Norwegian University
of Life Sciences

Master's Thesis 2024 30 ECTS
Faculty of Science and Technology

Exploring Breast Cancer Diagnosis: A Study of SHAP and LIME in XAI-Driven Medical Imaging

Ulrik Egge Husby
Data Science

Acknowledgements

I would like to deeply thank my supervisor Associate Professor in Data Science Fadi Al Machot at NMBU, for introducing me to this topic, and guiding me throughout my thesis. I appreciate the guidance, professional help and valuable insight provided throughout the thesis. I also want to acknowledge and thank my fellow students who spent the time together with me, making the writing of this thesis an enjoyable experience. A special thanks to my good friends Torjus Strandenes Moen and Kim Næss Kynningsrud for providing great feedback and support while writing my thesis, and for taking time out their day to read through my thesis.

Abstract

The motivation for this thesis is to enhance the interpretability and explainability of using Artificial Intelligence (AI) in healthcare, focusing on breast cancer images. Breast cancer is one of the leading causes of cancer-related deaths among women, making early detection and accurate diagnosis important to reduce mortality. To increase interpretability, explainability and trust of AI in healthcare, two Explainable AI (XAI) techniques SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are explored, and used to explain the underlying model EfficientNetV2B2. By employing metrics such as Intersect over Union (IoU), Precision, Recall and F1-score, this thesis evaluated the performance of these techniques in accurately identifying and localizing tumor regions in breast cancer images.

Through methodological insights, this thesis highlight that both SHAP and LIME enhanced the transparancy and interpretability of AI models, which is a crucial requirement in healthcare. They allowed for a detailed breakdown of decisions made by the underlying model by highlighting important features in images, contributing to a deeper understanding and trust in AI decisions. However, both techniques faced challenges such as computational complexity and inconsistency in performance, which limited their practical application.

The results indicated that SHAP generally provided higher precision than LIME, suggesting its useability in applications where reducing false positives is critical, which again could be useful in early diagnosis when capturing all positives is important. On the other hand LIME provided higher recall than SHAP, which could be essential in scenarios where reducing false negatives is vital. Reducing false negatives is essential in medical diagnosis since this can have fatal consequences for patients if a region is classified as non-cancerous while in reality it is cancerous.

The thesis underscores the potential of XAI to improve the interpretability and trust in AI models, especially in healthcare, as well as aiding in early diagnosis, which can result in higher survival rates when assessing breast cancer. Despite the variability in the techniques' performance, the ability of SHAP and LIME to provide visual and intuitive insights into model decisions marks a significant step towards integrating XAI techniques in critical healthcare applications. This study contributes to the ongoing focus on the need for trustable and interpretable AI models, suggesting areas for further research and development in XAI techniques.

Contents

List of Abbreviations

AI Artificial Intelligence

CNN Convolutional Neural Network

CT Computed Tomography

Grad-CAM Gradient-weighted Class Activation Mapping

IoU Intersection over Union

LIME Local Interpretable Model-agnostic Explanations

MRI Magnetic Resonance Imaging

PET Positron Emission Tomography

ReLU Rectified Linear Unit

SHAP SHapley Additive exPlanations

XAI Explainable AI

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In the complex landscape of healthcare, where data and diagnoses have uncertainties, the consequences of incorrect decisions can lead to detrimental outcomes for patients. Factors such as variability in patients' responses to treatments, varying symptoms and the complexity of medical data further complicate decision-making processes. AI has emerged in medical imaging, and has become an important tool offering the potential to analyze big datasets and assist in more accurate decisions, however the black-box nature of many AI models can complicate the logic behind outputs [1]. This is where XAI becomes important. XAI provides a method to explain AI decision-making, furthermore aiming to ensure transparancy, and increase efficiency, interpretability and explainability [2].

In the field of medical diagnosis, achieving accurate and early detection of breast cancer is critical for effective treatment and improved patient outcomes. Despite advancements in machine learning and their use in medical imaging tasks, the models often appear as black-box models, offering little insight into how predictions are made for patients. The lack of transparancy, explainability and interpretability can present obstacles when applying these methods in the medical domain [3]. Therefore, this thesis aims to provide insight into model decisions by leveraging XAI techniques, namely SHAP and LIME, to further enhance the black-box nature of models made for prediction, as well as aiming to contribute to more trust in AI systems in the medical domain.

1.1.1 Breast Cancer

Cancer is a huge threat to public health worldwide, being one of the leading causes of death in many countries [4]. In 2024 in the United States, there was an estimated number of 2,001,140 new cancer cases in 2024. The estimated number of deaths from cancer was 611,720. [5].

Breast cancer refers to cancerous nodules that origin from breast tissue. Breast cancer covers 11.7% of all cancer incidences, making it one of the single highest reasons for cancer mortality worldwide, and is the single leading cause of death among women between the

ages of 40 and 55 worldwide [6] [7].

Several risk factors are associated with the development of breast cancer, especially among women. Some risk factors are sex, age, medical history in family, estrogen levels and life style, which are considered as important factors when determining the risk of development of breast cancer [8].

In some developed countries, the five-year survival rate for patients with breast cancer are above 80%, due to advances in early detection and preventative measures. The rate refers to percentage of people alive after five years after the diagnosis. This underscores the critical importance of early diagnosis in enhancing survival outcomes [8].

1.1.2 Ultrasound

To detect and combat breast cancer, various screening and imaging techniques are used. One of the techniques used is ultrasound [9], which is a crucial tool that complements mammography. While mammography is known for its effectiveness in detecting early-stage breast cancer, ultrasound plays an important role, especially aiding in scenarios that need a more comprehensive approach [9].

The role of ultrasound in medical imaging is that it evaluates tumor characteristics like shape, margins and consistency by using sound waves to create images of tissues inside breasts. Unlike other methods like X-rays, ultrasound does not use radiation, making it a safer method, without excessive side effects [10].

1.2 Research Questions

For this thesis the following research questions are to be answered.

RQ1 How effective are XAI techniques like SHAP and LIME in predicting outcomes using breast cancer data?

RQ2 How does XAI techniques like SHAP and LIME contribute to the prediction performance of early diagnosis for breast cancer?

The first research question (RQ1) aims to explore the effectiveness of applying XAI in medical imaging with breast cancer data. The core of this question is to assess accuracy, reliability and usability of the output given by SHAP and LIME, which are two distinct XAI methods used to explain AI model's decision-making processes. By focusing on breast cancer data, the question impacts a critical area of healthcare.

Furthermore, the second research question (RQ2) focuses on the specific contribution of XAI for improving performance of early diagnosis for breast cancer patients. This involves not only looking at accuracy of predictions, but also interpreting the models, especially the output provided from the XAI techniques, which can further enhance decision-making, aiming to lead to earlier detection of breast cancer. This could contribute to more efficient and successful treatment.

1.3 My Objectives

This thesis aims to systematically explore the impact and utility of XAI models within the healthcare domain, focusing on outcome predictions and early diagnosis of breast cancer. The thesis will conduct a review of current XAI models, SHAP and LIME, especially highlighting their methodologies, strengths, and limitations. Furthermore, to evaluate the effectiveness, strengths and limitations, several predefined metrics are leveraged to achieve numerical results to compare and discuss. Further evaluation and investigation of the role of XAI, especially in enhancing model prediction performance is needed, assessing their ability to provide insights to healthcare professionals. Several challenges and limitations are introduced in later chapters, where identification and discussion of these are included, focusing on model transparency, interpretability and explainability. This thesis also aims to explore potential improvements and innovative approaches in XAI, that could enhance its application in early diagnosis and treatment planning. Finally this thesis aim to provide guidelines and best practices for integrating XAI models into healthcare settings, to improve patient outcomes through more accurate and interpretable predictions.

By addressing these objectives, this thesis aims to contribute to the field of healthcare by enhancing the understanding and application of XAI models for better model decision-making and patient care.

1.4 Scope and Limitations

While aiming to provide comprehensive insights into the role of XAI in healthcare, this research is bounded by certain scope and limitations, which are that the research will primarily analyze existing XAI models namely SHAP and LIME that are publicly documented and recognized within the AI community. The evaluation of XAI models will be based on their application to breast cancer imaging data. When working with breast cancer data, a domain where early diagnosis is critical to treatment success, indicating that the focus of early diagnosis is important. Due to the magnitude of the field, the research will not cover all existing XAI models, but will select the representative models SHAP and LIME, which are based on their relevance to the problem at hand which is healthcare and images. The availability of breast cancer datasets that are both comprehensive and publicly accessible is limited due to privacy concerns and regulations. This may restrict the depth of empirical analysis possible. The thesis' conclusions will be examined upon the quality and diversity of the dataset used, which might not fully capture the complexity of real-world scenarios. To further evaluate the XAI techniques a longer scope of study and collaboration with healthcare professionals is needed, which might be beyond the scope of this thesis. The rapidly evolving nature of XAI and machine learning technologies means that some aspects of the research may become outdated. Therefore continuous updates is needed to maintain relevance. Despite this scope and limitations, this thesis aims to provide insights into the potential of XAI in enhancing healthcare outcomes through more accurate predictions and diagnoses. By acknowledging these boundaries, the thesis seeks to contribute to the field

of the integration of AI in healthcare.

Chapter 2

Literature Review

In this chapter, a literature review of existing techniques used in various fields are reviewed, exploring their metrics, results and discussion of the techniques. This chapter aims to introduce existing research in the field of XAI.

2.1 Model-Agnostic Explanation Methods

Various approaches to XAI have been explored in recent literature. The study in [11] presents a comprehensive methodology to evaluate the explainability of AI models using the Kernel SHAP approach, demonstrating its effectiveness in approximating Shapley values for individual predictions across a multitude of features. The method deconstructs a prediction into an additive feature attribution model, providing explanations for model predictions that are both interpretable and reflective of the model's decision-making process. Furthermore, the authors acknowledge potential issues with feature dependence, where traditional Shapley value computations may include non-representative data instances, resulting in false explanations. They enlighten the need for future research to refine the estimation of model output expectations, particularly in scenarios where feature dependence is present [11].

The study [12] utilizes the Kernel SHAP method to optimize a network anomaly detection model, showcasing a feature selection technique that leverages SHAP values. Kernel SHAP shows strength when working with unsupervised learning models, where class labels are absent. Using the CICIDS2017 dataset, the demonstration of Kernel SHAP's ability to quantify feature importance was shown. Through application of Kernel SHAP, the researchers were able to construct an optimized model with accuracy and F1-score of 0.90 and 0.76 respectively. The use of Kernel SHAP offers a robust framework for interpretability that can augment the explanation, and further strengthen the trustworthiness of AI systems, especially in critical applications.

However it is made clear that the Kernel SHAP technique is not devoid of challenges. The method's computational complexity stands out as a limitation, with the potential to require hours of processing time for high-dimensional datasets. This imposes practical

constraints on the method's applicability in real-time or large-scale scenarios. Additionally, the selection of an appropriate background set is crucial for the accurate computation of SHAP values, yet selecting such a set from comprehensive datasets remains a complex task that can significantly impact the results [12].

Furthermore, the study [13] uses Kernel SHAP to enhance the interpretability of implemented models, which was trained on information from ultrasound images and numbers from 952 breast cancer lesions. The study implemented several models where they found that the XGBoost model performed the best of all models tested, across a variety of metrics such as Accuracy, Sensitivity, Specificity and F1-score. The XGBoost model achieved an Accuracy of 0.846, Sensitivity of 0.870, Specificity of 0.862 and F1-score of 0.826 [13]. By leveraging SHAP they found specific features (ultrasound signs) that showed the greatest impact on the model's decision-making process, aiming to strengthen the trust for clinicians.

2.2 Ontologies and Decision Trees

The study conducted in [14] discovers the possibility to enhance the human understandability of post-hoc explanations by using ontologies, this is performed by using the form of decisison trees.

Ontologies In the context of AI, the paper [14] refers to an ontology like a detailed map of a particular subject area. It describes things that exist in the domain and how they are related to each other.

In the paper [14], the authors use a technique called `TREPAN_RELOADED`, which is an approach that extracts surrogate decision trees from black-box models.

Surrogate Decision Trees is a tool that acts as a stand-in for a more complex AI model, in the paper's case the black-box [14]. Since the black-box is difficult to understand, the essence of utilizing ontologies to create these decision trees is to capture the steps and decisions of the black-box model, making it more transparent, accessible and understandable, especially for humans.

The study experiment with human subjects, which demonstrated that decision trees augmented by ontologies were not only more understandable compared to their neural network counterparts, but also did not sacrifice the accuracy of the original models. Subjects interacting with these trees provided more accurate responses, quicker, and reported higher confidence in their understanding of AI decisions [14].

2.3 Local Surrogate Model and Model Distillation

Similar to Kernel SHAP, LIME offers another method for making complex models more understandable. LIME provides clear insights into why models make certain decisions, one

instance at a time. It simplifies the model's decision-making process by creating a model that is easier to understand, nearby the data point in question [15] [16]. Although this method has its complications, it helps to clarify how different features impact individual predictions. This clarity is important for making high-stakes decisions more transparent when using AI. LIME's usefulness has been shown across different machine learning algorithms, highlighting its adaptability and its role in building trust in AI by making its decisions less of a mystery. The paper [15] also explores how LIME works in practice, looking at its effectiveness from the perspective of both experienced, and new users. It points out LIME's advantage in providing detailed local explanations and recognizes the issues with understanding the overall model and the amount of time it takes to apply LIME. The final assessment of LIME shows its important place in the area of XAI, taking into account how user-friendly it is, and the importance of focusing on the user when such tools are put into use [15].

Using these techniques in real-life applications can be crucial, especially in domains like healthcare. In the following research paper by [17], a proposed method that combines transfer learning with the XAI technique LIME to enhance interpretability and transparancy of AI systems in healthcare is discussed. The model used is a CNN-based model which was pretrained, VGG16, to classify chronic wounds. The researchers found that the said model achieved a Precision of 95%, recall of 94% and F1-score of 94% [17].

Similarly, another study was conducted by [18] where detection of brain tumor was the target. A VGG16 model combined with the XAI technique LIME was utilized for the task. The model achieved an accuracy of 97%.

Furthermore, a paper by [19], uses LIME to segment out important features of lungs of COVID-19 patients, aiming to provide insight to researchers when applying XAI in medical imaging. The underlying model's are VGG16 and MobileNetV2. The VGG16 achieved an accuracy of 98.5% when trained on a dataset containing X-ray scans of 400 images. Furthermore, the MobileNetV2 model received an accuracy of 98.5% when trained on the same X-ray dataset. The authors also tested the MobileNetV2 against a CT-scan dataset with 400 images and also a dataset of Mixed-data containing 2591 images, achieving an accuracy of 94% and 95% respectively.

