou Y. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, USA; 2018, 303-310.

[43] Bastani O, Kim C, Bastani H. Interpretability via Model Extraction. CoRR 2017;abs/1706.09773:

[44] Van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022 Jul;79:102470. doi: 10.1016/j.media.2022.102470. Epub 2022 May 4. PMID: 35576821.

[45] Zhou, B., Khosla, A., Lapedriza, A ., Oliva, A ., Torralba, A ., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929.

[46] Selvaraju, R.R., Cogswell, M., Das, A. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis 128, 336–359 (2020). https://doi.org/10.1007/s11263-019-01228-7

[47] Springenberg, J.T., Dosovitskiy, A., Brox, T., & Riedmiller, M.A. (2014). Striving for Simplicity: The All Convolutional Net. CoRR, abs/1412.6806.

[48] Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.

[49] Wang, C.J.; Hamm, C.A.; Savic, L.J.; Ferrante, M.; Schobert, I.; Schlachter, T.; Lin, M.; Weinreb, J.C.; Duncan, J.S.; Chapiro, J.; et al. Deep learning for liver tumor diagnosis part II: Convolutional neural network explanation using radiologic imaging features. Eur. Radiol. 2019, 29, 3348–3357.

[50] Wang, H.; Feng, J.; Bu, Q.; Liu, F.; Zhang, M.; Ren, Y.; Lv, Y. Breast Mass Detection in Digital Mammogram Based on Gestalt Psychology. J. Healthc. Eng. 2018, 2018, 4015613. [CrossRef]

[51] Ha, R.; Chin, C.; Karcich, J.; Liu, M.Z.; Chang, P.; Mutasa, S.; Van Sant, E.P.;Wynn, R.T.; Connolly, E.; Jambawalikar, S. Prior to Initiation of Chemotherapy, Can We Predict Breast Tumor Response? Deep Learning Convolutional Neural Networks Approach Using a Breast MRI Tumor Dataset. J. Digit Imaging 2019, 32, 693–701. [CrossRef]

[52] Yang, X.; Wang, Z.; Liu, C.; Le, H.M.; Chen, J.; Cheng, K.T.T.; Wang, L. Joint Detection and Diagnosis of Prostate Cancer in Multi- Parametric MRI Based on Multimodal Convolutional Neural Networks. In Proceedings of the Medical Image Computing and Computer Assisted Intervention, 20th International Conference, Quebec City, QC, Canada, 11–13 September 2017; pp. 426–434.

[53] L. Liu, F. Yang, Y. Wei, Y. Wang, H. Qin, and Q. Yang, "A survey of deep neural network architectures and their applications," Neurocomputing, vol. 234, pp. 11-26, 2017.

[54] Shia W-C, Chen D-R. Classification of malignant tumors in breast ultrasound using a pretrained deep, residual network model and support vector machine. Compute Medical Imaging Graph 2021;87:101829. https://doi.org/10.1016/j.compmedimag.2020.101829.

[55] Xie J, Song X, Zhang W, Dong Q, Wang Y, Li F, et al. A novel approach with dual-sampling convolutional, neural network for ultrasound image classification of breast tumors. Phys Med Biol 2020. https://doi.org/10.1088/1361-6560/abc5c7.

[56] Zhuang Z, Yang Z, Zhuang S, Raj ANJ, Yuan Y, Nersisson R. Multi-Features-Based Automated Breast , Tumor Diagnosis Using Ultrasound Image and Support Vector Machine. Comput Intell Neuroscience, 2021;2021:9980326:1-9980326:12. https://doi.org/10.1155/2021/9980326.

[57] Zhang G, Zhao K, Hong Y, Qiu X, Zhang K, Wei B. SHA-MTL: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification. Int J Comput Assist Radiology, Surg 2021. https://doi.org/10.1007/s11548-021-02445-7.