In the paper [20], LightBTSeg is introduced, which is a knowledge distillation approach, essentially meaning that the original model (often called the "teacher") transfer knowledge to another simpler model (often called "student"). The purpose of this is to reduce computational cost of the teacher model while still possessing the learned information that the teacher model has learned. This approach is used to reduce the complexity of the original model, making it more transparent and less complex. They use this approach to segment benign and malignant breast cancer tumors using breast cancer imaging dataset [21]. They achieve promising results, where the student model recieved a Precision of 0.8509, a Recall of 0.8623, a mIoU of 0.7778 and an Accuracy of 0.9526 [20].

2.4 Other approaches

Further analysis of XAI driven tools in healthcare is done by [22], where XAI techniques are leveraged to analyze breast cancer data and provide visual interpretations that enhance model decisions and clinical significance. The paper emphasizes the importance of XAI in medical applications, strengthening the decision-making process by understanding it, which is crucial for acceptance and trust within the medical domain. The paper also discusses the limitations of AI in the medical domain, due to their black-box nature, which results in difficult to interpret decision-making processes of these models. Therefore, the study highlights the need of explainable models to gain trust and confidence by medical professionals. The result of the XAI technique showed sensitivity ranging from 83 to 87%, and specificity ranging from 81 to 88% [22].

Furthermore, in [23], an EfficientNet model was used to classify breast cancer, several XAI techniques were also used to explore visualization. They state that the XAI techniques employed can produce confusing results, where weights had been interchanged. They also leveraged a method that involves perturbing images, and discovered that the method struggled with noisy images, which can lead to poor interpretations. The need for enhancements through XAI to improve model robustness by understanding the underlying model better is important. The paper also mentions that visualizations that highlight areas of interest can help understanding the model's decision making process and further improve the overall model.

In the study [24], the authors points out the importance of XAI in medial decision making, by building trust and clarity between AI and clinical users. The study focuses on breast cancer data, and by leveraging XAI, namely SHAP, the study is able to extract detailed visualizations that help determine the important features for the model's prediction, highlighting the need for clinical acceptance from domain experts to make decisions based on the offered transparency. The underlying model's XGBoost, Logistic Regression, Random Forest and SVM are evaluated using Precision, Recall and Accuracy. The study found the model with highest accuracy was XGBoost, achieving an accuracy of 85%. The paper also comments on the challenges of applied methods, where medical professionals may not be familiar with the model used. In addition, the paper also comments on how XAI techniques can be leveraged to better patient outcomes, through early detection based on model insights produced by XAI techniques.

Lastly, in the study [25], the authors use Gradient-weighted Class Activation Mapping (Grad-CAM) as XAI technique for improving explainability and interpretability of a pre-trained convnet model when identifying presense of metastases in lymph nodes. Grad-CAM is a type of saliency map based technique using class activation maps, that highlights regions of the image that are most influential to the model's prediction. The dataset used for this paper is a dataset of histopatholocical images of lymph nodes containing 220,000 training images and 57000 evaluation images. The model achieved an accuracy of 96.7%. By applying Grad-CAM, the authors found that they gained insights into the models decision

making process, which helped identifying features which were important for classification. The authors highlight the need and importance of XAI in medical imaging, as well as the potential trust model's recieves when combined with XAI.

Chapter 3

Theoretical Background

3.1 Breast Cancer and Medical Imaging

Breast cancer is a disease indicated by when cells within the breast begin to grow uncontrollably, leading to the formation of tumors [26]. These tumors can be benign or malignant, with the latter posing a significant threat as they have the ability to invade surrounding tissues and spread to other parts of the body [27].

Breast cancer is considered a significant health concern globally, being the leading cause of cancer deaths among women in the world [28]. According to World Health Organization (WHO), in 2022, approximately 2.3 million women were diagnosed with breast cancer, and there were 670,000 deaths from breast cancer globally [29].

Breast cancer is detected through various methods of medical imaging, and the need for sufficient and accurate equipment and methods is vital [30]. Many types of cancer are detected at an advanced stage, with poor prognosis and options for treatment.

In breast cancer one refers to cancerous nodules, which are small, solid mass of tissue that have different shape, size and can be either malignant (cancerous), or benign (non-cancerous) [31] [32]. The ability to detect cancerous nodules early is important for improving prognosis and treatment options in stages where the cancer has not yet developed to become advanced. However, the vision of early detection comes with a lot of challenges, such as finding existence of early cancer, and determination of who is at risk of developing cancer through genetics or demographic information [27] [30].

There are a lot of medical imaging techniques used in cancer imaging, each with specific uses [33]. The various techniques are dependent on a lot of factors such as malignancy, size of cancer, or where it is located in the body. Methods like Magnetic Resonance Imaging (MRI), mammography, ultrasound and Computed Tomography (CT) scans are common methods used for breast cancer imaging [34].

Ultrasound imaging is an important breast cancer imaging technique due to its ability to provide extensive analysis and diagnosis of breast cancer [34] [35]. The benefit of ultrasound over other techniques is the safety in examining dense breast tissues in situations where mammography might not capture everything. Ultrasound uses high-frequency sound waves

to create images of breasts, which is done without any radiation that could potentially harm the body. Ultrasound also benefits from being a cheaper and faster alternative than mammography [36].

3.1.1 Lesions

Lesions are defined as areas of abnormal tissue that have been identified by medical imaging or physical examination [37] [38]. Lesions often vary in their nature being either benign or malignant. Identification and detection of lesions play a critical role in early detection and diagnosis of breast cancer. Lesions serve as indicators for potential malignancy, and the patients needs for further investigation through additional imaging.

Detection of lesions can be challenging due to the variability in its nature, appearance, size and location [38]. These challenges can potentially complicate the diagnostic process and could potentially affect the accuracy of breast cancer diagnosis. Detection of lesions is typically done by various medical imaging strategies mentioned in 3.1.

3.1.2 Malignancy

Malignancy refers to the presence of cancerous or malignant cells that have the ability to grow rapidly and uncontrollably, and spread to other parts of the body [39]. The degree of malignancy is used as a measurement of how cancerous the lesions are, where a low degree of malignancy means the cancer cells can look and behave like normal cells, and with a high degree of malignancy, the cancer cells are usually more aggressive, look abnormal and have the ability to grow rapidly or spread quicker. Early diagnosis is therefore important in early stages of malignancy, due to being able to combat the illness before it reaches too high of a malignancy degree [35].

Breast cancer staging is a process that determines to which extent cancer has spread in the body [40]. This is assessed using a system called TNM, where T stands for Tumor size, N is the involvement of nearby lymph Nodes, and M is the presence of Metastasis. Stages range from early stage (I) to advanced stage (IV), with each stage having its own implications for prognosis and treatment planning. Staging aims to help medical professionals tailor the treatment strategies to the patient's needs and situation, as well as ensuring effectiveness within the existing methods while minimizing side effects [41].

The degree of malignancy, along with other information and tools such as biological and molecular markers, and cancer staging, helps decide what treatment decisions is made [42]. Early-stage, low-malignancy tumors may be treated with surgery followed by radiation, possibly avoiding chemotherapy. High-grade tumors may require a more aggressive approach, including chemotherapy, targeted therapy, and more extensive surgery [42].

3.2 Deep Learning

Deep learning is based on a layered structure of algorithms to process data, enabling models to learn and make complex decisions from large datasets by feature extraction and pattern

identification [43].

The development of more complex neural networks, such as the multilayer Perceptron and backpropagation introduced by D. E Rumelhart et al. in 1986 [44], marked a significant advance in machine learning. Adding multiple layers to machine learning models was introduced to solve more complex problems where the input potentially could consist of images or speech, where each layer would serve its purpose to perform specific types of transformations on the input data such as feature extraction and decomposition [45].

The general architecture of a neural network usually consists of an input layer, one or more hidden layers, and an output layer [45]. The input layer is fed the input data, the hidden layers process and transform this data by extracting complex representations of it. Non-linear activation functions, which is explained in 3.2.1, are often used in these layers due to their ability to introduce non-linearity and maintain gradient flow during training, which allows the model to be able to learn more complex representations of the data, and learn efficiently. The output layer produces the final predictions or classifications.

In Figure 3.1, a simple multilayer Perceptron is visualized. The multilayer Perceptron has fully connected layers, which means that every neuron is connected to its input neurons [45] [44].

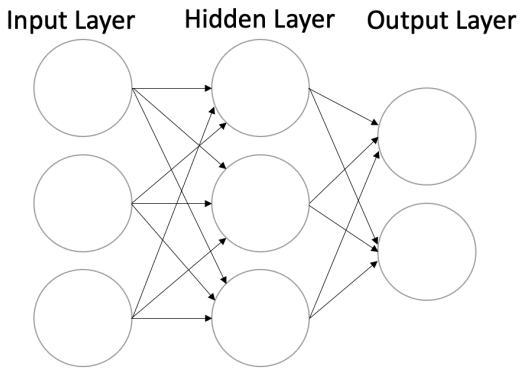


Figure 3.1: Mutlilayer Perceptron, consisting of an Input Layer, one Hidden Layer and an Output Layer, the arrows indicate the fully connected layers, where each neuron is connected to each neuron in the next layer

When data is passed through these layers, the network identifies important features, making it strong and efficient at recognizing for example patterns [45]. The process, often referred to as feature extraction, makes the network being able to learn complex representations in the data, not only simple shapes but also complex patterns.

One key feature of deep learning, and what makes it different from other machine learning methods, is the depth of its layers, meaning how many numbers of layers in a neural network, allowing efficient learning throughout its training process. Early layers tend to focus on simple patterns, such as edges and shapes, while deeper layers in the network focus on more complex structures, such as the presence of objects. This is referred to as a hierarchical learning process [46] [45].

Neural networks come in different variations and complexities to serve the purpose of different problems, and there is a wide range of possibilities when it comes to the network architecture, as the depth of the neurons and skip connections [47]. This allows for great versatility when working with various problems where the complexity of the task at hand may vary. With more complex models or problems comes the need for heavier computational resources. A critical aspect of more complex models is the potential of overcomplicating the model making it generalize poorly to unseen data, called overfitting. On the other hand is underfitting, which is a phenomena when a model is too simple to learn the underlying pattern of the data [45].

Many deep learning architectures are referred to as black-box models where the models internals are either hidden or unknown to the observer, or even if known, cannot be understood by humans [48]. Such black-box models can pose ethical and practical dilemmas when used in specific scenarios, especially in critical sectors like healthcare, where decision-making processes can affect patient outcomes. The inability to fully understand or interpret how these models arrive at their conclusions complicates efforts to ensure fairness, accountability, and transparency. In healthcare, where decisions can directly influence patient health, diagnosis, and treatment options, the need for methods like XAI becomes important. Addressing these challenges involves developing models that not only perform with high accuracy but also being interpretable and transparent in their decision-making processes. This ensures that in vital areas like healthcare, it could aid professionals to trust and effectively integrate AI tools into their decision-making framework, enhancing patient care while upholding medical standards [49].

3.2.1 Convolutional Neural Network (CNN)

Feature extraction is fundamental for machine learning algorithms [50]. Convolutional Neural Network (CNN) is designed to learn features automatically from raw data that are most useful for the particular task at hand, often used in image recognition and computer vision [45]. The development of CNNs was advanced by Yann LeCun when the creation of the LeNet-5 architecture took place, which is one of the first applications of CNN [50].

CNN are often referred to as feature extractors. The early layers of the CNN extract low-level features such as basic shapes and edges. As it works its way through the layers, the deeper one progresses through the network, the more complex and high-level features are extracted. CNNs construct a feature hierarchy, meaning it combines the low-level features to form high-level features such as objects like buildings or humans [50] [45].

A typical CNN consist of unique layers being convolutional layers and subsampling layers [45]. The convolutional layer consist of convolutional operations, which is a core idea of CNNs. The convolutional operation is the technique of sliding a filter over the input data, then computing the dot products. This creates a feature map of the convolutions, which are done to downsample images focusing on the relevant features. The formula for convolution in 2D is given by Equation 3.1.

$$\boldsymbol{Y} = \boldsymbol{X} * \boldsymbol{W} \rightarrow Y[i,j] = \sum_{k_2=-\infty}^{+\infty} \sum_{k_1=-\infty}^{+\infty} X[i-k_1, j-k_2] W[k_1, k_2] \tag{3.1}$$

where $\boldsymbol{Y} = \boldsymbol{X} * \boldsymbol{W}$ is the 2D convolution between the input X and filter W [45]. The subsampling layers consist of pooling techniques, usually max- or mean-pooling [45]. The pooling layer's task is to perform the operation of the n_1, n_2 array of pixels denoted as $P_{n_1 \times n_2}$. The aim of the subsampling layer is to decrease the size of features, resulting in higher computational efficiency and potentially reducing the risk of overfitting. In Figure 3.2 an example of a 3x3 max pooling is displayed.

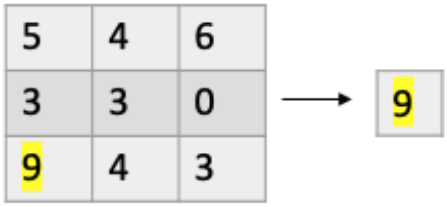


Figure 3.2: Example of a single 3x3 max pooling operation, the highlighted yellow number indicate the number that was chosen after performing the max pooling operation

Dense Layers, also known as fully connected layers are essential to CNNs, primarily used for classification after feature extraction from previous convolutional and pooling layers [45]. In dense layers each neuron receives input from all neurons from the previous layer, which is why it is also called a fully connected layer. Due to their fully connected nature, dense layers are prone to overfitting, especially when working with large inputs. To lower the risk of overfitting and reduce complexity of a model, L2 regularization is often applied [45]. L2 Regularization shown in Equation 3.2, works by adding a penalty term to the cost function that is proportional to the sum of the squares of the weights w_j. This essentially motivates the weights to stay small by penalizing large weights. This leads to a simpler model, and prevents overfitting.

$$L_2 : \|\mathbf{w}\|^2 = \sum_j w_j^2 \tag{3.2}$$

The dense layers utilize activation functions to introduce non-linearity to the system [45]. Three common activation functions are displayed; Rectified Linear Unit (ReLU), Logistic (sigmoid) and Hyperbolic tangent (tanh) respectively. The ReLU function, all values below 0 is set to 0 and all values above 0 remains unchanged, given by Equation 3.3, where z is the input to the activation function, and $\phi(z)$ is the activation function ReLU [45].

$$\phi(z) = max(0, z) \tag{3.3}$$

The sigmoid activation function in Equation 3.4 outputs values between 0 and 1 where $\phi(x)$ is the sigmoid activation function, and x is the input to the activation function [45].

$$\phi(x) = \frac{1}{1 + e^{-x}} \tag{3.4}$$

The Hyperbolic Tangent Function (tanh) in Equation 3.5 outputs values between -1 and 1, it is used for large negative or positive values, where $\tanh(x)$ is the activation function and x is the input to the activation function [45].

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3.5}$$

3.2.2 EfficientNetV2

Transfer learning is a powerful technique in deep learning where a model developed for a certain task is reused as the backbone for a model on a another task. This type of technique differs from traditional machine learning pipelines, where training and testing data have the same input feature space and data distribution. In transfer learning, the backbone of the model is already made. This approach is especially beneficial in scenarios where there is little labeled data, or where training from scratch is computationally expensive. It also potentially provides efficiency over traditional machine learning methods due to its pre-trained nature [51] [52].

EfficientNetV2 utilizes transfer learning and builds on advancement in convolutional neural networks designed for both improved training speed and enhanced parameter efficiency. EfficientNetV2 models were developed using a training-aware neural architecture search and scaling approach, focusing on the optimization of both training speed and parameter efficiency. This essentially means that the development of this model focused on both maximizing accuracy or minimizing error and optimization of the training speed and efficiency [53].

EfficientNetV2 aims to improve EfficientNet by scaling up the network's depth, width and resolution based on fixed scaling coefficients [53]. By focusing on reduced training time, EfficientNetV2 adjusts these coefficients to achieve better balance between accuracy and efficiency. EfficientNetV2 utilizes a progressive training approach where it starts off by training on small images, and images are increased in size the further down the training process it goes. Early stages is therefore sped up by training on smaller images [53].

3.3 Explainable AI

XAI aims to convert the black-box nature of machine learning algorithms into a transparent system [2]. XAI aims to make the machine learning models able to explain their predictions by leveraging techniques like post-hoc interpretability such as feature importance and SHAP values, or visualization techniques provided by LIME, where visualization of changes in input affects the models output or highlights of parts of images that are significant for

predictions [54] [16].

Explainable AI can be seen as an advanced topic due to its non-existent universal definition of interpretability [55]. Because of this, a uniform definition of interpretability and explainability will be used throughout the thesis.

3.3.1 Interpretability vs. Explainability

The terms interpretability and explainability are often used vaguely missing a uniform agreement of the terms in the literature surrounding artificial intelligence, while they cover unique concepts within the domain of XAI. Understanding the differences between them is crucial for the development and evaluation of XAI systems, as it directly impacts how humans can interact with and trust AI outputs [56] [57].

Interpretability involves the degree to which a human can comprehend the cause-and-effect within a system. In the context of machine learning, this means understanding how the model's input variables relate to its predictions or decisions. An interpretable model allows users to predict the model's outcome with reasonable accuracy, given a set of inputs. For simple models such as linear regression or decision trees, interpretability can often be simple due to their straightforward and logical structure [58].

Explainability goes a step further by providing the reasons behind specific decisions or predictions. It involves generating human-understandable explanations for the operations and results of complex models. XAI seeks to explain the model's decision-making process, translating it into a form that is understandable by the users. These explanations may include the identification of key features that influence the output or provide insights into the model's inner workings [59] [58].

3.3.2 SHAP

SHAP is an XAI technique that uses concepts from game theory to assign each feature in a model importance, based on the contribution of the feature on a prediction [54]. The method calculates how much each feature contributes to the prediction of an instance, which essentially means that SHAP's framework uses local explainability, and for simple models, is based on the idea that the original model provides the best explanation for the model itself [54]. For more complex models, like deep learning models, the model itself is too complex, and a simpler version to gather its explanations is needed. The simpler model is an approximation of the original model and is represented by an additive feature attribution model given by Equation 3.6.

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{3.6}$$

Here, f is the original model we aim to explain and $g(z')$ is the simpler explanation model of $f(z)$, ϕ_0 is the base value that would be predicted if all features were missing.

ϕ_i is the SHAP value for each feature. z'_i are the binary variables indicating either the presence defined as 1 or absence defined as 0 of the ith feature. M is the total number of simplified input features. This is meant to give a measure of each prediction and their contribution to each individual feature [54].

To compute SHAP values, three properties, derived from game theory must be fulfilled [54]. The objective of these properties is to allocate the prediction outcome fairly among the features based on their individual contributions.

The first property given by Equation 3.7 is the property of **local accuracy**. This property aims to ensure the explanation model is accurate locally [54].

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{3.7}$$

Local accuracy in Equation 3.7, essentially means that the sum of all SHAP values from the explanation model $g(x')$ for input x', exactly match the sum of the SHAP values from the original model $f(x)$ for input x, where x' is the simplification of the input x. ϕ_i is the SHAP values for feature i, and ϕ_0 is the base SHAP value. M is the total number of simplified input features [54].

The second property is **missingness**, given by Equation 3.8 [54].

$$x'_i = 0 \implies \phi_i = 0 \tag{3.8}$$

The property of missingness in Equation 3.8, states that if a feature x'_i is missing from the input, then that feature should not contribute to the model's prediction for that specific input. The SHAP values ϕ_i for the specific feature x'_i should therefore be equal to 0 [54].

The third and last property is **consistency**. The property of consistency is used to compare how different models behave when excluding a feature. Consistency states that if it makes a bigger difference when removing a feature from one model f's predictions than the other model g's predictions, then the respective SHAP value for the model f should be greater or equal to the SHAP value for the model g. This ensures consistency in SHAP values with respect to features in different models [54].

The classic SHAP value calculations are given by Equation 3.9.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right) \tag{3.9}$$

SHAP values are calculated from a model by evaluation of how the prediction changes when a feature i is added or removed to the subset S. S represents the subset of all features F when excluding feature i. The difference in prediction is given by $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$, where x_S is the input values for the subset of features S. The ! symbol represents the factorial operation, used to calculate number of permutations, F represents the set of all features in the model [54].

3.3.3 LIME

LIME is an XAI technique introduced in [16]. LIME is used in machine learning to provide explanations for the predictions made by complex machine learning models, particularly black-box models. LIME focuses on providing local interpretability, aiming to explain the prediction of a model for a specific instance or data point [16]. LIME is model-agnostic, which means it can be applied to explain the predictions of any machine learning model, regardless of its complexity or type. LIME achieves this by approximating the behavior of the black-box model locally with a simple, interpretable model, such as a linear regression model. LIME constructs a dataset for training the simple model by changing input features and watching the resulting variations in the predictions of the black-box model. The resulting explanations highlight the importance of specific features and their contributions to the prediction, making it easier for humans to understand why the model made a particular decision. LIME is a valuable tool for improving the transparency and trustworthiness of machine learning models, making complex black-box models more adaptable and understandable for humans [16].

For instance, consider an input x and a model f, LIME generates a new dataset of n perturbed samples around x and obtains predictions for these samples using the model f [16]. After fitting on the perturbed samples, it weighs these based on their closeness to x and fits a simpler, interpretable model to these samples. The coefficients of this new model, g, serve as the explanations for the prediction at x. The formula is shown in Equation 3.10.

$$\text{argmin}_{g \in G} \, \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{3.10}$$

where $\mathcal{L}$ is the loss function that explains how well the local explanation model g approximates the predictions of the full model, f, in the locality defined by π_x. π_x is a similarity measure that weighs the slightly perturbed samples according to their closeness to the prediction instance x. $\Omega(g)$ is a complexity measure of the explanation model g, aiming to encourage simpler models for easier interpretation by humans. The optimization seeks to find g, which often is a linear model, such that it both accurately approximates f around x and remains interpretable [16].

3.4 Evaluation Metrics

In this section the various evaluation metrics considered in this thesis will be described. TP, FP, TN, FN are acronyms for True Positives, False Positives, True Negatives and False Negatives respectively.

TP, TN, FP and FN are calculated by an analysis of images processed using XAI techniques. The process applies these techniques to each image to identify features which are considered regions of pixels that significantly influence the model's prediction.

These values are then converted to a binary format suitable for calculation, where a fixed threshold value is specified. Scores exceeding this threshold are classified as 1, and

scores below are classified as 0. Furthermore, the actual mask of the image, which holds information about the true area of interest, is binarized similarly to the XAI output array. Areas within the mask are marked as 1, and areas outside are marked as 0. TP, TN, FP and FN are then calculated as follows:

- TP: The sum of instances where both the actual mask and the predicted mask correctly identify an area of interest (both are 1).
- TN: The sum of instances where both the actual mask and the predicted mask correctly identify an area as not of interest (both are 0).
- FP: The sum of instances where the predicted mask incorrectly identifies an area as of interest while the actual mask does not (predicted is 1, actual is 0).
- FN: The sum of instances where the predicted mask fails to identify an area of interest that the actual mask does (predicted is 0, actual is 1).

3.4.1 Accuracy

Accuracy represents the proportion of predictions that the model correctly predicts out of all the predictions it makes, given in Equation 3.11 [60].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{3.11}$$

3.4.2 Precision, Recall and F1-score

For evaluation of the XAI techniques leveraged in this thesis, four metrics will be considered. These are F1-score, Precision, Recall and IoU respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.12}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.13}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.14}$$

Precision, given by Equation 3.12 aims to capture how precise the XAI output is. It measures the models ability to accurately predict relevant areas within the image. It measures the proportion of the predicted area of pixels which falls inside the correct mask [60].

Recall, given by Equation 3.13 aims to provide a measurement of the XAI output's effectiveness in capturing the total significant area given by the mask. It evaluates to what extent the predicted mask by the XAI technique targets the correct relevant area of pixels defined by the ground truth mask [60].

F1 Score, given by Equation 3.14 is used as a balanced mean of both Precision and Recall. This metric provides a balanced measure that both uses the precision of identified features and the model's ability to capture relevant areas of features [60].

3.4.3 Intersection over Union

Intersection over Union (IoU) is a metric that calculates the Intersection of the Union which calculates the rate of overlap between two areas [61] [62]. It is used in this thesis to measure the area of overlap of the XAI techniques output and the correct mask provided. IoU provides a measure of overlap that can consistently be applied across different images and XAI output, helping evaluate the analysis of XAI's effectiveness and correctness.

Furthermore, a strong feature of IoU is its simplicity and robustness, making it easily adaptable to a variety of XAI techniques, regardless of the underlying model complexity. The formula for IoU is given by Equation 3.15 [62].

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{3.15}$$

where Area of Intersection is the overlapping area between the ground truth and the predicted area, and Area of Union is the total area of both ground truth and prediction minus the Area of Intersection in Equation 3.15.

In this chapter a comprehensive overview of the theoretical background was presented, setting the stage for the analysis conducted in the thesis. The upcoming chapter will provide methodology, which demonstrates how the theoretical framework helped analytical processes done in this thesis.

Chapter 4

Methodology

This chapter outlines the methodology employed in this study, focusing on investigating if XAI can be efficiently leveraged for enhancing predictions in medical imaging. The design was chosen due to its ability to provide insights into the decision-making process of AI models, enabling an evaluation of their reliability and accuracy. This approach aligns closely with the study's objectives to enhance interpretability and explainability in black-box models working with imaging data.

The study was conducted using real healthcare data, targeting a population of medical images from patients with cancerous and non cancerous nodules in breasts [21].

4.1 Data Management

4.1.1 Data Collection and Description

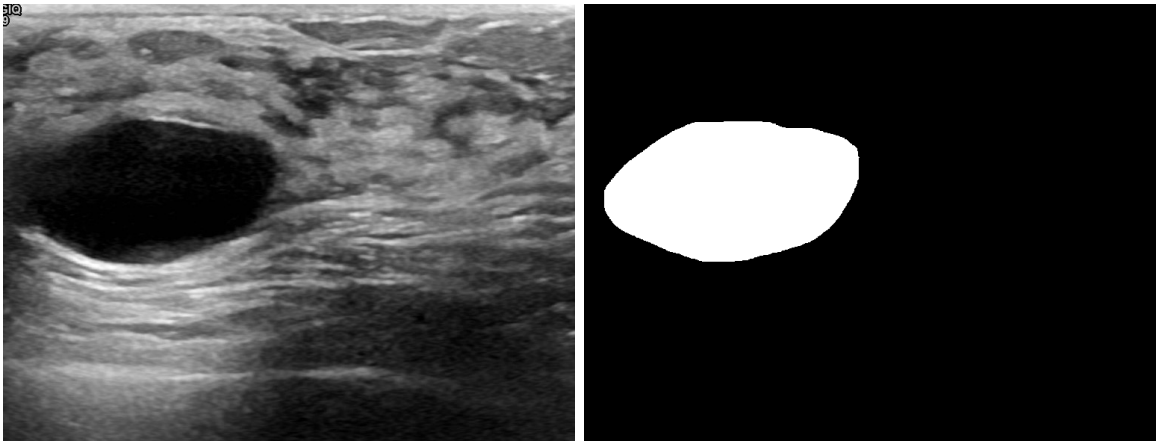
Data collection was carried out through direct acquisition from the publicly available dataset acquired from medical imaging equipment, utilizing images in PNG format [21]. The dataset used for this thesis was the Breast Ultrasound Images dataset [21], which consisted of 2D ultrasound scan images for breast cancer diagnosis, as well as Positron Emission Tomography (PET) masks for identifying the location of the tumor, done by professionals. The samples are classified as one of three classes: *normal, benign and malignant* cases. Table 4.1 showcase the image distribution among the three classes. The dataset consisted of 780 images with an average size of 500 x 500 pixels from a total of 600 female patients between the ages 25 and 75 years old. A total of 437 images are classified as benign, 210 images are classified as malignant and 133 is classified as normal. A benign case and its respective mask is shown in Figure 4.1, as well as an malignant and its respective mask is shown in Figure 4.2 [21].