[58] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

[59] Kim, J., Lee, J., Kang, W., & Kim, W. (2021). XAI-Net: Explainable artificial intelligence network for breast cancer diagnosis. IEEE Transactions on Medical Imaging, 40(1), 256-266.

[60] Yang, Y., Sun, X., Zhang, Y., Wu, S., Liu, Z., & Xu, B. (2021). SNet: A segmentation network for breast cancer diagnosis using mammography images. Medical Image Analysis, 68, 101905.

[61] Y. Shen, Y. Jiang, J. Sun, Z. Li, and Y. Zhang, "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," Medical Image Anal., vol. 68, p. 101908, 2021.

[62] J. Wu, C. Zhang, T. Xu, and L. Zhang, "DeepMiner: Discovering Interpretable Representations for Mammogram Classification and Explanation," ArXiv, vol. abs/1805.12323, 2018.

[63] S. T. Kim, H. Lee, H. G. Kim, and Y. M. Ro, "ICADx: interpretable computer aided diagnosis of breast masses," in Medical Imaging: CAD, Houston, Texas, USA, 2018, p. 1057522.

[64] Zhang E, Seiler S, Chen M, Lu W, Gu X. BIRADS features-oriented semi-supervised deep learning for breast ultrasound computer-aided diagnosis. Phys Med Biol. 2020 Jun 12;65(12):125005. doi: 10.1088/1361-6560/ab7e7d. PMID: 32155605.

[65] A. Rezazadeh, Y. Jafarian, and A. Kord, "Explainable Ensemble Machine Learning for Breast Cancer Diagnosis Based on Ultrasound Image Texture Features," Forecasting, vol. 4, no. 1, pp. 262-274, Feb. 2022.

[66] H. Lee, S. T. Kim, and Y. M. Ro, "Generation of Multimodal Justification Using Visual Word Constraint Model for Explainable Computer-Aided Diagnosis," in Interpretability of Machine Intelligence in MIC and MLCDS, 2019, pp. 21-29.

[67] Zhang, B., Vakanski, A., & Xian, M. (2021). Bi-Rads-Net: An Explainable Multitask Learning Approach for Cancer Diagnosis in Breast Ultrasound Images. In 2021 IEEE 31st International Workshop on MLSP, 1-6.

[68] Zhang B, Vakanski A, Xian M. BI-RADS-NET-V2: A Composite Multi-Task Neural Network for Computer-Aided Diagnosis of Breast Cancer in Ultrasound Images With Semantic and Quantitative Explanations. IEEE Access. 2023;11:79480-79494. doi: 10.1109/access.2023.3298569. Epub 2023 Jul 25. PMID: 37608804; PMCID: PMC10443928.

[69] American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS). 5th ed. Reston, Va: American College of Radiology; 2013

[70] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” Data in Brief, vol. 28, p. 104863, Feb. 2020.

[71] M. Xian et al., “A Benchmark for Breast Ultrasound Image Segmentation (BUSIS),” CoRR, vol. abs/1801.03182, 2018.

[72] T. Geertsma, Ultrasoundcases.info, FujiFilm. (2014). https://www.ultrasoundcases.info/.

[73] Shapley, L. S. (1953). A Value for n-Person Games. In H. W. Kuhn & A. W. Tucker (Eds.), Contributions to the Theory of Games (Vol. 2, pp. 307-317). Princeton University Press.

[74] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[75] Ferguson, Max & ak, Ronay & Lee, Yung-Tsun & Law, Kincho. (2017). Automatic localization of casting defects with convolutional neural networks. 1726-1735. 10.1109/BigData.2017.8258115.

[76] U-Net: Convolutional Networks for Biomedical Image Segmentation by Olaf Ronneberger, Philipp Fischer, and Thomas Brox

[77] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 424-432). Springer, Cham.

[78] Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., ... & Maier-Hein, K. H. (2018). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 16(4), 319-326.

[79] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2018). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2097-2106

[80] Zhou, T., & Tajbakhsh, N. (2019). UNet++: A nested U-Net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (pp. 3-11). Springer