4.1.2 Data Preparation

Various methods for preparing the data for training is explained in this chapter. To feed the data into the model, a uniform shape of $(224, 224, 3)$ were needed for correct input shape for the chosen model. Since the images in the Breast Ultrasound Images dataset has

Table 4.1: Distribution of number of images by case, Breast Ultrasound Images dataset

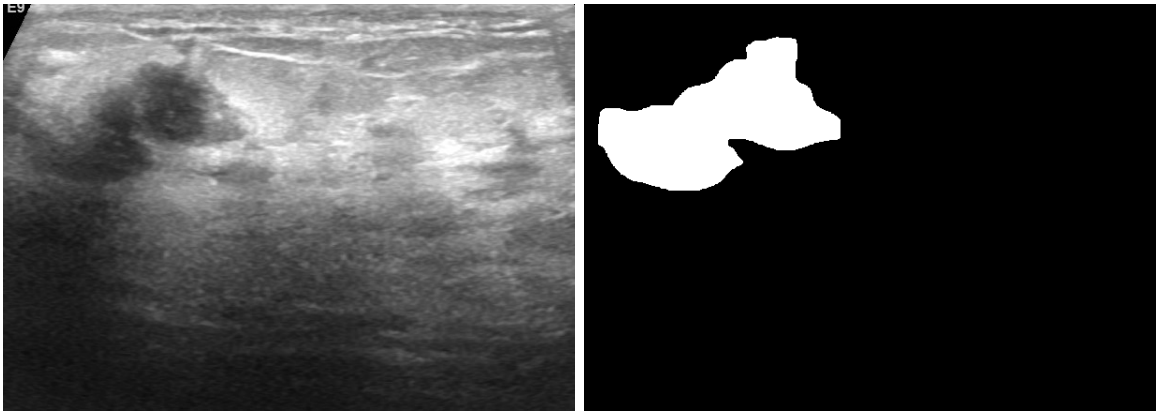
Case	**Number of Images**
Benign	437
Malignant	210
Normal	133
Total	780



(a) Original Image of a Benign Case.

(b) Masked Image Showing Affected Benign Areas highlighted in white

Figure 4.1: Comparison of a benign case with and without a diagnostic mask. Figure (a) shows the original medical imaging of a Benign case, while Figure (b) demonstrates the mask applied to the same image to highlight the affected areas in white



(a) Original Image of a Malignant Case.

(b) Masked Image Showing Affected Malignant Areas highlighted in white

Figure 4.2: Comparison of a benign case with and without a diagnostic mask. Figure (a) shows the original medical imaging of a malignant case, while Figure (b) demonstrates the mask applied to the same image to highlight the affected areas in white

varying sizes, the images were resized to the respective shape to ensure consistency in the results and input throughout the workflow.

Data Augmentation Data augmentation techniques were used to expand the datasets size and variability, including horizontal flipping, vertical flipping, rotation, width and height shifting, and zooming. A set of specific augmentation strategies was implemented to minority classes being Malignant and Normal, to address class imbalance. These techniques ensured a more robust model, that generalized better with unseen data. The data augmentation was carried out through the use of **ImageDataGenerator** from TensorFlow's Keras [63] [64].

Data Splitting The data was split into masks and images, which were later used for quantifying the performance, explainability and interpretability of our XAI techniques. The images were further split into training and testing subsets, where 75% were used for training, and 25% for testing. The splitting was performed by using the traintestsplit function from the sklearn module of the scikit-learn library [65]. This ensured a sufficient amount of data for both training and evaluation of model performance.

Class Weight Calculation Since the dataset used contained class imbalances, class weights were introduced to count for class imbalances in the dataset. This ensured the model's performance did not bias towards the majority class being Benign case. To compute this the *compute_class_weight* function from *sklearn* to generate balanced class weights was used [65]. This process was applied before model training.

4.1.3 Ethical Considerations

Ethical considerations were an important topic when working with medical imaging data. Recognizing the sensitive nature of medical data, the needed measures were taken to uphold ethical standards, ensuring respect, privacy, and protection for all individuals whose data were involved.

The potential impact of this thesis' research was considered, focusing on contributing positively to the field of medical imaging. The study aims to improve patient outcomes and support clinical decision-making, while also raising awareness of ethical implications of AI in healthcare.

4.2 Model Development

In this chapter the development of the deep learning model used for experiments in the thesis is showcased through a detailed description of the workflow of developing the model.

4.2.1 Model Workflow

A transfer learning approach of a CNN based model EfficientNetV2B2 was implemented for image classification of ultrasound medical images [53]. The motivation for the choice of this specific model was that it offers efficient training, while also maintaining optimal performance [53]. The model workflow displayed in Figure 4.3 shows the architecture of

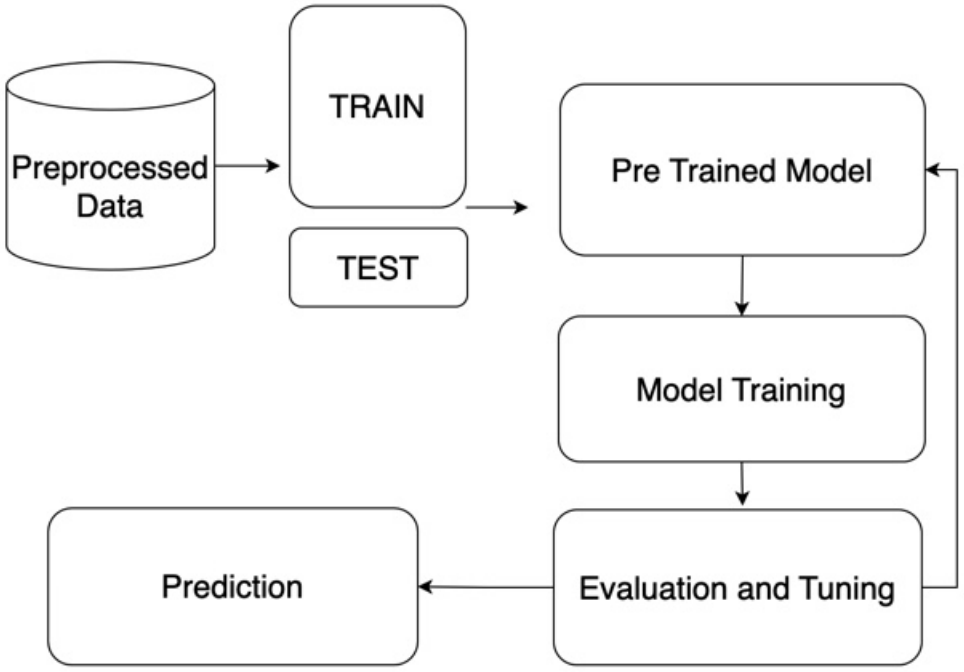


Figure 4.3: Model workflow, Preprocessed Data is the starting point where the original dataset was preprocessed, Train Test Split involved splitting data into Train and Test set of 75% and 25% respectively. Pre Trained Model; initialize pre-trained EfficientNetV2B2 using transfer learning. Model Training, train model on test data. Evaluation and Tuning; Evaluation of parameters to optimize performance. Prediction; make predictions on test data.

building the model. The first step in model development was splitting the preprocessed data which were resized to the uniform shape of $224 \times 224 \times 3$, where the ultrasound images were loaded as well as their corresponding masks belonging to the three unique categories benign, malignant and normal. To further split the data into train and test sets as described in Chapter 4.1.2, the data was extracted to a full dataset by combining all three classes into one dataset in Figure 4.3. Preprocessing also involved augmenting images, as described in Chapter 4.1.2, as well as calculating class weights for minority classes as discussed in Chapter 4.1.2. Now the data was ready for model deployment. The next part was where the pre-trained Model was initialized using transfer learning, the EfficientNetV2B2 model was employed, which was pre-trained on ImageNet. This was followed by Model Training in Figure 4.3, both described in chapter 4.2.2. The next step of the model workflow is the Evaluation and Tuning where the model's metrics described in Chapter 4.2.2, were evaluated. Finally after these steps, model prediction was done shown as Prediction in Figure 4.3.

4.2.2 Implementation details

The model architecture leveraged both the robustness of the pre-trained EfficientNetV2B2 [53] architecture, which was pretrained on ImageNet by setting **weights=imagenet**, providing a robust architecture for image classification tasks [66], and custom layers described. The model was used without its top layers to allow construction of custom layers to improve model generalization. The input shape was set to $224 \times 224 \times 3$ to match the dimension requirement for the EfficientNetV2 model [53] [67].

The custom top layers included the additional pooling layer, **GlobalAveragePooling2D**, used to reduce dimensionality as well as minimizing the risk of overfitting. Dense layers were also added for penalizing large weights. These were added with both ReLU activation function and L2 regularization respectively. Dropout layers were also introduced with rates of 0.5 and 0.2 respectively to further help minimize the risk of overfitting. The hyperparameters 0.5 and 0.2 were chosen due to these provided the highest metrics when trying different hyperparameters.

The final layer was a dense layer with three output units to match our classification task. The final dense layer used the **'softmax'** activation function to generate a probability distribution among the three classes.

The model was compiled with the **Adam** optimizer, which contributes to efficient convergence [68], with a learning rate of 0.001. 0.001 was at default. The model training was done with the sparsecategoricalcrossentropy loss function, which is suitable for multi-class classification [64] [63].

Software and Libraries Used For the model and XAI development, Google's Colaboratory platform, Google Colab was used. Google Colab provides a cloud-based environment, with powerful GPUs for efficient training of complex models. Google Colabs Pro version was leveraged to access more powerful hardware for extensive training and model development, especially the High-RAM feature. All development was done using Python 3.10.12.

In Python, several libraries were used to develop the experiments. The library NumPy version 1.25.2 was used to perform operations and transformation of arrays [69]. Tensorflow version 2.15.0 was used to build and train the model [63]. Keras version 2.15.0 was key to implement the model, as well as transform and augment data [64]. Scikit-learn was leveraged to evaluate model and other machine learning utilities [65]. SHAP version 0.45.0 [54] and LIME version 0.2.0.1 [16] were used to implement the XAI techniques.

Model Evaluation Metrics Consistent evaluation of model performance was important during development. For the model development the following metrics were used to ensure consistency between results. These were; Accuracy, Precision , Recall and F1-score [60]. These scores were calculated using the classification report from sklearn [65].

4.3 Explainable AI (XAI) Techniques

In this chapter, the XAI techniques SHAP and LIME, their implementation details and workflow is outlined. These techniques were chosen to answer the research questions due to a wide variety of previous literature covering their applications in the domain of XAI as discussed in Chapter 2.

4.3.1 XAI Evaluation Metrics

For the evaluation of the respective XAI output, four evaluation metrics were used to measure the XAI output. These were Precision, Recall, F1-score, and IoU. The components TP, TN, FP and FN of Precision and Recall for the XAI evaluation were calculated as shown in Algorithm 1.

Algorithm 1 Calculate TP, TN, FP, FN for XAI evaluation Precison and Recall

function CALCULATETPTNFPFN($actualMask, predictedMask$)
    $TP \leftarrow \text{SUM}((actualMask = 1) \wedge (predictedMask = 1))$
    $TN \leftarrow \text{SUM}((actualMask = 0) \wedge (predictedMask = 0))$
    $FP \leftarrow \text{SUM}((actualMask = 0) \wedge (predictedMask = 1))$
    $FN \leftarrow \text{SUM}((actualMask = 1) \wedge (predictedMask = 0))$
    return $\{TP, TN, FP, FN\}$

The components of IoU being Area of Intersection was calculated as the overlapping area between the ground truth mask and the predicted area by the XAI output. The other part of IoU being Area of Union was calculated by the total are of both ground truth mask and the predicted area from XAI output minus the Area of Intersection. These two components were then divided as shown in Equation 3.15.

4.3.2 SHAP Implementation

The first proposed explainability technique was SHAP. SHAP uses Shapley Values to calculate feature importance [54]. For images, a feature is considered as a pixel, or group of pixels. The SHAP values for a model predicting on image data, is given as a measurement of the impact of including or excluding a specific region of the image in its predictions.

The workflow for the SHAP implementation is given by Figure 4.4. A subset of 128 total images of benign and malignant cases were used to calculate the SHAP values, 128 images were chosen due to the complexity of time when calculating SHAP values. To calculate the SHAP values, a SHAP explainer was initlialized using a predefined model EfficientNetV2B2, as well as a masker to apply blur effects on the images. The masker aims to help manage how features in input data were hidden or altered when computing SHAP values. The masker used in this experiment was "blur(32,32)". This essentially blurs the image, providing a balance between detailed information and maintaining visual features for meaningful predictions. The input is prepared in uniform shape to match underlying models input shape, and a test set of 128 images of malignant and benign

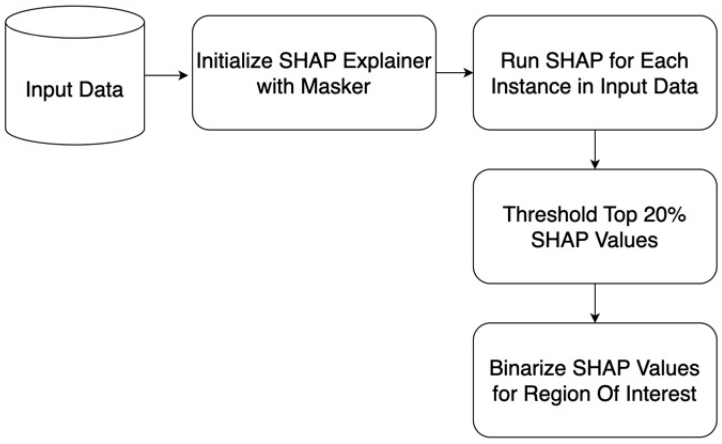


Figure 4.4: SHAP workflow, Input Data; initial dataset used for SHAP calculations. Initialize SHAP Explainer with Masker; Set up the SHAP explainer to gather explanations. Run SHAP for Each Instance in Input Data; Apply the SHAP explainer to each instance in the input data to compute respective SHAP values, which quantify the impact of each feature on the prediction. Threshold Top 20% SHAP values to get the most meaningful features. Binarize SHAP Values for Region Of Interest; Convert the SHAP values to binary format to gain presence of significant features to further analyse with metrics.

cases was chosen. For each image, the respective SHAP values were calculated using the initialized SHAP explainer. To focus on the most influential regions, a threshold of the highest 20% SHAP values were set for binarization; the highest 20% were set to 1 and the rest were set to 0. This was done to select the most influential regions of pixels positively to contributing to the correct classification of the model and to calculate the metrics to evaluate its performance in Figure 4.4. The threshold of 20% was empirically chosen after testing with various thresholds. This threshold provided the best explanation without including unnecessary noise.

4.3.3 LIME Implementation

The second proposed explainability technique was LIME [16]. LIME focuses on approximating the model locally around the prediction by creating a surrogate model to approximate the behavior of the original model near the instance being explained. This technique uses interpretable models to explain individual predictions [16].

The workflow for LIME implementation is illustrated by Figure 4.5. The first step in the LIME workflow was Data Preperation in Figure 4.5, which was crucial for matching the requirements such as shape of the underlying model. The same set of images were chosen for LIME as for SHAP, where a total of 128 images of malignant and benign cases were considered, this were chosen due to the limitation considering complexity of time as well as limitation of available data.

The second step in the workflow was initializing an explainer. The explainer initialized was the LimeImageExplainer [16], using the predefined model EfficientNetV2B2 as the

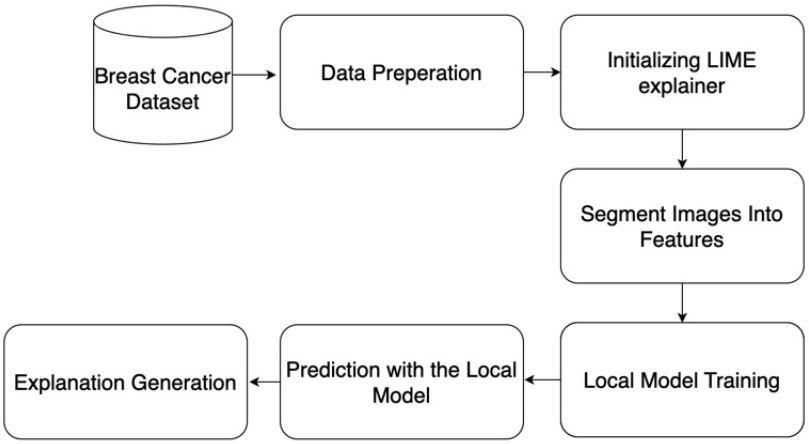


Figure 4.5: LIME Workflow, Breast Cancer Dataset; This is the input data, Data Preperation; In this step the data is processed to match requirements of underlying model being 224x224x3 . Initializing LIME Explainer; LIME Explainer is used to generate interpretable explanations from underlying model. Segment Images Into Features; Segmented by LIME to understand each effect of the segmentations on the prediction. Local Model Training; Training the simple interpretable model from LIME on new dataset. Predictions With the Model; Local model makes predictions on the perturbed data to provide insights into how different features contribute to prediction. Explanation Generation; Based on local model's predictions, explanations that detail each feature's contribution to the final prediction is generated.

underlying classifier from which predictions are explained. A perturbation method was employed by LIME's built in explaininstance method, which segmented the image into features and systematically created a new dataset of perturbed images. The new dataset was then used to train a local model around the prediction being explained and observe the variations in its predictions for each instance. It helped quantify the impact of each feature on the model's prediction. This helped LIME to calculate how each feature contribute to the final decision, producing an interpretable model that locally approximated the behavior of the underlying classifier model.

For each image the top 10 highlighted regions chosen by LIME for being influential in the models underlying predictions, specified by calling toplabels=10. This was chosen after testing various thresholds, where it was found that 10 was the ideal number for gathering the most influential regions that influenced the models prediction, without including too much noise which made the explanations too vague.

These highlighted regions were saved as masks to further analyze LIME's output compared to the original mask when calculating the evaluation metrics.

Chapter 5

Results

This chapter presents results of the analysis done by leveraging the XAI techniques SHAP and LIME. The findings presented in this chapter shows the possibilities of using XAI techniques like SHAP and LIME for improving the broader understanding and explanation in complex deep learning models. By analyzing the performance metrics IoU, Precision, Recall and F1-score, this study demonstrates how XAI techniques can be efficiently leveraged to uncover decision-making processes of complex deep learning models. These insights contribute to uncovering black-box models, aiming to make them more transparent and able to explain their decisions, which is particularly important in critical domains such as healthcare.

5.1 Model Evaluation and Performance

This chapter provides insight on the models results, as well as the data used for training and testing.

Table 5.1: Detailed Classification Performance Metrics for EfficientNetV2B2 Model on Breast Cancer Imaging Dataset, with Three Classes Benign, Malignant and Normal, displaying Precision, Recall, F1-score for all three classes and Macro and Weighted Average of these.

Class	Precision	Recall	F1-Score	Support
0	0.90	0.93	0.92	112
1	0.89	0.80	0.85	51
2	0.82	0.88	0.85	32
Accuracy			0.89	195
Macro Avg	0.87	0.87	0.87	195
Weighted Avg	0.89	0.89	0.89	195

Table 5.1 shows the classification scores of the model. The model received a high predictive accuracy on the test set of 0.89 shown as Weighted Avg in the Table 5.1, which consisted of 195 unique patients. This indicated that the model correctly predicts the correct class in 89% of cases across an unbalanced dataset.

	Predicted Labels			
	0	**1**	**2**	
Benign	104	4	4	**0**
Malignant	8	41	2	**1**
Normal	3	1	28	**2**

Table 5.2: Confusion Matrix of the EfficientNetV2B2 Model Classification on Breast Cancer Dataset, detailing the correct and false predictions for each class Benign, Malignant and Normal.

The classification scores shown in Table [5.1] shows the precision, recall and F1-score for each class. Class 0, which was patients classified as benign, received precision of 0.90, recall of 0.93 and F1-score of 0.92. Class 1, being patients classified as malignant, shows a precision of 0.89, recall of 0.80 and an F1-score of 0.85. Class 2, which was the patients classified as normal, received a precision of 0.82, a recall of 0.88 and an F1-score of 0.85. The weighted average of precision, recall and F1-score is 0.89 for all three metrics.

In Table 5.1 Support is mentioned, which refers to the number of samples in the specified class in the dataset, this was to provide insight into the dataset's distribution among classes. In Figure [5.2] the various predictions among classes is displayed. It is observed that the test set is unbalanced, and has the majority of patients in class 0 with 104 correct predictions in class 0, 41 correct predictions out of 51 for class 1 and 28 out of 32 correct predictions for class 2.

5.2 Explainable AI (XAI) Evaluation

In this chapter the output and performance metrics from both XAI techniques, SHAP and LIME, is displayed. Tables 5.3 and 5.5 summarize the evaluation metrics for SHAP and LIME respectively. These tables provide mean, median, and standard deviation values for the following metrics: IoU, Precision, Recall, and F1-Score. Table 5.4 and Table 5.6 presents the best performance metrics for one sample for each technique.

5.2.1 SHAP Analysis

The SHAP evaluation metrics are displayed in Table 5.3, which offer insights in reliability of the explanation outputs. The average IoU of 0.14 indicated a poor overlap between the model's prediction and the actual ground truth. An average of 14% indicated that the model might look at some parts of the actual mask, but not the full region of pixels. Average Precision at 0.60 indicated that 60% of the area the model classified as important falls under the correct mask. This indicates that the model was able to capture the correct region of pixels in the image and classify it as an important feature 60% of the time. A low

Table 5.3: SHAP Evaluation Metrics for the EfficientNetV2B2 model on Breast Cancer Images. The Table Includes Metrics IoU, Precision, Recall, and F1-score, where all metrics' Average, Median, and Standard Deviation is displayed except for the F1-score which displays only the Average.

Metric	IOU	PRECISION	RECALL	F1-SCORE (Average)
SHAP				
Average	0.14	0.60	0.20	0.30
Median	0.12	0.58	0.15	
Standard deviation	0.12	0.30	0.19	

Table 5.4: The top SHAP Evaluation Metric for the EfficientNetV2B2 model on Breast Cancer Images. The Table Includes Metrics IoU, Precision, Recall and F1-score.

Metric	IOU	PRECISION	RECALL
SHAP			
Top 1 evaluation metrics	0.43	1.0	0.89

average Recall at 0.20 is shown for SHAP, which indicated that the model often included areas as important features which did not fall under the ground truth mask provided. An average F1 score of 0.30 suggested that SHAP often missed important areas of pixels.

Standard deviation values for each metric is displayed in Table 5.3 to demonstrate the variation observed by the explanation technique. It was observed a relatively high standard deviation among the predictions, meaning that the different metrics varied depending on the relative image. For IoU, the average standard deviation was 0.12, which means that the predicted values and the ground truth overlap were inconsistent. In Precision an average standard deviation of 0.30 was observed. This highlights potential fluctuations in some cases regarding the XAI techniques' ability to identify areas that were correct when compared to the ground truth mask (True Positives).

Recall's standard deviation was 0.19, which showcased a relatively lower variability in Recall compared to Precision.

A high standard deviation indicated that the performance might be influenced by different factors like image quality or presence of similar characteristics in benign and malignant cases.

In Table 5.4, the best performance metrics for SHAP are displayed for a single image. These are meant to display the positive possibilities for SHAP in healthcare imaging. The best IoU of the test set was observed at 0.43, this indicated a moderate overlap between the predicted and ground truth mask, meaning that 43% of the predicted important region of pixels were within the ground truth mask. Essentially this means that 43% of the region predicted by model and SHAP was considered significant for making a decision where the decision accurately aligned with the ground truth mask.

A Precisison of 1.00 is displayed for the best prediction in Table 5.4, this indicated that the whole predicted area by SHAP was considered significant when compared to

the ground truth mask. There were no false positives in the SHAP output, showcasing a 100% predicted area of pixel, which was inside the ground truth mask, with no areas of significance falling outside the ground truth. This described the techniques ability to capture essential information about the actual mask under optimal conditions.

A Recall of 0.89 is observed for the top best prediction, this showcased the SHAP techniques ability to capture the majority of critical areas.

This was indicating that SHAP under optimal conditions are able to achieve an near optimal balance between Recall and Precision, focusing on avoiding false positives, and capturing significant true regions especially important in critical domains such as healthcare for accurate diagnosis classification.

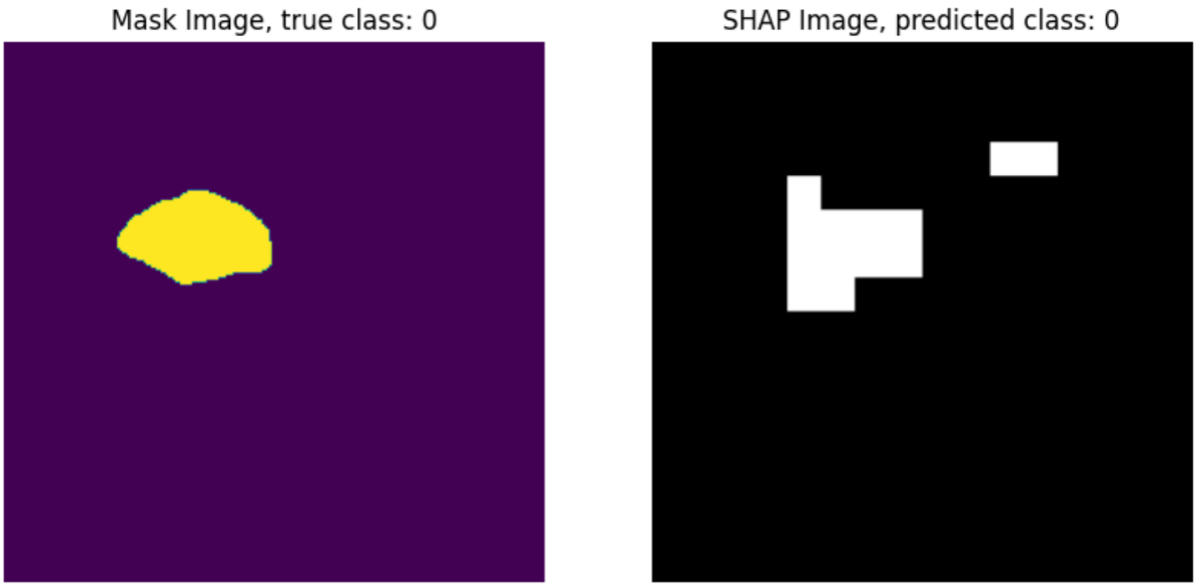


Figure 5.1: On the left, the Actual mask (ground truth) is displayed in yellow. On the right, the SHAP predicted mask is displayed in white. This displays a correct classification of class 0 (benign) and its corresponding predicted mask for class 0 (benign)

In Figures 5.2 and 5.4 the output of the SHAP technique is visualized. Here the corresponding SHAP values is shown on a scale, where the blue color represents negative contribution to the corresponding prediction, the color red represents positive contribution to the corresponding prediction. The leftmost image is the actual raw image that the explanation instance is based of, as well as the different classes (0, 1 and 2). The outputs displays the positive and negative contribution to the corresponding prediction highlighted in red and blue respectively, given in SHAP values. For the unique classes one can see what areas the model highlights to contribute to the model's decision making process for the unique classes respectively.

In Figures 5.1 and 5.3 the ground truth region displayed in yellow is compared to the regions of interest given by the SHAP displayed in white over a fixed threshold of top 20% SHAP values.

In figure 5.2 the model correctly classified the patient as class 0 (benign). Figure 5.2

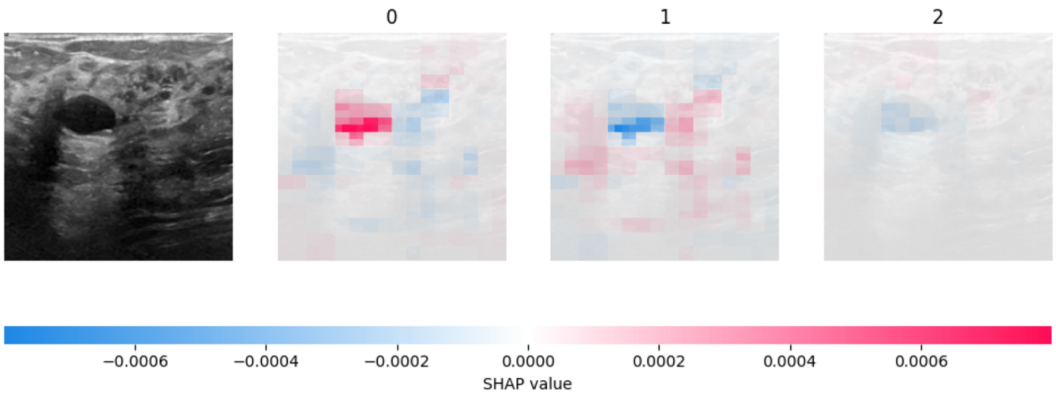


Figure 5.2: Output of the SHAP technique, where class 0, 1 and 2 is displayed and its corresponding SHAP values, Red SHAP values highlight positive contribution to the respective class, blue values corresponds to negative contribution to the corresponding class. This instance is classified correctly by the underlying model.

displays the predicted mask from SHAP against the actual mask of the patient. In contrast, in Figures 5.3 and 5.4, a scenario where the model failed to correctly classify the patient, predicting the patient as class 2 (normal), when the correct class was 0 (benign) is shown. This misclassification highlighted the challenges in applying SHAP for complex models where features may lead to explainability errors. Figure 5.3 displays the actual mask of the misclassified instance and the SHAP output. From figure 5.3 it was observed that the technique highlights insignificant regions as significant.

5.2.2 LIME Analysis

Table 5.5: LIME Evaluation Metrics for the EfficientNetV2B2 model on Breast Cancer Images. The Table Includes Metrics IoU, Precision, Recall, and F1-score, where all metrics' Average, Median, and Standard Deviation is displayed on the left except for F1-score which displays only the Average.

Metric	IOU	PRECISION	RECALL	F1-SCORE (Average)
LIME				
Average	0.11	0.16	0.31	0.21
Median	0.09	0.13	0.21	
Standard deviation	0.11	0.17	0.28	

In this chapter, exploration of the performance metrics of the LIME technique was done, as summarized in Table 5.5.

The LIME evaluation metrics shown in Table 5.5, shows an average IoU of 0.11. This indicated a relatively poor overlap between the regions of interest by the LIME technique compared to the ground truth mask. This means that 11% of the area the model believed

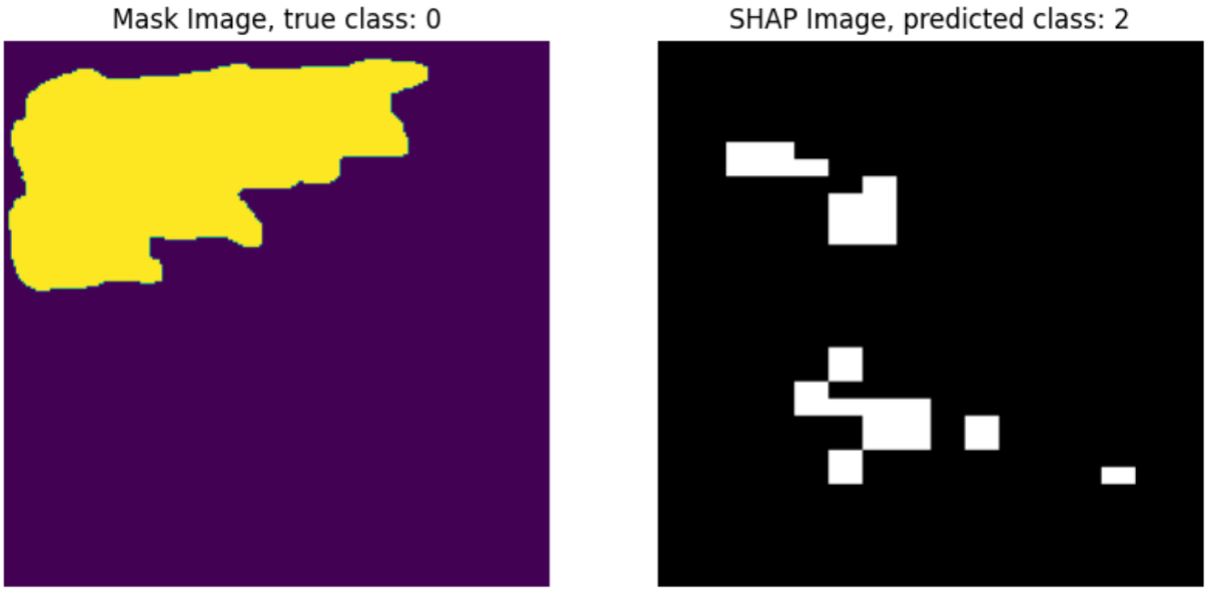


Figure 5.3: On the left, the Actual mask (ground truth) is displayed in yellow. On the right, the SHAP predicted mask is displayed in white. This displays a wrong classification of class 0 (Benign) and its corresponding predicted mask for class 2 (Normal)

was relevant overlap with the actual ground truth. The technique did not classify the majority of actual areas of interest as significant.

The average Precision is observed at 0.16, indicating a significant amount of false positives, where the technique only captured 16% of the ground truth displayed by correct mask. This means that the technique was classifying regions not belonging to the ground truth as important for model decision.

The Recall's average value is 0.31, which means that the technique captured 31% of the actual significant regions. The relatively low Recall means that there is a high number of false negatives, meaning that the technique classified important regions as not important for its decision making process.

An average F1 score of 0.21 is displayed in Table 5.5. This underscores the overall effectiveness of LIME, showcasing poor performance on the full test set.

Standard deviation was included in Table 5.5 to provide a measurement of how much the metrics changed over the entire test set. This metric helped indicate to what extent the technique fluctuates around the mean value. A standard deviation of 0.11 is shown for IoU, indicating a moderate fluctuation of IoU score over the test set. This means that the overall consistency of IoU was somewhat vague, which means that the overlapping areas between the LIME output and ground truth varied depending on the respective image. The standard deviation for Precision is observed at 0.17 with an median of 0.13. This displays a relatively high standard deviation compared to median and average indicating variability in the technique and model's ability to identify true positives, leading to varying predictions.The average standard deviation for recall is observed at 0.28 with a median of 0.21, which displays poor overall performance of the models ability to capture all relevant

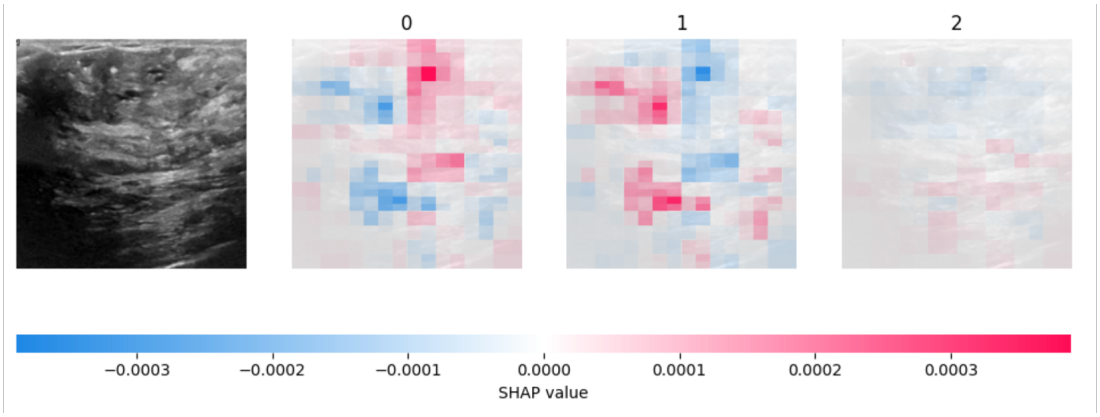


Figure 5.4: Output of the SHAP technique, where class 0, 1 and 2 is displayed and its corresponding SHAP values, Red SHAP values highlight positive contribution to the respective class, blue values corresponds too negative contribution to the corresponding class. This instance is classified wrong by the underlying model.

regions of interest when compared to the ground truth, due to the high standard deviation. It had potential to discover the relevant regions of interests under optimal conditions for certain images.

The top performance metrics for LIME is displayed in Table 5.6, where an IoU of 0.43 is showcased. This displays the technique's ability to capture relevant areas compared to ground truth, where the model captured 43% of overlapping areas between LIME output and actual mask.

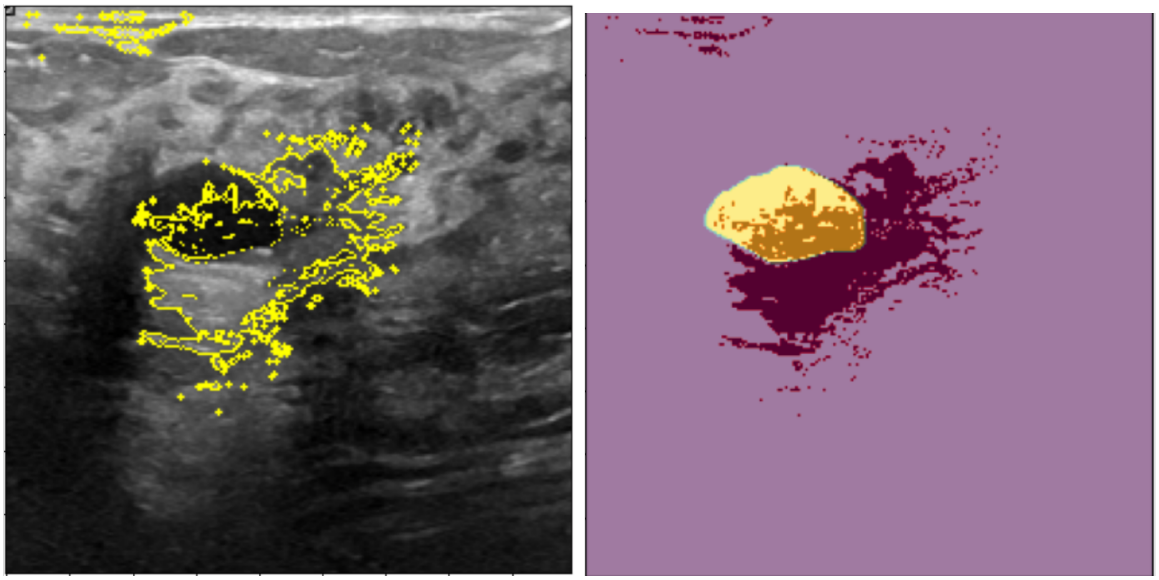
Table 5.6: The top LIME Evaluation Metric for the EfficientNetV2B2 model on Breast Cancer Images. The Table Includes Metrics IoU, Precision, Recall and F1-score

Metric	IOU	PRECISION	RECALL
LIME			
Top 1 evaluation metrics	0.46	0.75	1.0

A precision of 0.75 for the top output in Table 5.6 indicate that 75% of the region of significance for the model's decision-making process were inside the area marked by the ground truth mask, meaning that there were some false positives in the LIME output for the top output.

A recall of 1.0 is displayed for the top LIME output. This displays that there were no false negatives in the LIME output, indicating that the technique was able to identify all areas of interest determined by the ground truth mask.

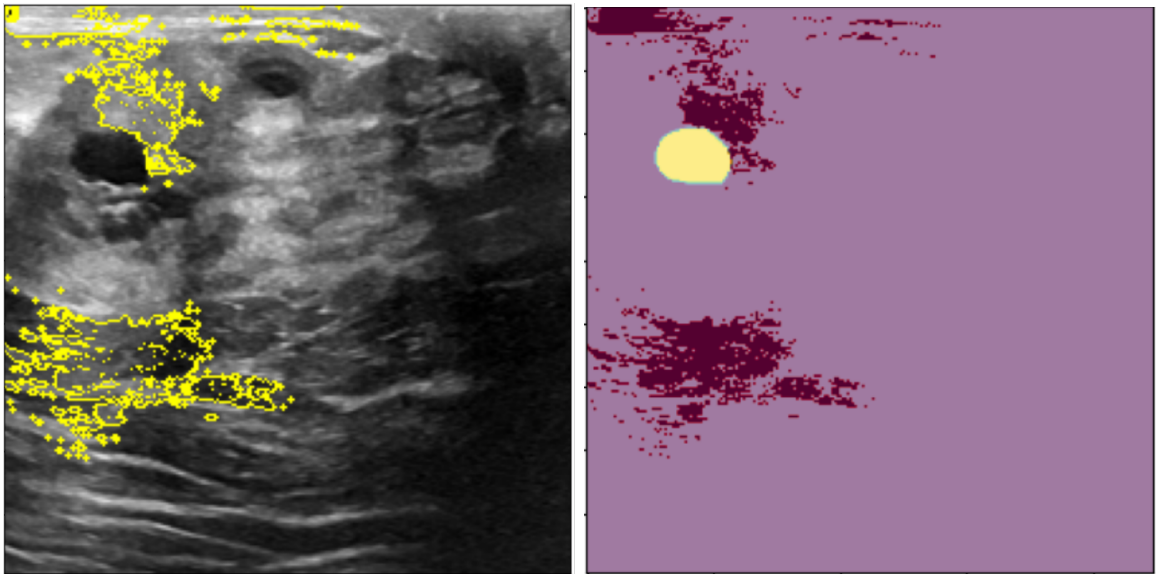
In Figure 5.5a and 5.6a LIME output is displayed of two randomly chosen patients. The regions highlighted in yellow is where the technique show where it based the decision-making process of the model's predictions of off. The yellow region which was the LIME output is displayed on top of the actual raw image of the lesion, displaying which parts of the image the technique highlighted. Figures 5.5b and 5.6b displays the same marked regions, including both the actual mask and the areas of interest marked by the LIME technique. Where the marked yellow region is the ground truth and the red regions are



(a) LIME output region of interest, highlighted with yellow, with Breast Lesion in the background.

(b) LIME output highlighted with the colour red displayed with actual mask highlighted in yellow

Figure 5.5: LIME output with both original image (a) and actual mask (b)



(a) LIME output region of interest, highlighted with yellow, with Breast Lesion in the background.

(b) LIME output highlighted with the colour red displayed with actual mask highlighted in yellow

Figure 5.6: LIME output with both original image (a) and actual mask (b)

the output from the LIME technique.

Chapter 6

Discussion

In this chapter, results obtained from the analysis are discussed and interpreted in the context of the research questions and the thesis' objectives. Key findings are discussed with respect to their implications.

6.1 Analysis of Findings

In this chapter the following XAI techniques SHAP and LIME results was discussed and compared. The discussion focused on the performance metrics of SHAP and LIME in the medical domain, highlighting the implications of their difference in performance across the metrics. It was observed that SHAP performed better overall than LIME on breast cancer data. SHAP is based on Shapley values from game theory, which assures that the contribution of each feature was fairly allocated. By considering all combination of features SHAP ensured that the attributions were more accurate and robust than LIME. Since SHAP considers all possible combinations of features, SHAP is able to capture complex interactions between features. LIME provide local explanations by perturbing data and observing changes in predictions. This can introduce noise by creating unrealistic data points, which may lead to less accurate explanations, in contrast, SHAP's method of using actual data combinations rather than perturbations resulted in a more consistent explanation over the entire dataset without unnecessary noise. These points captures the essence in why SHAP was better than LIME when dealing with complex data such as medical images, however both showed distinct strengths and weaknesses [54] [16].

6.1.1 SHAP vs. LIME

The performance metrics showed distinct strengths and weaknesses when comparing SHAP and LIME. The metrics suggested that SHAP generally achieved better precision than LIME, but over all remained moderately low. SHAP's precision was shown at an average of 0.60, which was significantly higher than LIME's 0.16. This indicated that SHAP is more reliable in identifying true positive instances with low false positives, making SHAP better at actually marking the true region which is crucial in critical sections like healthcare,

where the cost of false positives can be high [70].

Furthermore, it is observed that LIME performed slightly better than SHAP in the Recall metric, where LIME achieved an average of 0.31, which was higher than SHAP's average at 0.20. This indicated LIME's capability in capturing relevant areas, with the cost of a higher rate of false positives. LIME's higher recall might be more desirable and preferred over SHAP's higher precision in certain situations such as early diagnosis in the healthcare domain.

The contrast between the two techniques became notable when comparing F1-scores for both, which provides a balance between both precision and recall. SHAP achieved an average F1-score of 0.30 and LIME was observed at 0.21. This indicated that SHAP's slightly higher F1-score, offered a better balance over the entire dataset than LIME.

When considering the average IoU, SHAP provided a slightly higher value at 0.14 compared to LIME's average IoU at 0.11. The IoU metric provided a measurement of how much each output overlaps the ground truth. This metric quantified the techniques ability to represent the precise locations of the infiltrated region. The highest average of the two techniques were SHAP with 0.14, this indicated a poor localization of region which could be due to several reasons such as complexity of data.

When comparing the IoU with the Precision for SHAP, this indicated that SHAP was often able to find the tumor area, but struggled to cover the entirety of the area, rather giving indications of where the infiltrated region potentially could be.

Furthermore LIME provided a generally low precision and IoU but a slightly higher Recall than SHAP, which indicated that LIME may not localize the ground truth as precisely as SHAP, but rather indicated the presence of infiltrated areas broadly. The slightly higher Recall potentially indicated that LIME was better at ensuring that fewer relevant areas were missed.

The best instance for SHAP in Table 5.4 recieved an IoU of 0.43, with a precision of 1.0 and recall of 0.89. This indicates that the technique for the instance was able to achieve perfect precision, meaning no false positives. The technique's areas of importance were all within the region of interest. This is important in fields such as medical imaging where minimizing false positives is critical [70]. A strong recall of 0.89 suggested that SHAP identified 89% of all relevant instances within the data without missing many true positives. An IoU of 0.43 was substantial, however a higher IoU indicates more overlap between the XAI output and ground truth, where the IoU means that 43% of the SHAP output is overlapping with the ground truth. This showed that SHAP can approximate the area of interest, but not capture the entirety of the area. With all metrics considered, SHAP was able to identify highly likely correct areas, however struggled to capture the entirety under ideal conditions.

Furthermore the best metrics for a single instance for LIME provided a slightly higher IoU at 0.46, which showed the methods ability to capture relevant areas. An IoU of 0.46 implied that 46% of the LIME output lied within the ground truth. A Precision of 0.75 and a Recall of 1.0 was shown. When compared to SHAP it was noticed that LIME is able

to perform better when considering minimizing the risk of not including True Positives indicated by the perfect Recall for the optimal instance. This high recall is especially crucial in critical fields such as cancer detection, where failing to capture a true positive could lead to detrimental outcomes. A technique's ability to identify all actual positives can be seen as more important than being able to exclude all negatives in the respective domain. In scenarios where this is important, LIME can be a valuable tool for capturing true positives under ideal conditions, reducing the risk of overlooking critical information.

To compare the metrics observed in this study, previous research [71], where several XAI techniques were employed and measured in the domain of medical images. The IoU of SHAP and LIME were 0.06 and 0.1 respectively. This indicates that while the results presented in this thesis showed similarity, the localization capability of both SHAP and LIME remains a challenge, aligning with earlier research. This posed difficulty of achieving high scores in several metrics when using complex datasets such as medical imaging. In previous research, the XAI techniques described failed to meet medical clinical requirements and standards [71].

6.2 Methodological Insights

In this chapter, insights in the applied methods are discussed, highlighting the techniques' abilities to visualize important features aiming to aid in development of AI models and healthcare decision-making. The chapter also contains discussion about the challenges and limitations of applied XAI techniques.

6.2.1 Advantages of Applied Methods

Both SHAP and LIME offered a method to enhance transparency of AI model's decisions. This is especially crucial in departments such as healthcare, where understanding the process behind diagnosis classification and detection of cancer is important. By breaking down each prediction to show the impact of each region of pixel to the respective prediction is an important element in understanding the model's decision-making process. Therefore both techniques offer a great explainability measurement. Both have a hierarchical process of selecting the top chosen number positive or negative features to the respective prediction, which offered a robust method of explainability.

Both techniques offered good visual outputs which are significant advantages in domains such as medical imaging, since it is comparable to what is actually being done by medical professionals, which is visual inspection of images from medical tools, as well as other techniques. SHAP and LIME offer visualizations that translate complex inner workings of a model to visible boundaries of regions of importance. This makes humans able to visually inspect the techniques output and quantify its correctness to further develop and trust models. These visualizations also offer more intuitive assessment of which regions are most influential aiding to better understand the model's behavior.

SHAP offers a set of values that determine the contribution of each region of pixel

to the prediction by leveraging game theory, this to provide consistent explanations. By leveraging this method, each region of pixel is fairly allocated, which can strengthen the reliability of the interpretations [54].

LIME's model agnostic approach allows it to be a versatile technique, being able to be used on any model. This versatility can be seen as important in fields where various models are employed [16].

With clear explanations of model outputs, XAI can aid in increasing confidence among medical professionals such as radiologists when leveraging AI. With increased confidence of AI decisions, professionals are potentially able to detect diagnosis earlier, leading to higher survival rates in the healthcare domain [72] [73].

The benefits from leveraging XAI are not limited to professionals in the respective domain, but also developers who can improve models by understanding which features are underperforming or overperforming in a models decision-making process, leading to a broader understanding and further refinement of models.

6.2.2 Disadvantages and Limitations

A limitation when considering both SHAP and LIME was their computational complexity, particularly when dealing with complex models and large datasets. SHAP encountered computational complexity when calculating the contribution for each feature over all possible combinations [12]. Furthermore, LIME encountered computational complexity when generating and test with perturbed samples of original instances, which is needed for all instances being explained [15]. These perturbed samples of the original image can face challenges when working with complex medical images where LIME might not always generate meaningful or realistic perturbed samples, potentially leading to explanations that are oversimplified or misleading as discussed in the paper [11]. This limitation is crucial in healthcare, where accuracy and reliability of interpretations are important.

Both techniques SHAP and LIME can suffer from inconsistency in their explanations, where similar cases might receive varying explanations, which could be due to model-specific settings or distribution of data.

LIME's ability to be model-agnostic allows it to function with any model, however this ability can encounter potential issues because the technique is not specifically tailored to a specific type of model, potentially leading to less precise explanations for complex models [15].

Finally, the metrics used to evaluate the output from the techniques might not capture the relevance of the technique's outputs compared to the actual problem at hand.

6.3 Discussion of Research Questions

In this chapter the results are discussed with regards to the research questions for this thesis. These questions aim to guide the structure and focus of this thesis.

6.3.1 Discussion of Research Question 1 (RQ1)

To answer RQ1, several metrics were introduced, namely IoU, Precision, Recall and F1-score. An observed IoU of 0.11, as well as Precision of 0.16 for the LIME technique were measured. These metrics suggested that LIME struggled to accurately overlap the predicted region with the actual tumor region, as well as poor performance in identification of true positives without including false positives. The Recall for LIME was measured at 0.31, which reflected that LIME was able to capture a reasonable number of actual positives, but at the cost of including some false negatives. The corresponding average F1-score for LIME was observed at 0.21, which reflected an overall poor performance when considering the balance between precision and recall, which is seen as crucial for medical imaging applications, where minimizing false negatives and identifying true positives is vital. The best individual LIME metrics showed more promising results with an IoU of 0.46, precision of 0.75 and a perfect recall of 1.0. While these results indicated that LIME performed well under optimal conditions, the variability showed by standard deviation, median and average values indicate the need for further advancements in the implementation of LIME.

Furthermore SHAP received a slightly higher IoU and significantly higher precision at 0.14 and 0.60 respectively compared to LIME. This indicates that SHAP inherited the ability to more effectively identify true positives. The recall for SHAP was observed at 0.20, which was lower than LIME's, which indicated that SHAP missed a lot of true positives in its output. SHAP received an F1-score of 0.30, which was better than LIME, but still indicated poor effectiveness. The top SHAP metric for a single instance received an IoU of 0.43, perfect precision and recall of 0.89. These values suggest that SHAP under ideal conditions, can provide highly accurate segmentations. Similar to LIME, the variability in the overall metrics showed that optimal performance was not achieveable for a multitude of input images. The variability in performances across the metrics and instances for both techniques underscores the challenge of leveraging XAI in healthcare as stated in [12]. The variability can complicate deployment of these techniques, especially in critical sectors like healthcare.

Inconsistent performance suggest that careful selection and tuning of these models is needed to fit the specific problem at hand. The evaluation also shows the varying results of both techniques, underscoring their limited usability in segmentation of tumor regions whilst classifying. Building on the XAI techniques' impact on early diagnosis, the paper [74], used a light CNN model with LIME and SHAP as XAI techniques for enhancing predictions in kidney abnormalities. The paper reported an overall test accuracy of 99% when distinguishing between cysts, stones and tumors. The model's effectiveness was enhanced by XAI techniques. SHAP and LIME were implemented to highlight the influence of specific pixels in the image of specific predictions. The ability to visualize the impact of individual features aligns with the need for clarity and precision in medical diagnostics. The results were further shown to medical professionals who verified the results, which confirmed the correctness of the employed XAI techniques. These findings are important when discussing RQ1, and showed promising results for employing XAI techniques to efficiently

segmenting correct regions in medical imaging. Furthermore, the effectiveness of SHAP and LIME is discussed in the paper [75], where it was found that SHAP was generally better at providing more accurate explanations of the models predictions than LIME. SHAP also highlighted more correct regions without much noise such as false positives, which was verified by medical professionals. The effectiveness of the various XAI techniques was similar to what was observed in this thesis, where SHAP had a significantly higher precision than LIME, underscoring SHAP's ability to correctly identify correct regions.

6.3.2 Discussion of Research Question 2 (RQ2)

XAI's ability to provide transparent and interpretable decision-making can become important especially in early disease diagnosis, where accurate decisions can significantly affect patient outcomes.

Many deep learning models act like black-boxes, meaning their inner workings are unseen and not able to be understood by humans. These provide little insight into how decision-making for respective predictions are made. This limitation is potentially critical in complex fields like medical imaging where understanding the reasoning behind medical decisions is crucial for trust, medical relevance and acceptance. Previous studies discussed in Chapter 2, have shown that XAI techniques can bridge this gap by offering visualizations and insight into the inner workings of complex models, potentially enhancing the predictive performance of machine learning driven tools [16]. Previous research has also shown that visualization can aid researchers in identifying wrong reasoning in classification problems that were previously overlooked [76]. This underscores the point of importance of visualization tools in decision-making, and further strengthens the point of enhancing predictive performance by leveraging XAI tools.

By integrating XAI techniques, namely SHAP and LIME, medical professionals have the possibility to be informed on what diagnostics were made and why these decisions were made. In tumor classification, these techniques will highlight the feature that was most influential in its respective classification. The detailed explanation can enhance decision-making, and being able to align the output of techniques to professionals in the field potentially leading to more precise outputs over all.

Despite the advantages of XAI application, they also introduce challenges. Inconsistency in results align with previous research and these inconsistencies lead to variability in trust especially when trying to meet clinical standards [71]. By addressing these variabilites through visualization from XAI techniques, and including feedback from clinical outcomes could further improve reliability of such systems.

The inconsistency in the performance metrics such as varying IoU and recall across different instances reflect how XAI techniques might influence clinical decision-making. Inconsistent outputs lead to low metrics which reflects the degree of trust and reliance of these techniques. Despite challenges, the ability to provide visual interpretations of models can be considered important by assisting medical professionals by quantifying what feature or area of an image influence predictions.

Early diagnosis of diseases such as cancer has the possibility to reduce mortality as mentioned earlier. XAI techniques like SHAP and LIME are increasingly vital due to their potential of enhancing predictive performance and improve decision-making processes by understanding the reasoning behind it. This could possibly improve decision-making by aligning XAI outputs with clinical prognosis by professionals. In special cases where either benign or malignant tumor exists, SHAP and LIME can provide detailed reasoning and explanations of areas influencing the predictions by visualizing the areas potentially leading to more accurate assessments and earlier detection.

Chapter 7

Conclusion

7.1 Summary of Findings

In this thesis, the implementation and efficiency of XAI techniques SHAP and LIME in medical imaging were explored, focusing on improving diagnostic accuracy of breast cancer detection and classification. Furthermore, the findings display the limitations and underperformance of the XAI techniques across all metrics evaluated being IoU, Recall, Precision and F1 score. SHAP's average Precision score was an important finding of this thesis, displaying SHAP's potential at locating the tumor in the image, however not fully the entire area. Therefore SHAP was better than LIME at accurately finding the correct areas. The low Recalls for both SHAP and LIME displayed that these techniques would not be able to be used in a real world setting based on this thesis' results, especially in healthcare domain where false negatives are critical.

The overall underperformance of the XAI techniques suggested limitations of usability in medical imaging, since it is considered a critical domain, where correctness is key and the outcome of obeying from this has potentially detrimental outcomes. This raised important questions about the use of these techniques in such critical domains, stating clear need for further development and validation of XAI implementations before integrating them to the medical workflow.

7.2 Recommendations for Future Research

Future research should focus on several key areas to enhance application of XAI techniques. As seen in previous studies the need for improving XAI frameworks in medical imaging is needed. The need of developing better models that integrate both AI and XAI without compromising on performance metrics such as IoU, Recall, Precision and F1-score is clear. A possibility to enhance efficiency could be found by leveraging multitude of XAI techniques to provide best of both worlds scenarios. Furthermore, neural symbolic learning for enhancing XAI in medical imaging is a field to dive in to, which might better fit the problem at hand. Future research should explore the use and integration of such systems

with regards to XAI, to possibly receive better results of interpretability and explainability of AI models.

Another area to dive into is investigation of integration of XAI where data from different medical image tools are leveraged and combined. Different XAI techniques could also be used, where they might be better suited for medical images. Both areas could potentially increase the robustness of interpretations by XAI techniques.

In this thesis only one model was considered as the XAI techniques input. However, testing XAI techniques with a variety of models for comparison to find the most appropriate model is important to determine the best possible approach for application of XAI techniques.

Finally the need for extensive validation of XAI techniques by medical professionals to ensure outputs is clear and makes sense, as well as meeting medical and clinical standards will ensure the implementation in real-world scenarios.

Bibliography

[1] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and the Precise4Q consortium, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 310, 2020. DOI: 10.1186/s12911-020-01332-6.

[2] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable ai: A brief survey on history, research areas, approaches and challenges," in Sep. 2019, pp. 563–574, ISBN: 978-3-030-32235-9. DOI: 10.1007/978-3-030-32236-6_51.

[3] M. Gulum, C. M. Trombley, and M. Kantardzic, "A review of explainable deep learning cancer detection models in medical imaging," *Applied Sciences*, vol. 11, p. 4573, 2021. DOI: 10.3390/APP11104573.

[4] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17–48, 2023. DOI: 10.3322/caac.21763.

[5] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 1, pp. 12–49, 2024. DOI: https://doi.org/10.3322/caac.21820.

[6] G. Khuwaja and A. Abu-Rezq, "Bimodal breast cancer classification system," *Pattern Analysis & Applications*, vol. 7, pp. 235–242, 2004. DOI: 10.1007/BF02683990.

[7] H. Sung, J. Ferlay, R. Siegel, *et al.*, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, pp. 209–249, 2021. DOI: 10.3322/caac.21660.

[8] Y. S. Sun, Z. Zhao, Z. N. Yang, *et al.*, "Risk factors and preventions of breast cancer," *International Journal of Biological Sciences*, vol. 13, no. 11, pp. 1387–1397, 2017. DOI: 10.7150/ijbs.21635.

[9] W. A. Berg, A. I. Bandos, E. B. Mendelson, D. Lehrer, R. A. Jong, and E. D. Pisano, "Ultrasound as the Primary Screening Test for Breast Cancer: Analysis From ACRIN 6666," *JNCI: Journal of the National Cancer Institute*, vol. 108, no. 4, djv367, Dec. 2015, ISSN: 0027-8874. DOI: 10.1093/jnci/djv367.

[10] Y. Guo, Y. Hu, M. Qiao, *et al.*, "Radiomics analysis on ultrasound for prediction of biologic behavior in breast invasive ductal carcinoma," *Clinical Breast Cancer*, vol. 18, no. 3, e335–e344, 2018, ISSN: 1526-8209. DOI: 10.1016/j.clbc.2017.08.002.

[11] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to shapley values," *Artificial Intelligence*, vol. 298, p. 103 502, 2021, ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2021.103502.

[12] K. Roshan and A. Zafar, "Using kernel shap xai method to optimize the network anomaly detection model," in *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2022, pp. 74–80. DOI: 10.23919/INDIACom54597.2022.9763241.

[13] G. Zhang, Y. Shi, P. Yin, *et al.*, "A machine learning model based on ultrasound image features to assess the risk of sentinel lymph node metastasis in breast cancer patients: Applications of scikit-learn and shap," *Frontiers in Oncology*, vol. 12, 2022. DOI: 10.3389/fonc.2022.944569.

[14] R. Confalonieri, T. Weyde, T. R. Besold, and F. Moscoso del Prado Martín, "Using ontologies to enhance human understandability of global post-hoc explanations of black-box models," *Artificial Intelligence*, vol. 296, p. 103 471, 2021, ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2021.103471.

[15] J. Dieber and S. Kirrane, *Why model why? assessing the strengths and limitations of lime*, 2020. DOI: 10.48550/arXiv.2012.00093. arXiv: 2012.00093 [cs.LG].

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016. DOI: 10.48550/arXiv.1602.04938.

[17] S. Sarp, M. Kuzlu, E. Wilson, U. Cali, and O. Guler, "The enlightening role of explainable artificial intelligence in chronic wound classification," *Electronics*, vol. 10, no. 12, 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10121406.

[18] M. Rucco, G. Viticchi, and L. Falsetti, "Towards personalized diagnosis of glioblastoma in fluid-attenuated inversion recovery (flair) by topological interpretable machine learning," *Mathematics*, vol. 8, no. 5, 2020, ISSN: 2227-7390. DOI: 10.3390/math8050770.

[19] M. M. Ahsan, R. Nazim, Z. Siddique, and P. Huebner, "Detection of covid-19 patients from ct scan and chest x-ray data using modified mobilenetv2 and lime," *Healthcare*, vol. 9, no. 9, 2021, ISSN: 2227-9032. DOI: 10.3390/healthcare9091099.

[20] H. Guo, S. Wang, H. Dang, *et al.*, "Lightbtseg: A lightweight breast tumor segmentation model using ultrasound images via dual-path joint knowledge distillation," *ArXiv*, vol. abs/2311.11086, 2023. DOI: 10.48550/arXiv.2311.11086.

[21] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104 863, 2020, ISSN: 2352-3409. DOI: 10.1016/j.dib.2019.104863.

[22] M. Idrees and A. Sohail, "Explainable machine learning of the breast cancer staging for designing smart biomarker sensors," *Sensors International*, vol. 3, p. 100 202, 2022, ISSN: 2666-3511. DOI: https://doi.org/10.1016/j.sintl.2022.100202.

[23] M. Rodriguez-Sampaio, M. Rincón, S. Valladares-Rodriguez, and M. Bachiller-Mayoral, "Explainable artificial intelligence to detect breast cancer: A qualitative case-based visual interpretability approach," in *Artificial Intelligence in Neuroscience: Affective Analysis and Health Applications*, J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López, and H. Adeli, Eds., Cham: Springer International Publishing, 2022, pp. 557–566, ISBN: 978-3-031-06242-1. DOI: 10.1007/978-3-031-06242-1_55.

[24] F. Silva-Aravena, H. Núñez Delafuente, J. H. Gutiérrez-Bahamondes, and J. Morales, "A hybrid algorithm of ml and xai to prevent breast cancer: A strategy to support decision making," *Cancers*, vol. 15, no. 9, 2023, ISSN: 2072-6694. DOI: 10.3390/cancers15092443.

[25] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, *Is grad-cam explainable in medical images?* 2023. DOI: 10.48550/arXiv.2307.10506. arXiv: 2307.10506 [eess.IV].

[26] U. Veronesi, P. Boyle, A. Goldhirsch, R. Orecchia, and G. Viale, "Breast cancer," *The Lancet*, vol. 365, pp. 1727–1741, 2005. DOI: 10.1016/S0140-6736(05)66546-4.

[27] T. Mahmood, J. Li, Y. Pei, F. Akhtar, A. Imran, and K. U. Rehman, "A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities," *IEEE Access*, vol. 8, pp. 165 779–165 809, 2020. DOI: 10.1109/ACCESS.2020.3021343.

[28] O. Ginsburg, F. Bray, M. P. Coleman, *et al.*, "The global burden of women's cancers: A grand challenge in global health," *The Lancet*, vol. 389, no. 10071, pp. 847–860, 2017, ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(16)31392-7.

[29] World Health Organization, *Breast cancer*, https://www.who.int/news-room/fact-sheets/detail/breast-cancer, Retrieved April 10, 2024, 2022.

[30] D. Crosby, S. Bhatia, K. M. Brindle, *et al.*, "Early detection of cancer," *Science*, vol. 375, no. 6586, eaay9040, 2022. DOI: 10.1126/science.aay9040.

[31] A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney, "Solid breast nodules: Use of sonography to distinguish between benign and malignant lesions.," *Radiology*, vol. 196, no. 1, pp. 123–134, 1995, PMID: 7784555. DOI: 10.1148/radiology.196.1.7784555.

[32] G. Xue-ha, "Ultrasonographic characteristics of benign and malignant calcification of breast," *Chinese Journal of Medical Ultrasound*, 2007. [Online]. Available: https://api.semanticscholar.org/CorpusID:76635232.

[33] National Cancer Institute, *How cancer is diagnosed*, https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis, Last updated on January 17, 2023, Jan. 2023.

[34] A. Redman, S. Lowes, and A. Leaver, "Imaging techniques in breast cancer," *Surgery (oxford)*, vol. 34, pp. 8–18, 2016. DOI: 10.1016/J.MPSUR.2015.10.004.

[35] L. Wang, "Early diagnosis of breast cancer," *Sensors (Basel, Switzerland)*, vol. 17, no. 7, p. 1572, 2017. DOI: 10.3390/s17071572.

[36] B. O. Anderson, R. Shyyan, A. Eniu, *et al.*, "Breast cancer in limited-resource countries: An overview of the breast health global initiative 2005 guidelines," *The Breast Journal*, vol. 12, no. s1, S3–S15, 2006. DOI: https://doi.org/10.1111/j.1075-122X.2006.00199.x.

[37] National Cancer Institute, *Definition of lesion - nci dictionary of cancer terms*, https://www.cancer.gov/publications/dictionaries/cancer-terms/def/lesion, Accessed: 2023-04-18, 2023.

[38] H. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognition*, vol. 43, no. 1, pp. 299–317, 2010, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2009.05.012.

[39] National Cancer Institute, *Definition of high-grade*, https://www.cancer.gov/publications/dictionaries/cancer-terms/def/high-grade, Retrieved April 10, 2024, n.d.

[40] E. A. Asare, E. G. Grubbs, J. E. Gershenwald, F. L. Greene, and T. A. Aloia, "Setting the "stage" for surgical oncology fellows: Pierre denoix and tnm staging," *Journal of Surgical Oncology*, vol. 119, no. 7, p. 823, 2019. DOI: 10.1002/jso.25404.

[41] S. E. Singletary and J. L. Connolly, "Breast cancer staging: Working with the sixth edition of the ajcc cancer staging manual," *CA: A Cancer Journal for Clinicians*, vol. 56, no. 1, pp. 37–47, 2006. DOI: https://doi.org/10.3322/canjclin.56.1.37.

[42] G. N. Hortobagyi, "Treatment of breast cancer," *New England Journal of Medicine*, vol. 339, no. 14, pp. 974–984, 1998. DOI: 10.1056/NEJM199810013391407.

[43] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015. DOI: 10.1038/nature14539.

[44] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986. DOI: 10.1038/323533a0.

[45] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019.

[46] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Learning hierarchical categories in deep neural networks," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, Cognitive Science Society, vol. 35, 2013. [Online]. Available: https://escholarship.org/content/qt2fv5q3hn/qt2fv5q3hn.pdf.

[47] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, "Sensitivity and generalization in neural networks: An empirical study," *arXiv preprint arXiv:1802.08760*, 2018. DOI: 10.48550/arXiv.1802.08760.

[48] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, Aug. 2018, ISSN: 0360-0300. DOI: 10.1145/3236009.

[49] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Incorporating explainable artificial intelligence (xai) to aid understanding of machine learning in the healthcare domain," Oct. 2020, pp. 169–180. DOI: 10.13140/RG.2.2.24754.02246.

[50] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, ISSN: 0001-0782. DOI: 10.1145/3065386.

[52] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 9, 2016. DOI: 10.1186/s40537-016-0043-6.

[53] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 18–24 Jul 2021, pp. 10 096–10 106. DOI: 10.48550/arXiv.2104.00298.

[54] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, 2017. DOI: 10.48550/arXiv.1705.07874. arXiv: 1705.07874.

[55] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018, ISSN: 1542-7730. DOI: 10.1145/3236386.3241340.

[56] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018. DOI: 10.1109/ACCESS.2018.2870052.

[57] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.12.012.

[58] D. Broniatowski, *Psychological foundations of explainability and interpretability in artificial intelligence*, en, 2021-04-12 04:04:00 2021. DOI: https://doi.org/10.6028/NIST.IR.8367.

[59] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, *A survey of methods for explaining black box models*, 2018. DOI: 10.48550/arXiv.1802.01933. arXiv: 1802.01933 [cs.CY].

[60] M. Hossin and S. M.N, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining Knowledge Management Process*, vol. 5, pp. 01–11, Mar. 2015. DOI: 10.5121/ijdkp.2015.5201.

[61] P. Jaccard, "Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines.," *Bulletin de la Societe Vaudoise des Sciences Naturelles*, vol. 37, pp. 241–72, Jan. 1901. DOI: 10.5169/seals-266440.

[62] L. da F. Costa, *Further generalizations of the jaccard index*, 2021. DOI: 10.48550/arXiv.2110.09619. arXiv: 2110.09619 [cs.LG].

[63] Martín Abadi, Ashish Agarwal, Paul Barham, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

[64] F. Chollet *et al.*, *Keras*, https://keras.io, 2015.

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[67] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. DOI: 10.48550/arXiv.1905.11946. arXiv: 1905.11946 [cs.LG].

[68] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980 [cs.LG].

[69] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2.

[70] "Factors associated with rates of false-positive and false-negative results from digital mammography screening: An analysis of registry data," *Annals of Internal Medicine*, vol. 164, no. 4, pp. 226–235, 2016, PMID: 26756902. DOI: 10.7326/M15-0971.

[71] W. Jin, X. Li, and G. Hamarneh, "Evaluating explainable ai on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements?" *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 11 945–11 953, Jun. 2022. DOI: 10.1609/aaai.v36i11.21452. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/21452.

[72] N. Hawkes, "Cancer survival data emphasise importance of early diagnosis," *BMJ*, vol. 364, 2019, ISSN: 0959-8138. DOI: 10.1136/bmj.l408.

[73] F. Sadoughi, Z. Kazemy, F. Hamedan, L. Owji, M. Rahmanikatigari, and T. T. Azadboni, "Artificial intelligence methods for the diagnosis of breast cancer by image processing: A review," *Breast Cancer: Targets and Therapy*, vol. 10, pp. 219–230, 2018. DOI: 10.2147/BCTT.S175311.

[74] M. Bhandari, P. Yogarajah, M. S. Kavitha, and J. Condell, "Exploring the capabilities of a lightweight cnn model in accurately identifying renal abnormalities: Cysts, stones, and tumors, using lime and shap," *Applied Sciences*, vol. 13, no. 5, 2023, ISSN: 2076-3417. DOI: 10.3390/app13053125.

[75] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, "Explainable ai for retinoblastoma diagnosis: Interpreting deep learning models with lime and shap," *Diagnostics*, vol. 13, no. 11, 2023, ISSN: 2075-4418. DOI: 10.3390/diagnostics13111932.

[76] S. Lapuschkin, S. Wäldchen, A. Binder, *et al.*, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019. DOI: 10.1038/s41467-019-08987-4.

Appendix A

AI statement

In this thesis, the use of Chat-GPT, a language model by OpenAI has been used. Throughout the work on this thesis, current regulations from Faculty of Science and Technology (REALTEK) has been followed. The regulations can be found at https://www.nmbu.no/en/faculties/faculty-science-and-technology/kunstig-intelligens-ved-realtek. ChatGPT was used throughout this thesis for the design of latex tables and aiding in the formatting of such tables. Example prompts used for this include giving latex tables to ChatGPT, "make table header on top of table", "add a new column with X numbers" and "collapse column in Table X". ChatGPT was also used for suggestions of better flow of already written text and aiding in rewriting sentences based on already written text to clarify points. Careful examination of rewritten text were done as well as insuring that no extra information was added. Example prompts are giving ChatGPT a sentence, or paragraph of already written text, and making it rewrite to clarify more efficiently, while ensuring flow. This was carefully examined and checked to make sure the output provided by ChatGPT was correct without adding new information.

ChatGPT was also used when developing code, especially aiding by debugging helping with debugging existing code. Example prompt include providing error messages with the code.

Lastly, ChatGPT was also used to summarize existing literature. This was done to get a quick overview of existing literature, finding out if the literature provided actually was worth diving into, based on the context of this thesis. Every claim made by ChatGPT was carefully fact checked to ensure the following of guidelines and regulations. Example prompt were providing already found literature as PDF files, and asking ChatGPT to summarize the contents.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway