EXPLAINABLE AI FOR BREAST CANCER DETECTION IN MAMMOGRAPHY

L.L.M. Wickremesinghe

229406T

Degree of Master of Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

January 2025

EXPLAINABLE AI FOR BREAST CANCER DETECTION IN MAMMOGRAPHY

L.L.M. Wickremesinghe

229406T

Dissertation submitted in partial fulfillment of the requirements for the degree
Degree of Master of Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

January 2025

DECLARATION

I declare that this is my own work and this Dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 30-06-2025

The supervisor should certify the Dissertation with the following declaration.

The above candidate has carried out research for the Degree of Master of Computer Science Dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Thanuja D. Ambegoda

Signature of the Supervisor: Date: 30-06-2025

ACKNOWLEDGEMENT

I am grateful to Dr. Thanuja D. Ambegoda for his invaluable assistance in identifying an engaging research topic and for his unwavering support and encouragement. His guidance was pivotal in setting goals and enhancing my study engagement. I extend my heartfelt thanks to the Department of Computer Science and Engineering at the University of Moratuwa for their assistance in overcoming challenges. Lastly, I want to express deep appreciation to my family for their steadfast support throughout this journey.

ABSTRACT

Breast cancer remains a significant global health concern among women. This research introduces an explainable AI-assisted breast cancer detection system aimed at improving both the accuracy and interpretability of mammogram-based diagnoses. The study utilizes high-quality mammographic datasets, CBIS-DDSM and the RSNA Screening Mammography dataset, to train and validate the models.

The system uses two powerful deep learning models: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The InceptionResNetV2 CNN achieved an accuracy of 92%, while the ViT model reached 96% accuracy by effectively focusing on important regions in the mammogram images. To make the system more transparent, several Explainable AI (XAI) methods were applied, including Grad-CAM, SIDU, Attention Maps, and Ablation-CAM. Among these, SIDU provided the clearest and most accurate visual explanations, which are valuable for medical decision-making.

To further improve the reliability and clinical value of the system, this study introduces a Dual-Stage Ensemble Diagnosis and Decision Fusion Framework. This approach combines the diagnostic strengths of both models to deliver a more confident and balanced final decision, supported by detailed visual explanations. The platform consist with a user-friendly web application that allows doctors and patients to easily upload mammogram images and receive AI-based predictions with clear and interpretable outputs. This research helps advance the development of trustworthy AI tools for breast cancer detection in real clinical settings.

Keywords: Explainable AI(XAI), Breast Cancer, Convolutional Neural Networks(CNN), Vision Transformers (ViT), Mammogram, Medical Imaging

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation and Background

Cancer is a multifaceted disease influenced by a variety of genetic, environmental, and lifestyle factors, which complicates both its diagnosis and treatment planning. The International Agency for Research on Cancer (IARC) reported that, globally, around 18.1 million new cancer cases and 9.6 million cancer-related deaths occurred in 2018 [1]. Among these, breast cancer has seen a consistent rise in incidence over the past few decades and remains a major cause of mortality among women. As illustrated in Figure 1.1, breast cancer is the most commonly diagnosed cancer among women across all Human Development Index (HDI) categories [2]. With approximately 2.1 million new cases and 627,000 deaths reported in 2018 alone, breast cancer continues to be a major public health concern worldwide [3]. Its development is not attributed to a single cause but rather a combination of various risk factors, including genetic mutations (such as those in the BRCA1 and BRCA2 genes), advancing age, family history, obesity, alcohol consumption, sedentary lifestyle, and extended exposure to hormones [4].

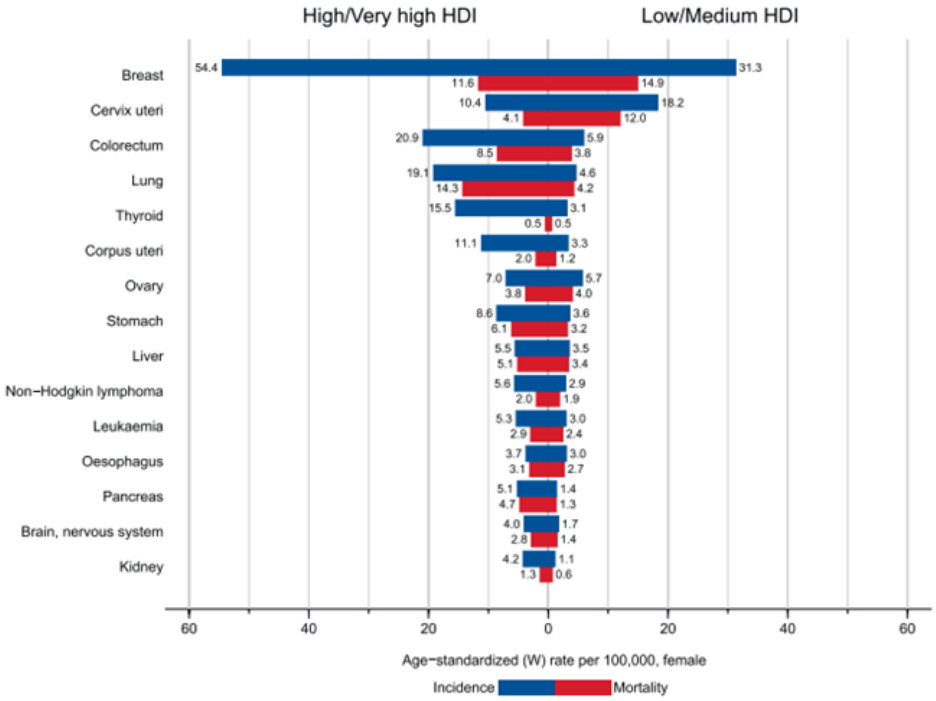

Fig. 1.1: Bar chart presents comparing age-standardized incidence and mortality rates (per 100,000 person-years) for the 15 most common cancer types in 2018, contrasting countries with high or very high Human Development Index (HDI) against those with low or medium HDI, separately for women (top) and men (bottom).

The American Cancer Society categorizes breast cancer into several types based on the specific cells where it originates, primarily ductal carcinoma starting in breast ducts and lobular carcinoma in milk-producing glands. It distinguishes between in situ cancers, restricted to ducts or lobules, and invasive cancers that expand into surrounding tissue. Invasive Ductal Carcinoma (IDC) is the predominant form, constituting roughly 80% of breast cancer cases, while Invasive Lobular Carcinoma (ILC) represents 10% to 15% of cases. Recognizing these differences is critical for customizing treatment strategies that can enhance patient survival rates [5].

Early identification of breast cancer significantly improves the chances of successful treatment and increases survival rates. Recognizing symptoms like new lumps in the breast or armpit, changes in breast size, shape, or texture, skin dimpling resembling an orange peel, and nipple abnormalities like discharge or inversion can prompt timely medical evaluation. These symptoms may indicate various types and stages of breast cancer, including rare forms like inflammatory breast cancer and Paget's disease. Prompt recognition and diagnosis allow for early intervention, which can significantly impact treatment efficacy and patient prognosis. Therefore, staying vigilant about breast health and seeking medical attention for any concerning changes are essential practices in reducing the impact of breast cancer [6]. Identifying cancer in its initial stages greatly enhances treatment success and survival outcomes, as symptoms are often absent until the disease becomes advanced.

Mammography is one of the most accessible and cost-effective method for initial breast cancer screening and diagnosis. It has significantly lowered mortality rates by enabling early cancer identification. Mammography is a low-dose X-ray procedure where the breast is gently flattened between two plates, capturing two distinct images of each breast [7]. The mammograms are carefully examined for abnormalities in size, shape, and contrast, with a focus on distinguishing between benign masses, fibroadenomas, or complex cysts, and potentially cancerous lesions. Additional views or diagnostic tests like biopsies may be recommended for further evaluation if suspicious areas are detected. Mammograms can be used to detect both benign and malignant lumps or changes in the breast, allowing for early intervention and treatment. Women are usually recommended to have their first mammogram at around 40 years old, though individuals with a family history of breast cancer may require earlier evaluation.

According to guidelines from the American Cancer Society, mammography and physical examinations are crucial for early detection of breast cancer. While there are various other screening methods available, each differing in the clarity of breast imaging they provide, mammography refers as the most effective technique for breast cancer screening. One of the primary advantages of mammography is its cost-effectiveness, making it feasible for screening large populations. However, maintaining consistency and accuracy in interpreting mammograms, given the high volume reviewed by radiologists daily, poses challenges. Consequently, computer-aided diagnostic systems

hold promise in improving breast cancer detection and reducing associated morbidity. These systems offer the best potential for enhancing diagnostic accuracy and ensuring more effective management of the disease. A single mammogram can consist of multiple views, each generating high-resolution images that require meticulous examination. For instance, a typical screening mammogram may produce four views per breast, leading to approximately 8 to 10 images per patient study. The manual review of such extensive datasets is time-consuming and mentally taxing, increasing the risk of fatigue and potential oversights.

This is where computer-aided design (CAD) frameworks excel. They are capable of efficiently processing and analyzing vast quantities of mammography images using various ML techniques, providing valuable assistance to radiologists. Leveraging Artificial Intelligence (AI) algorithms, CAD frameworks excel in pattern recognition, systematically scanning through numerous images to identify suspicious areas and highlight potential abnormalities [8]. As a result, CAD frameworks not only save time but also enhance the precision and consistency of breast cancer diagnosis. Early detection of breast cancer can greatly increase the chances of recovery for millions of women worldwide. Hence CAD system act as a main role in the sector of diagnostic radiology and medical imaging, particularly in breast cancer detection. These frameworks have emerged as preferred tools among healthcare professionals, notably radiologists, due to their potential to interpret medical images, especially mammography scans.

CAD systems provide detailed visualizations of the patient's condition and allow doctors to quickly and accurately identify issues. Additionally, these systems contribute to making the diagnostic workflow more efficient, shortening the time required for physicians to reach a diagnosis. This can be especially useful in emergencies, where every second can make a difference. CAD has become a significant research topic in diagnostic radiology and medical imaging. CAD frameworks, powered by AI algorithms, can reduce this burden and enhance the accuracy of breast cancer diagnosis. Typically, these frameworks consist of two primary components: a computer-aided detection (CADe) subsystem and a computer-aided diagnosis (CADx) subsystem. The CADe subsystem focuses on identifying and marking potential lesions and abnormalities in mammograms, while the CADx subsystem then analyzes these findings to classify the detected abnormalities and assess their likelihood of malignancy [9].

The main stages typically involved in mammography-based CAD systems are illustrated in Figure 1.2 [10].This system begins by detecting suspicious regions, which is typically achieved using either pixel-based and region-based methods [11]. Pixel-based methods require intensive computational resources. In contrast, region-based techniques utilize segmentation to extract regions of interest (ROIs) based on morphological characteristics, offering lower computational complexity. In mammography, common signs of breast cancer include masses and microcalcifications [12]. Mass detection algorithms primarily involve identifying suspicious regions and classifying

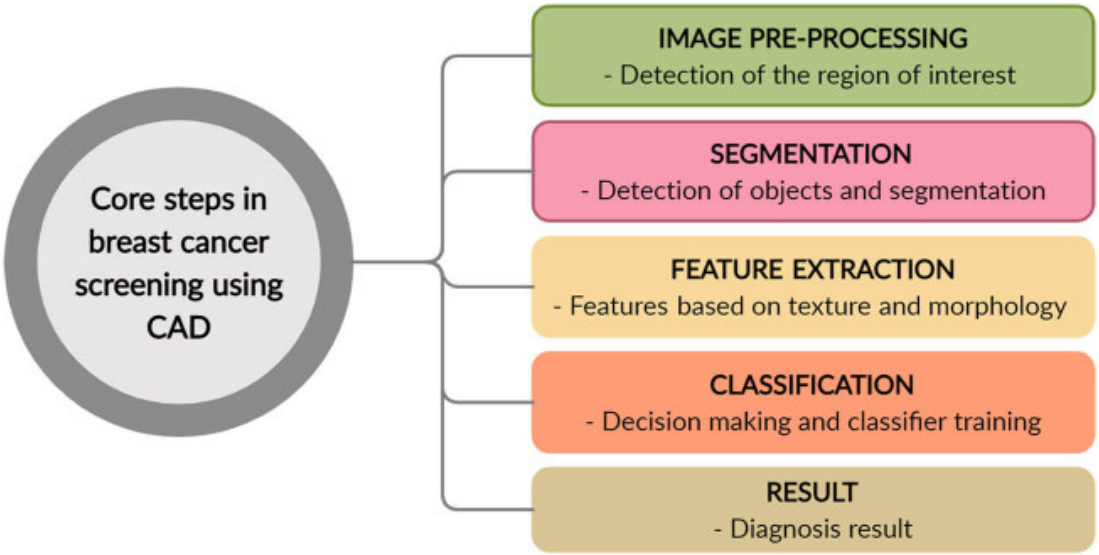

Fig. 1.2: Mammography-based CAD system Process

them from normal structures [13, 14]. The assessment typically relies on features such as shape and boundary definition, with irregular or spiculated edges often suggesting a greater risk of malignancy. Techniques such as edge orientation calculation, statistical analysis of pixel orientation maps, and wavelet transform for multi-resolution feature extraction are employed for spicule detection [15, 16]. Multiple threshold algorithms and region-based methods focusing on specific margin characteristics have also been developed for mass detection.

The next step in mass detection is to decide if the suspicious areas are masses or normal tissue. Researchers utilize various image characteristics, such as contrast, intensity, and spatial location, to develop classifiers capable of distinguishing between masses and normal tissue [17, 18].Overall, CADe systems integrate advanced algorithms and image processing techniques to enhance the accuracy and efficiency of breast cancer detection from mammograms, offering valuable support to radiologists in clinical practice. Here are a few benefits of CAD frameworks in breast cancer diagnosis:

- **Enhanced Accuracy:** CAD systems are highly effective at recognizing patterns and can spot minor abnormalities in breast tissue that may not be visible to the human eye. This enhances diagnostic accuracy and reliability, boosting the chances for early and successful treatment.

- **Consistency:** CAD frameworks provide consistent results, reducing variability in interpretations between different radiologists. This standardization ensures that all patients receive a uniform level of care, regardless of the interpreting radiologist.

- **Time Efficiency:** CAD systems can quickly process and analyze large volumes of mammography images. This efficiency helps radiologists focus on more challenging cases and handle their workload more effectively, ultimately improving overall productivity.

- **Early Detection:**CAD frameworks are adept at identifying early-stage breast lesions, such as microcalcifications and small tumors. Early detection is crucial for timely intervention, which can significantly enhance patient outcomes and survival rates.

- **Reduced Fatigue:** The systematic approach of CAD frameworks can reduce some of the cognitive load on radiologists, reducing the risk of fatigue-related errors. This helps in maintaining high diagnostic accuracy over extended periods.

- **Accessibility:** CAD systems can be deployed in various healthcare settings, including smaller clinics and remote areas. This broader accessibility helps in extending advanced diagnostic capabilities to underserved regions, improving overall healthcare equity.

While computer-aided design frameworks provide many benefits, they also have challenges and limitations. Addressing these issues is crucial for further improving their ability in breast cancer diagnosis.

- **Diverse Training Data:** For CAD frameworks to perform effectively, they must be trained on diverse and representative datasets. This ensures that the system can recognize all types of breast cancer, including rare and atypical forms. Efforts must be made to develop comprehensive datasets that cover a wide spectrum of breast cancer types and demographics.

- **Handle Complexity:** CAD systems face limitations in handling complex or subtle patterns and poor generalization across diverse datasets. These bottlenecks arise because CAD systems rely on predefined features and rule-based algorithms.This may fail to encompass the complete diversity of medical images, potentially lowering accuracy when encountering unfamiliar or varied cases.

- **False Positives and Negatives:** CAD systems may occasionally generate incorrect results, either flagging benign areas as suspicious, which can cause unneeded tests and stress, or missing actual cases of cancer. Ongoing improvement and rigorous testing of these algorithms are essential to reduce such errors and improve their reliability in diagnosis.

- **Integration into Clinical Workflow:** Incorporating CAD frameworks into clinical practice presents logistical challenges. Radiologists need training to use

these systems effectively, and healthcare institutions must address integration issues, including workflow adaptation and cost considerations.

- **Ethical and Privacy Concerns:** Integrating artificial intelligence into healthcare brings forward key ethical concerns, particularly around safeguarding patient privacy, securing sensitive data, and ensuring AI technologies are used responsibly. It is crucial to establish robust safeguards and regulations to protect patient information and ensure ethical AI deployment.

Despite these challenges, computer-aided design frameworks offer significant potential in breast cancer diagnosis. They could revolutionize the detection and classification of breast abnormalities, ultimately improving treatment outcomes and reducing breast cancer mortality rates.

Maximizing the effectiveness of CAD frameworks and addressing their existing blockers, active research and collaboration among medical professionals, researchers, and AI specialists are essential. This collaboration drives the development of more robust algorithms, facilitate better integration into clinical workflows, and lead to the creation of more comprehensive and diverse training datasets. In this setting, DL algorithms have demonstrated significant potential and widespread collaboration in image processing and data analysis. The transition from CAD systems to DL systems in medical imaging marks a significant improvement in medical diagnostic capabilities. DL particularly, CNN offers end-to-end learning, where models automatically learn hierarchical features directly from raw image data [19, 20]. This eliminates the need for manual feature engineering and allows DL models to capture more advanced patterns, and the number of hidden layers that are added to the network, make higher accuracy and better generalization across various imaging conditions.

However, DL still faces challenges, such as need of large annotated datasets, computational resource demands and the "black-box" nature of the model that can reduce trust and slow clinical adoption [21]. Improved transparency is needed to translate automated decision making to clinical practice. To overcome these obstacles, research is increasingly focusing on incorporating explainable AI (XAI) methods, which enhance transparency by clarifying how deep learning models make decisions [22].

This research focuses on using advanced technologies to improve healthcare in countries with limited financial resources, like Sri Lanka. In such regions, access to advanced medical tools and expert knowledge is often lacking, making it difficult to diagnose diseases like breast cancer accurately and quickly. Detecting breast cancer at an early stage is essential for improving treatment outcomes, but resource limitations can lead to delays and incorrect diagnoses.

Given the financial challenges in Sri Lanka's healthcare system, it's essential to explore new solutions that make the best use of available resources. XAI improves the understandability of AI models, allowing their decision-making process clearer and

more understandable, which helps build trust among healthcare providers.Even with limited resources, AI-powered tools can assist doctors identify and classify breast cancer more precisely, enabling earlier diagnosis and more effective treatment planning.

1.2 Problem Statement

Breast cancer continues to be one of the most dangerous forms of tumors, often advancing through multiple stages before becoming critical. Detecting the disease early significantly improves treatment success, as breast cancer is generally more manageable in its initial stages [23]. Although mammography remains the gold standard for the detection and diagnosis of breast cancer, it still depends heavily on human interpretation, which can introduce inconsistencies. Potentially resulting in false positives which may cause unnecessary interventions or false negatives, which could delay treatment and worsen patient outcomes.

CAD systems were developed to assist radiologists by improving diagnostic consistency and identifying potential abnormalities in mammograms. While these systems have reduced the burden on radiologists and improved accuracy in certain cases, they also exhibit several limitations. CAD systems can produce false alarms, increasing the rate of unnecessary follow-ups, biopsies, and patient anxiety. On the other hand, false negatives present an even more serious risk by allowing undetected malignancies to progress. These limitations reveal the need for enhanced diagnostic solutions that reduce errors and improve early detection capabilities.

In recent years, DL models have demonstrated remarkable success in breast cancer detection by automatically extracting complex patterns from mammograms. CNNs and ViTs have emerged as the most promising DL architectures in this domain. CNN models excel at detecting localized image features, while ViTs are particularly effective at capturing long-range dependencies.

Despite their impressive accuracy, these models are often perceived as "black-box" systems, providing predictions without clear explanations. This lack of interpretability raises significant concerns for medical professionals who rely on clear, evidence-based reasoning to make informed decisions. Consequently, healthcare providers may hesitate to adopt DL-based tools in clinical practice without a reliable way to interpret their predictions. In addition, a mechanism for understanding the factors that influence the risk of developing the disease, as well as identifying the characteristics that are most important to predict whether a person will develop the disease, was essential with the rapid increase in the number of breast cancer cases. Hence, Explainable AI(XAI) [24, 25] can be used as a solution for this necessity.

Figure 1.3 illustrates how XAI introduces new capabilities to traditional AI by addressing the "wh" questions that earlier models could not answer [26]. XAI also plays a significant role in offering more individualized predictions regarding a patient's like-

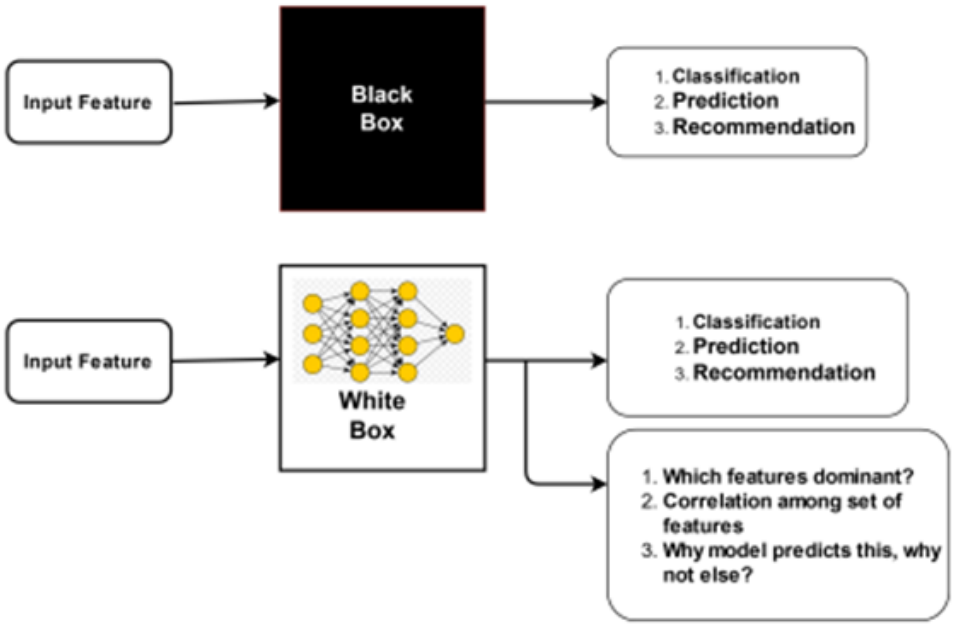

Fig. 1.3: AI vs XAI

lihood of developing breast cancer. Furthermore, it can uncover patterns linked to the early stages of the disease, helping to create models that are more effective at identifying individuals at higher risk. Additionally, XAI enables medical professionals to gain deeper insights into the underlying factors influencing breast cancer development, paving the way for more personalized and targeted treatment plans.

Furthermore, this research focuses on using advanced technologies to improve healthcare in countries with limited financial resources, like Sri Lanka. In such regions, access to advanced medical tools and expert knowledge is often lacking, making it difficult to diagnose diseases like breast cancer accurately and quickly. Timely and precise detection of breast cancer is essential for improving patient outcomes; however, limited resources can result in delays and diagnostic errors.

Given the financial challenges in Sri Lanka's healthcare system, it's essential to explore new solutions that make the best use of available resources. XAI improves the clarity and interpretability of AI models, making their decision-making processes more understandable, which helps build trust among healthcare providers. Despite resource constraints, AI-driven tools can support doctors in precisely identifying and categorizing breast cancer, enabling earlier diagnoses and more effective treatment strategies.

In summary, this research addresses several critical problems in breast cancer detection. Key issues include the risk of false positives and false negatives in diagnostic methods, inconsistent performance in CAD systems, and the "black-box" nature of

deep learning models that limits their interpretability. Additionally, limited access to medical expertise in resource-constrained regions like Sri Lanka further delays diagnosis, affecting patient outcomes. These obstacles highlight the pressing need for improved diagnostic tools that are accurate, reliable, and interpretable to support clinical decision-making and enhance breast cancer detection worldwide. The following section highlights the key research gaps that arise from these challenges.

1.3 Research Gap

Correct identification and classification of breast cancer are critical for early diagnosis and effective treatment strategies. With advancements in technology, DL techniques have demonstrated remarkable promise in the field of medical imaging, especially for breast cancer detection [20, 27, 28]. DL models, such as CNNs and ViTs, have shown impressive performance in tasks like classifying images, detecting tumors, and segmenting breast tissue in mammograms, ultrasounds, and MRIs. These models are capable of learning complex patterns directly from large volumes of data, identifying subtle patterns that might be overlooked by human observers. Thanks to their high precision, deep learning systems are increasingly supporting radiologists and oncologists by enabling earlier detection of breast cancer, leading to more effective treatments and improved patient survival rates.

In recent years, CNNs have become the dominant architecture for analyzing medical images. Nevertheless, despite the success of CNN-based models, there are still limitations, such as difficulty in capturing long-range relationships and context within images. Additionally, CNNs depend on design choices like kernel sizes and pooling, which may not always capture the complex patterns in medical records. Conversely, ViT,a recent architecture adapted from the transformer models used in natural language processing, have showcase promising results in image recognition. Unlike CNNs, ViTs are capable of capturing connections among all sections of an image and global context in images through the self-attention mechanism. In contrast, ViTs use self-attention mechanisms, which allow them to consider relationships between all parts of an image, regardless of their spatial distance.The scalability of the ViT models particularly attractive for tasks where large annotated datasets are available, such as in medical imaging or high-resolution object recognition tasks [29]. Nevertheless, ViTs are still not well-explored in medical imaging, especially compared to CNNs, and research on their use in breast cancer detection is limited. A thorough comparative analysis between CNN and ViT in breast cancer detection could reveal critical insights about which architecture is more effective for identifying malignancies in medical imaging.

Despite their potential, both CNN and ViT models face a common limitation: the "black-box" nature of their decision-making processes. While these models provide accurate predictions, they often lack transparency, posing a challenge for clinical adop-

tion. Understanding the rationale behind model predictions is crucial in medical contexts where trust in diagnostic decisions is paramount. This is where Explainable Artificial Intelligence (XAI) methods can play a transformative role. XAI techniques, such as heatmaps and attention maps, can visually highlight regions of interest within mammogram images, providing medical professionals with clear visual cues on how predictions were formed. By incorporating XAI, models become more interpretable, facilitating alignment with clinicians' observations and enhancing trust in AI systems.

This research aims to address the current gap by conducting a comparative study of CNN and ViT architectures, integrated with XAI techniques, to improve the transparency and reliability of breast cancer detection tools. The goal is to support the clinical adoption of AI-driven solutions by ensuring model decisions are both accurate and explainable. By achieving this, medical professionals will be better equipped to understand and trust these advanced diagnostic tools, fostering improved decision-making in clinical practice.

To address these research gaps, this study outlines specific objectives focused on developing clinically explainable AI models that not only enhance diagnostic accuracy but also improve usability for medical professionals. The next section presents these research objectives in detail.

1.4 Research Objectives

This project aims to develop an interpretable classification model for the detection of breast cancer through mammogram analysis. The key objectives are outlined below.

1. Investigate the performance of CNN and ViT to detect medical image analysis of breast cancer.

   The research will explore how effective of CNNs and ViTs for detecting breast cancer from mammogram images. Key architectures such as ResNet50, ResNet101, VGG16, VGG19, MobileNetV2, DenseNet121, InceptionResNetV2, and ViT will be implemented and evaluated. This objective aims to assess the strengths and weaknesses of each architecture, identifying the most effective model for accurate and efficient breast cancer detection.

2. Integrate Explainable AI techniques for Model Transparency and evaluate the effectiveness of XAI techniques in medical application.

   Another key objective is to integrate advanced Explainable AI (XAI) techniques to make the classification model more transparent and interpretable. Techniques such as Grad-CAM, SIDU, Attention Maps, and Ablation-based Class Activation Mapping (Ablation CAM) will be utilized. A comparative analysis will also be conducted to measure the performance gains achieved by integrating XAI methods.

3. Design an intuitive web application to support both medical professionals and patients, facilitating the integration of AI models into clinical decision-making processes.

   A significant focus of this research is the development of a user-friendly web application that will allow both medical professionals and patients to upload and analyze mammogram images. This platform will integrate the XAI-enhanced breast cancer detection model and provide clear, understandable explanations of the AI's predictions. Furthermore, it will help reduce diagnostic time, minimize false negative rates, and enhance the overall patient experience by providing transparent, interpretable results. Additionally, the integration of XAI methods will make it easier for patients to understand their diagnosis and provide them with more control over the decision-making process, while also reducing the costs associated with manual oversight in breast cancer detection.

By achieving these objectives, this research aims to contribute to the development of trustworthy, interpretable, and clinically viable breast cancer detection system in Mammography. With these objectives in place, the next chapter reviews existing literature on AI-driven breast cancer detection. The discussion will focus on screening methods, deep learning techniques, and the role of XAI in medical imaging.

CHAPTER 2

LITERATURE REVIEW

Breast cancer detection has witnessed significant advancements through the integration of medical imaging technologies and AI systems. This chapter reviews the existing literature on breast cancer screening methods, deep learning techniques applied to mammogram analysis, and the role of XAI in enhancing model transparency.

2.1 Breast Cancer and Screening Methods

Breast cancer remains one of the most widespread types of cancer globally, particularly affecting women. Early detection significantly improves treatment outcomes and survival rates. In order to support in early detection, screening has emerged as a popular method. This section outlines the various medical imaging techniques that are essential in the assessment and detection of breast cancer. Key imaging modalities used for early-stage diagnosis include "X-ray mammography (MG), breast thermography (BT), magnetic resonance imaging (MRI), positron emission tomography (PET), computed tomography (CT), three-dimensional ultrasound (US), and histopathological examination (HP)" [30].

Mammography is crucial for breast cancer screening, particularly for women over 40, and can also assist in monitoring existing breast conditions or detecting abnormalities. The process involves compressing the breast to spread out the tissue for clearer images, helping radiologists evaluate the presence of suspicious areas. Mammography is widely recognized in medical literature as the primary imaging technique used in breast cancer screening and computer-aided diagnosis, often referred to as the 'gold standard' for early detection [31, 32]. Utilizing low-dose X-rays, mammograms are quick to perform and commonly serve as the first-line method for identifying abnormalities in breast tissue. As illustrated in Figure 2.1, malignant tumors and calcium deposits typically appear as brighter regions on mammographic images. This contrast allows for effective interpretation by experienced radiologists or deep learning models trained on annotated mammogram data [32].

Furthermore, mammography serves as the initial screening tool for breast cancer in many healthcare systems due to its accessibility, speed, and proven effectiveness in detecting early breast cancer. However, it should be noted that mammography alone may not always detect all types of breast cancer, especially in women with dense breasts, where radiopaque tissue can obscure tumors. In such cases, additional imaging techniques such as ultrasound or MRI may be used to balance mammography.

Publicly available datasets such as CBIS-DDSM, INBreast, and the RSNA Breast Cancer Detection Dataset have been widely utilized in developing AI models for breast

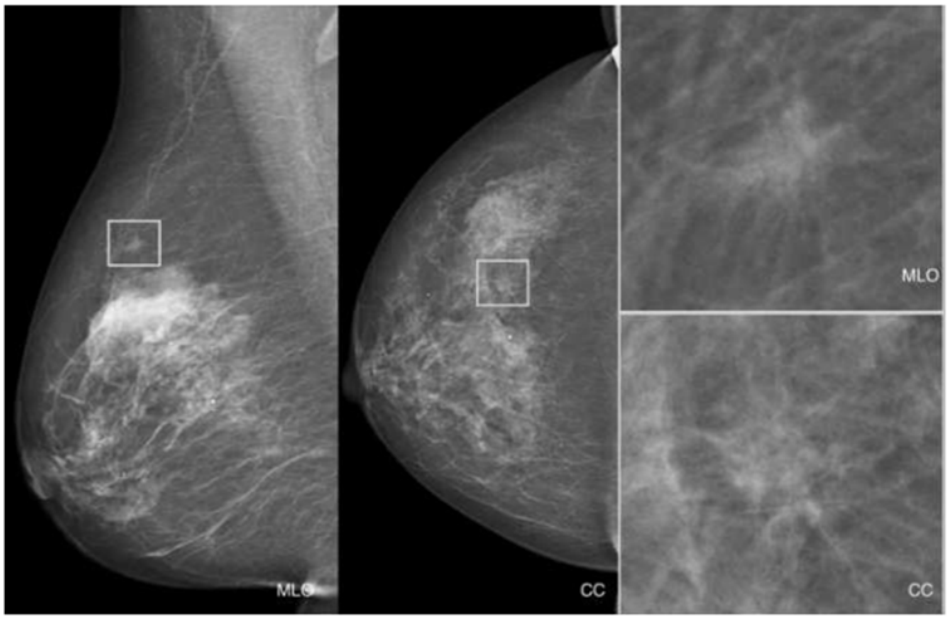

Fig. 2.1: Mammographic X-Ray of Human Breast
[30]

cancer detection. The CBIS-DDSM dataset provides annotated mammogram images for model training, while INBreast offers high-quality full-field digital mammograms. The RSNA dataset is designed for large-scale AI model evaluation. Additionally, datasets like Mini-MIAS, DDSM, and BCDR are frequently referenced in breast cancer detection research for algorithm development and validation. Table 2.1 provides more details about the publicly available mammogram datasets.

TABLE 2.1: Publicly Available Mammogram Datasets

Database Name	Link
Mini-MIAS	http://peipa.essex.ac.uk/info/mias.html
DDSM	http://marathon.csee.usf.edu/Mammography/Database.html
INBreast	https://biokeanos.com/source/INBreast
BCDR	https://bcdr.ceta-ciemat.es/information/about
CBIS-DDSM	https://wiki.cancerimagingarchive.net/pages
MIAS	https://www.repository.cam.ac.uk/handle/1810/250394

Although these data sets are invaluable, they also introduce challenges that researchers must address. Dataset biases, such as class imbalance (where benign cases outnumber malignant cases) and limited demographic diversity, can impair model generalization. In addition, inconsistencies in image quality, labeling errors, and variations in imaging protocols add complexity to model training and evaluation.

While traditional imaging techniques such as mammography are effective in de-

tecting breast cancer, they have some limitations, especially in detecting hidden abnormalities. Deep learning techniques have been extensively investigated to enhance accuracy and support radiologists in the detection of breast cancer. The next section discusses these techniques and their role in enhancing mammography-based diagnosis.

2.2 Evaluation of Deep Learning Techniques for Breast Cancer Detection Using Mammography

Machine learning (ML) has become a pivotal area of innovation in breast cancer detection, offering the potential to significantly enhance and modernize conventional diagnostic approaches. A wide range of ML algorithms have been explored for identifying and classifying breast cancer, contributing to improved diagnostic accuracy and early detection capabilities. In [33], features such as tumor radius, concavity, texture, and fractal dimensions were analyzed to distinguish between benign and malignant tumors using the Wisconsin Breast Cancer Database (WBCD). Several algorithms, including Logistic Regression, Nearest Neighbor, and Support Vector Machines (SVM), were employed in this study. Similarly, [34] leveraged the same dataset, applying ML techniques such as artificial neural networks (ANNs), SVM, decision trees (DTs), and k-nearest neighbors (k-NNs), demonstrating that ANNs have played a dominant role in the diagnosis and prognosis of breast cancer. This trend reflects the growing use of diverse ML methods in intelligent healthcare systems, offering a variety of diagnostic options to clinicians. Additionally, in [35], a hybrid approach combining feature selection and extraction methods was explored. Linear Discriminant Analysis (LDA) was used to reduce the feature space, followed by classification using SVM, Naive Bayes, and ANN models. Among these, combinations like SVM-LDA and NN-LDA achieved superior performance, with sensitivity, precision, and recall rates of 98.41% and an overall accuracy of 98.82

Identifying the most suitable features for machine learning (ML) models can be a complex and demanding process. It typically involves extensive domain expertise and meticulous preprocessing to transform raw data into formats that algorithms can effectively utilize. Furthermore, ML techniques often depend heavily on large volumes of accurately labeled data and subject matter knowledge to deliver reliable results. In scenarios where such data or expertise is limited, the performance of these models can be significantly compromised. To reduce the reliance on manual feature extraction and domain-specific engineering, researchers have increasingly turned to DL approaches. These models automatically learn feature representations from raw data, and have demonstrated exceptional success across diverse fields, including image analysis, speech recognition, and language processing. They have outperformed traditional machine learning algorithms in many applications, achieving state-of-the-art results.

The integration of AI,, especially deep learning techniques, has opened new path-

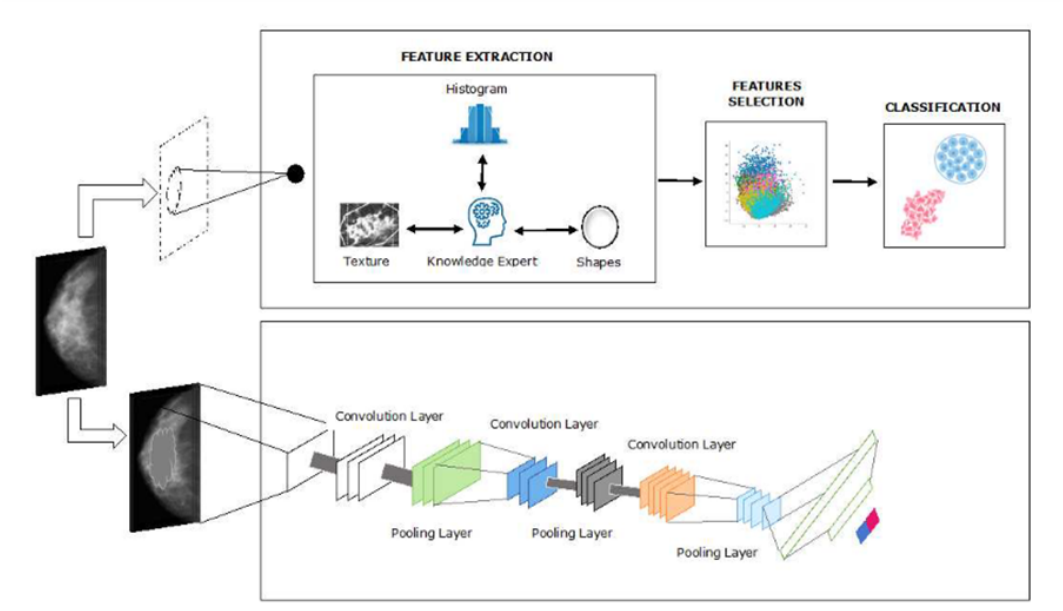

Fig. 2.2: ML and DL Approach
[30]

ways for enhancing traditional diagnostic practices. In Figure 2.2, the top row outlines the typical workflow of nl models, while the bottom row illustrates the structure of a CNN, a key architecture within DL. Several researchers, such as those in [36], have conducted comprehensive reviews of DL approaches for breast cancer detection, offering valuable insights into existing challenges and emerging trends that can assist both practitioners and researchers in advancing the field.

There has been a lot of research in recent years aimed at using deep learning algorithms to maintain the accuracy and productivity of breast cancer recognition. DL algorithms have been used to develop CAD systems that can analyze mammograms and assist radiologists in identifying potential breast cancer cases. Most of these systems use CNNs to identify patterns and images and make predictions about the presence of breast cancer.

Several studies, such as Araujo et al [37], Rakhlin et al. n.d. [38] showcase the efficacy of advanced computational approaches, particularly utilizing CNNs, for accurate and automated breast cancer histology image classification. While these methods achieved high performance, they lacked clear strategies for handling dataset biases.

The Zhu et al's research [39] introduced a hybrid CNN model integrating global and local branches with a Squeeze-Excitation-Pruning (SEP) block. Although the SEP block improved model robustness and reduced overfitting, the study lacked comprehensive evaluation on diverse datasets, leaving questions about its generalization.

Besides, the concentrate by Zheng et al. [40], proposed a "Deep Learning-Assisted Efficient Adaboost Algorithm (DLA-EABA) for breast cancer detection", which integrates deep convolutional neural networks with transfer learning and feature extrac-

tion. Their approach achieved notable results, with an accuracy of 97.2%, sensitivity of 98.%, and specificity of 96.5%, outperforming existing classification systems. However, the method's significant computational demands present challenges for real-world implementation.

Shen et al. [41], introduced an "end-to-end deep learning algorithm for breast cancer detection" in screening mammograms, demonstrating strong performance on "datasets such as CBIS-DDSM and INbreast". While the model effectively reduced false positives and false negatives, its reliance on these datasets raises concerns about potential bias, particularly due to their demographic limitations and class imbalance. Additionally, the study did not address strategies for handling imbalanced data, which risks overfitting to benign cases. Furthermore, the absence of interpretability methods like Grad-CAM or SHAP limits transparency, making it difficult for radiologists to understand the model's decision-making process. Future improvements should prioritize diverse datasets, enhanced interpretability methods, and computational efficiency to improve clinical integration.

Moreover, Suh et al. [42], developed and evaluated a deep learning model for breast cancer detection in digital mammograms, comprising 3002 merged images from 1501 subjects. Two CNNs, DenseNet-169 and EfficientNet-B5, exhibited high accuracy with mean AUCs of 0.952 and 0.954, surpassing a meta-analysis mean AUC. The model's effectiveness decreased with increasing breast density, highlighting its efficiency in screening breast cancer, particularly in breasts with lower density of the parenchyma.

Additionally, some researchers [43] proposed a lightweight detection and classification approach based on an enhanced version of YOLOv5 for identifying and categorizing breast tumors. Their work utilized all four YOLOv5 variants (YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x) with specific architectural improvements. Transfer learning was employed to adapt deep learning models pre-trained on large-scale datasets to the task, helping improve model performance on smaller breast cancer datasets. Similarly, the application of transfer learning using AlexNet for breast tumor classification and detection was explored in [44]. Although this technique effectively leverages prior knowledge for medical image analysis, it introduces a risk of overfitting when complex models are fine-tuned on limited or imbalanced data. The study did not extensively discuss mitigation strategies like data augmentation, dropout, or early stopping to address this issue.

In Summary, in Table 2.2, compares various studies on breast cancer detection using deep learning pre-trained models techniques applied to different publicly available datasets. The variation in performance may also be changed to the different datasets used in these studies. Datasets like CBIS-DDSM and INBreast tend to generate better results with more advanced models.Overall, ResNet50 and Inception-based models are among the top performers, delivering reliable results in breast cancer detection.

TABLE 2.2: Comparison of Breast Cancer Detection Techniques in Various Studies

Author (Year)	**Dataset**	**Technique**	**Testing Accuracy**	**AUC**
L. Shen et al. (2017) [20]	CBIS-DDSM Mini-MIAS	ResNet50 VGG16	97% 84%	0.98 0.95
N. Khan et al. (2019) [45]	Mini-MIAS CBIS-DDSM	VGG16,VGG19,ResNet-50 Inception-v3	93.73% 92.29%	0.93 0.90
J. Suh et al. (2020) [42]	Private Database	DenseNet-169 EfficientNet-B5	96.2% 95.2%	0.96 0.99
P. Xi et al. (2018) [46]	CBIS-DDSM	AlexNet,VGGNet, GoogleNet,ResNet	92.53%	-
Chougrad et al. (2017) [47]	DDSM INbreast BCDR	VGG16 ResNet50 InceptionV3	96.22% 92.00% 97.50%	0.98 0.97 0.97
Saber et al. (2021) [48]	MIAS	Inception V3 Inception V2 ResNet 50 VGG16 VGG19	96.19% 93.42% 95.27% 96.77% 94.35%	0.99 0.97 0.97 0.99 0.97
Al-Antari et al. (2020) [49]	INBreast	CNN from scratch ResNet50 InceptionResNet-V2	88.74% 92.33% 95.32%	0.87 0.97 0.93
A. Sahu et al. (2023) [50]	DDSM INBreast	CNN ResNet50 InceptionResNet-V2	94.50% 95.83% 97.50%	-
S. Dias et al. (2023) [51]	RSNA	VGG Googlenet EfficientNet Residual Networks	91% 93% 95% 94%	-

Additionally, while CNNs have played a pivotal role in advancing computer vision, a significant paradigm shift has emerged with the introduction of Vision Transformers (ViTs) [52]. ViTs mark a significant shift from traditional CNNs by utilizing the Transformer architecture, which was initially created for handling sequential data, in the context of image analysis. While CNNs process raw pixel data by recognizing local structures and spatial hierarchies, ViTs reframe images as sequences of patches, similar to how words are treated as tokens in NLP. This change in how image data is represented allows ViTs to capture the holistic context of an image, facilitating the learning of different patterns and relationships through the utilization of self-attention mechanisms [29, 53]. Table 2.3 depicts a comparison of ViT and CNN models.

Cantone et al.'s research [54] study compares CNNs and Vision Transformers for mammogram classification, finding Vision Transformers to be promising alternatives to CNNs. Through an extensive experimental analysis of the OMI-DB database, involving 33 models across various resolutions and lesion categories, the study demon-

strates the comparable performance of Vision Transformers to traditional CNNs like ResNet. However, modern convolutional networks like EfficientNet exhibit superior performance in certain scenarios, highlighting their continued relevance in medical imaging.

Also, a recent study [55] compared three deep learning architectures ViT, Compact Convolutional Transformer (CCT), and TokenLearner ViT (TVIT) — for binary classification of mammography images using the DDSM dataset. The models achieved high accuracies of 99.81% (ViT), 99.92% (CCT), and 99.05% (TVIT), outperforming existing methods. The outcomes emphasize the success of convolution-attention mechanisms in developing efficient and accurate computer-aided breast cancer diagnosis systems while reducing computational costs and decision time.

Moreover, G.Ayana et al's study [56] highlights that CNNs have primarily focused on specific portions of the mammogram,may be ignoring important areas. In contrast, ViTs demonstrate a more comprehensive approach by considering the entire image, allowing them to capture a broader context and potentially improving the detection of abnormalities across the full mammogram. When comparing the results, the ViT based transfer learning models achieve the highest AUC of 1 ± 0, while the CNN-based transfer learning models, with ResNet50 achieving a maximum AUC of 0.95 ± 0.01, perform significantly worse. This suggests that ViT outperform CNNs in analyzing mammograms.In summary, Table 2.4 compares few studies on breast cancer detection using ViT-based techniques, applied to different publicly available datasets.

TABLE 2.3: Comparative Analysis of ViT and CNN Architectures

Feature	ViT	CNN
Image visualization	Gradually getting bigger	Big vision right from the ground
Global feature information acquisitions	Shallow	Higher Level
Jump Connection	High impact	High impact
Space Information	Keep more	More reservations
Middle layer feature learning	Better learning results	Better learning results
Generalization Bias	No	There are

A recent study [63] tackles the challenge of early breast cancer detection by introducing a novel dual-track deep learning architecture designed to classify abnormalities such as masses and microcalcifications in mammograms. The approach combines a Dense-unified Multiscale Attention Fusion" (UMAF) track and a Data-efficient Image Transformer (DeiT) track. The DeiT which is a variation of ViT, track captures global, multiscale features across the entire image through patch embeddings, While Dense-UMAF is dedicated to extracting detailed localized features, leveraging DenseNet connectivity to enhance feature reuse and mitigate vanishing gradient issues. By capturing

TABLE 2.4: Comparison of Breast Cancer Detection Techniques Using ViT

Author (Year)	**Dataset**	**Technique**	**Accuracy**	**AUC**
L. Mouhamed et al. (2017) [57]	DDSM	VIT CCT TVIT	99.81% 99.81% 98.55%	99.99% 100% 99.90%
B. Gheflati et al. (2019) [58]	BUSI Ultra-Sound	ViT+ResNet	82%	92%
A. Dias et al. (2023) [59]	BreakHis	DeiT	98.17%	-
S. Tummala et al. (2022) [60]	BreakHis	Swin Transformer (SwinT)	99.4%	-
S. Boudouh et al. (2024) [61]	CBIS-DDSM	ViT++	96.12%	-
O. Tanimola et al. (2024) [62]	DDSM InBreast MIAS	SwinT	98.88%	-
O. Tanimola et al. (2024) [55]	DDSM	ViT CCT TVIT	99.81% 99.92% 99.05%	-

both local and global patterns crucial for diagnosing breast cancer, the model achieved a classification accuracy of 88.69% on the CBIS-DDSM dataset.

Following the successful application of ViT models for histopathology-based breast cancer classification, where deep learning approaches like ViT and Data-Efficient Image Transformer (DeiT) ensembles demonstrated high accuracy, recent studies have further explored deep learning advancements in mammography analysis. Specifically, a hybrid deep learning framework was proposed in [61], combining a Vision Transformer (ViT++) for capturing holistic information and a CNN based on transfer learning for extracting visual attributes. After a preprocessing phase involving noise reduction and enhancement, the model incorporated pretrained CNN architectures such as Xception, VGG16, and RegNetX002. Experimental results on the CBIS-DDSM dataset revealed that the integration of ViT++ with VGG16 achieved a classification accuracy of 99.22%, substantially outperforming individual models like VGG16 alone. These findings highlight the evolving trend of leveraging hybrid deep learning architectures, particularly Vision Transformers in combination with CNNs, to achieve superior performance in breast cancer detection tasks across different imaging modalities.

Notably, while ViTs are a powerful tool, they are still challenged by the same issues as CNNs when it comes to limited labeled data, complex image patterns, and computational demands. However, recently new techniques like self-supervised learning, interactive segmentation models, and combining different types of images have been introduced. These methods aim to solve these issues and make medical image analysis more accurate, scalable, and easier to use. Thus, while ViTs are promising, they still need some adjustments to be fully optimized for medical imaging tasks.

Self-supervised learning has become a valuable technique in medical image anal-

ysis, especially when annotated datasets are limited. DINO (Distillation with NO labels), a self-supervised learning model created by Meta AI, uses ViTs to learn detailed visual representations without requiring labeled data. DINO v2 builds on this idea, providing a more refined method for pretraining deep neural networks on large datasets, which can then be fine-tuned on specific medical imaging tasks such as breast cancer detection [64]. Self-supervised learning enables the model to automatically identify complex patterns in imaging data that may be difficult for human observers to notice, helping radiologists detect subtle indications of cancer.

Moreover, the Segment Anything Model (SAM), developed by Meta AI, represents a major advancement in interactive image segmentation. In the medical domain, the Medical-SAM extension builds upon the original SAM model, specifically tailored to handle medical images like mammography, CT scans, MRIs, and X-rays with little to no manual annotation required [65]. Recent studies have shown that Medical-SAM significantly reduces the time required for segmenting regions such as brain tumors in MRI scans or lung nodules in CT images, allowing radiologists to focus on diagnosis rather than annotation tasks. In the context of mammography, Medical-SAM can assist in the segmentation of critical areas like breast tissue, tumors, or microcalcifications, which are essential for detecting breast cancer.

In addition, TransUNet is a deep learning model that combines transformers with the popular U-Net architecture to improve medical image segmentation. By integrating transformer-based attention mechanisms with the U-Net architecture helps to find tumors or small changes in the breast tissue, which are important signs of early-stage breast cancer [66]. The model effectively captures detailed and broad features within the image allows it to more precisely highlight areas of concern, even in complex mammograms where tumor boundaries may be unclear. This makes TransUNet a valuable tool for improving the early detection and diagnosis of breast cancer, supporting radiologists in making faster and more accurate decisions.

In summary, the integration of self-supervised learning models like DINO v2, interactive segmentation models like Medical-SAM, and advanced segmentation models like TransUNet could provide clinicians with more powerful and intuitive tools for medical image analysis. However, challenges remain in terms of data privacy, bias in training data, and ensuring model interpretability for clinical decision-making. As these models continue to improve, future research should focus on complete integration of multi-modal data (combining Mammogram, CT, MRI, and other imaging techniques) to create more comprehensive and accurate diagnostic tools.

While deep learning models like CNNs and ViTs have shown promising results in breast cancer detection, understanding how these models make decisions is crucial for clinical adoption. Past research has explored various XAI techniques to address this challenge. The next section reviews existing literature on XAI techniques and their contribution to enhancing the interpretability of models used in breast cancer detection.

2.3 Explainable Artificial Intelligence (XAI) Methods

Although AI and DL algorithms are powerful and widely used approaches in machine learning for solving complex problems in areas such as medical image analysis and disease predictions, there are some limitations and challenges associated with deep learning. The main drawbacks is DL models require massive set of high-quality images to achieve good performance, making them less effective in low-data scenarios. Moreover, AI and DL algorithms Black Box Nature, due to the intricate nature of deep learning models, it becomes challenging to understand how they arrive at their predictions, limiting their Explainability [67–69]. Concurrently, deep learning enthusiasts have shown a growing interest in XAI techniques in their new researches.

Explainability is important for breast cancer detection because it enhances the transparency and trust in the system. It allows medical professionals to comprehend the key factors influencing the model's decisions. By making the model's behavior more interpretable, explainability contributes to improved accuracy and reliability, ultimately boosting clinicians' confidence in using such technology. Hence, the XAI method is going to be used in this proposed method.

According to [70, 71], XAI methods that are used to explain predictions made by models can be categorized into two classes namely Model-Agnostic and Model-Specific. The model Agnostic method is mostly appropriate to post-hoc analysis and not constrained to particular model architecture. These explanation methods work independently of DL models, relying solely on the models' inputs and outputs without altering their internal structures. Model-specific methods are limited to the specific model classes. Model-specific techniques are efficient and can quickly analyze a neural network in a single pass, focusing on its specific architecture. In contrast, model-agnostic explanation techniques are more general but tend to be computationally intensive as they involve making significant changes to input images to observe how the model reacts to these perturbations.

Moreover, image based XAI methods are categorized into two main categories: attribution and non-attribution [72].

1. Attribution-Based Methods: This method produces visual interpretations by emphasizing the areas within an image that most influence the model's prediction.It is further categorized into gradient-based, perturbation-based, and layer-wise relevance propagation techniques.

    (a) Gradient-Based Methods: In Gradient-Based Methods, gradients are used to understand which input features are most responsible for the final prediction. These techniques create attribution maps from input images by utilizing gradients or backpropagation to emphasize the regions most critical to the model's prediction. Typically, the explanations are model-agnostic and

applied after model training. Examples of such techniques include Class Activation Mapping (CAM), Gradient-weighted Class Activation Mapping (Grad-CAM), Grad-CAM++, Integrated Gradients (IG), SmoothGrad, and XRAI.

(b) Perturbation-based methods: This methods are a category of interpretability techniques used to explain machine learning model predictions by systematically altering the input data and observing how the model's output changes. Examples of such methods include LIME (Local Interpretable Model-agnostic Explanations), Randomized Input Sampling for Explanation(RISE), Similarity Diference and Uniqueness(SIDU), Ablation CAM and SHAP(SHapley Additive exPlanations).

(c) Layerwise relevance propagation: This method traces the model's output back to specific input features, revealing which parts of the input had the greatest impact on the prediction. This is done by redistributing the relevance score of the model's results across its layers, ultimately reaching the input features. Standard LRP, LRP-Z + and Deep Taylor Decomposition are some of the example method for this category.

2. Non-Attribution Methods: This approach emphasizes understanding the reasoning behind a decision and explains the prediction by exploring concepts, prototypes, and altered predictions. Instead of only focusing on pixel-level information, it examines model behavior, analyzes sensitivity and stability, and assists in model debugging.

   (a) Concept Based: These techniques that help explain a model's predictions by linking them to high-level concepts, rather than focusing on individual features or details. These concepts are easier for humans to understand and can represent things like "stripes" in an image or "positive sentiment" in a piece of text. Testing with Concept Activation Vectors(TCAVs), Automatic Concept-based Explanations(ACE), ConceptShap an,d Causal Concept Efect(CACE) can be defined as examples of this method.

   (b) Counterfactual-based: This aim to interpret model predictions by making small changes to an image that would result in a different outcome. These techniques reveal which slight alterations to the input could shift the model's decision, offering valuable insight into how the model reasons for individual cases. Examples of counterfactual approaches include Guided by Prototypes, the Contrastive Explanations Method (CEM), and Learning to Explain (L2X).

   (c) Prototype-based techniques: This helps users understand a model's prediction by presenting cases that are similar to the original input. These

representative examples, or "prototypes," serve as a way to categorize or make decisions, much like how humans use familiar examples to reason and make judgments. By showing similar instances, the model's reasoning becomes clearer and easier to interpret. Influence Functions, Prototypical Part Network (ProtoPNet), and Maximum Mean Discrepancy Critic (MMD-Critic) are techniques that are examples of this method.

Moreover, visualizing the regions of the input, such as medical images, that contributed to the model's output is an effective approach to providing this transparency. Several visual explanation methods have been proposed to address the challenge of interpreting deep learning models, and they often focus on producing saliency maps, heat maps that highlight the regions in the input image or data that contributed the most to the model's decision. These saliency maps can be used to identify and visualize the most important areas of the input data, making it easier for humans grasp the reasoning behind the model's specific prediction. Some of the most widely used visual explanation methods are as follows.

1. LIME- Local Interpretable Model-agnostic Explanations [73]

   This model-agnostic approach highlights that LIME can explain the predictions of any supervised learning model. It is versatile and can be applied to various types of data, including images, text, and videos. LIME operates by slightly altering the input data and analyzing the resulting changes in predictions, providing insights into the model's behavior within a specific local region. However, a key limitation of LIME is the difficulty in choosing an appropriate local surrogate model to accurately approximate the complex decision boundaries of the original model.

2. SHAP (SHapley Additive exPlanations) [73]

   SHAP aims to explain the prediction of a specific input by calculating how much each feature contributes to the final decision. It does this by applying Shapley values derived from coalitional game theory. Unlike LIME, which assigns weights to its local linear model based on the cosine similarity between the original and perturbed inputs, SHAP determines these weights using the Shapley value formula. Despite its effectiveness, SHAP can be highly computationally intensive, especially when dealing with models that have many features, making it less practical for large datasets or deep neural networks.

3. Grad-CAM (Gradient-weighted Class Activation Mapping) [74]

   Grad-CAM is a popular interpretability technique in deep learning that highlights the regions of an image most influential to a model's prediction. It produces a heatmap by calculating the gradients of the target class with respect to

the feature maps from the final convolutional layer.While Grad-CAM has proven useful in various tasks, it faces challenges in accurately localizing the salient regions, particularly when objects are fragmented or when the model's decision depends on fine-grained features.

4. Ablation-CAM (Ablation based Class Activation Mapping) [75]

   Ablation-CAM is an improved version of traditional visualization methods like Grad-CAM, aimed at providing more precise and focused explanations of a neural network's predictions. Unlike Grad-CAM, which relies on gradient information, Ablation-CAM works by systematically removing parts of the model's feature maps during computation to evaluate the output. This offers a more accurate view of the areas that affect the decision-making process.

5. RISE (Randomized Input Sampling for Explanation) [76]

   RISE is another visual explanation method that generates saliency maps by randomly changing the input and using the corresponding outcomes to compute a weighted importance score for each pixel. Unlike Grad-CAM, RISE does not depends on gradients and is more robust to noisy gradients in certain cases. However, it may not always provide precise localizations of salient regions, especially when complex features or spatial relationships are crucial for the decision-making process.

6. SIDU (Similarity Difference and Uniqueness) [77]

   SIDU is a novel method designed to improve the localization of salient regions in deep learning models. It addresses some of the drawbacks of older methods such as Grad-CAM and RISE, which face challenges with fine-grained localization. SIDU uses the final convolutional layer of a deep CNN model to give similarity differences and uniqueness masks, which are then combined to produce a saliency map that highlights more accurately the regions of the image contributing to the model's prediction. SIDU aims to provide improved localization, particularly when higher classification accuracy is essential, and thus contributes to greater trust in AI systems, particularly in sensitive fields like healthcare.

7. Saliency maps [78]

   Saliency maps is another widely used method, assign an importance score to each pixel in an image by calculating the gradient of the model's result based on the input image pixels. These gradients are visualized as a heat map, emphasizing the pixels that have the greatest impact on the decision of the model. While saliency maps are relatively simple to generate and easy to interpret, they

can sometimes be noisy and may not provide precise explanations, especially for complex models.

8. DeepLIFT (Deep Learning Important FeaTures) [79]

   DeepLIFT compares the activation of the model on a given input to the activation on a reference input (usually a baseline or zero input). The difference in these activations is used to calculate the contribution of each feature.

Many Explainable Artificial Intelligence algorithms have already been developed for medical image categorization. For instance, Imouokhome et al. ''s [80] research addresses the importance of early breast cancer detection, emphasizing the lack of visual interpretation in existing models. Utilizing the BreakHis dataset, a ResNet50 model achieved a remarkable 96.84% accuracy in classifying tumors as malignant or benign, surpassing previous deep learning-based studies. This study further employed Explainable Artificial Intelligence (XAI) techniques like Integrated Gradient, GradientShap, and Occlusion to interpret and visually explain the classification results. Among these, Occlusion stands out for its superior predictive results, offering insights into why specific histopathological scans are classified as benign or malignant.

Furthermore, Kashefi et al's study [81] investigates explainability methods for visual transformers, which is crucial for understanding their decision-making processes. ViT consists of attention mechanisms (self-attention layers) that assign pairwise attention values between two image patches. Cantone et al.'s medical image analysis [54], shows that XAI methods like CAM, Grad-CAM, and Grad-CAM++ are deemed crucial for interpreting CNN decisions in mammogram classification. While the effectiveness of applying Grad-CAM on transformer architectures is still debated, attention mechanisms intrinsic to transformers support explanations through inspection of attention matrix weights, as demonstrated by Attention Rollout. Additionally, hierarchical transformers, like NesT, can benefit from specialized XAI methods such as GradCAT, which exploits architecture-specific features for enhanced interpretability.

Talaat et al. [82] proposed a novel explainable AI framework (BCaXAI) based on the Inception-ResNet V2 architecture for noninvasive breast cancer diagnosis using mammography images. The model incorporates Grad-CAM to provide visual explanations, enhancing trust and interpretability for radiologists. Evaluated on the DDSM and CBIS-DDSM datasets, BCaXAI obtained superior results with 98.53% accuracy, 98.53% recall, 98.40% precision, 98.43% F1-score, and an AUROC of 0.9933, outperforming traditional models like ResNet50 and VGG16. The approach demonstrates significant potential for reducing diagnostic subjectivity and minimizing unnecessary biopsy procedures.

Moreover, Saliency Maps can be developed for each explainable model and can calculate the entropy measure. Qualitative and quantitative evaluation of the XAI attribution maps can be generated in terms of robustness and localization of findings with

bounding boxes and Intersection over Union(IoU) scores [83]. IoU is a measure utilized to evaluate how well the bounding box generated by an XAI method aligns with the true region of interest. The bounding box is drawn around the most salient area interpreted by the XAI technique, and IoU calculates the overlap between the predicted bounding box and the actual area. It is defined as the ratio of the area of overlap to the area of their union. A higher IoU score indicates a better match between the XAI method's highlighted region and the true trigger [84]. Together, the IoU and Bounding box offer a comprehensive evaluation of how well the XAI method identifies key features, helping to assess its effectiveness in different interpretations.

A novel explainable deep learning approach was proposed for breast cancer detection using enhanced DenseNet architectures combined with advanced image preprocessing and fine-tuning strategies. Modifications, including BN-ReLU-Conv and Block-End layers, were applied to DenseNet169, resulting in high accuracy scores across multiple histopathology datasets such as BreakHis (40X, 100X, 200X, 400X) and BACH, with a peak accuracy of 99.50% on BreakHis 40X [85]. To improve clinical interpretability, Class Activation Maps (CAM) and Saliency Maps were employed, providing visual explanations of model decisions. While the approach showed excellent performance in controlled settings, future work aims to translate the method into real-world clinical practice.

In conclusion, visual explanation methods are essential for improving the interpretability and transparency of AI models. Techniques such as LIME, Grad-CAM, RISE, SHAP and Saliency Maps provide valuable insights into model decision-making, each offering unique strengths and limitations. While these methods have made significant contributions to the field of XAI, challenges remain in achieving accurate localization of important features, particularly for complex tasks. New approaches, such as SIDU, continue to address these gaps, enhancing the ability to generate clearer and more reliable explanations. As the field evolves, ongoing research and development will further improve the trustworthiness and usability of AI systems, making them more transparent and applicable to a wider range of sensitive domains.

Traditional post-hoc XAI methods, such as LIME and SHAP, generate explanations after model training, potentially misaligning with the model's internal decision-making processes. In contrast, Self-eXplainable AI (S-XAI) integrates explainability directly into the model's architecture during training. This approach enhances transparency and trustworthiness in medical applications. A comprehensive survey in [86] reviewed over 200 papers, categorizing S-XAI methods into input, model, and output explainability, and highlighted their applications across various medical imaging modalities In summary, the intersection of deep learning models and XAI holds great promise for the future of breast cancer detection. The incorporation of XAI techniques in some approaches not only enhances model interpretability but also contributes to the crucial aspect of interoperability in medical applications. By providing insights into

the decision-making process of these models, XAI techniques bridge the gap between advanced algorithms and clinical practitioners, improving trust and understanding.

Building upon the insights gained from the literature review, this study adopts DL and ViT architectures to improve detecting breast cancer. The literature review highlighted the importance of XAI methods to strengthen model interpretability, confirming that healthcare professionals can trust and follow the model's results. Using these findings, the following methodology outlines the data preparation, experiment training and model selection, and training strategies employed in this research to develop an efficient, accurate, and transparent diagnostic platform.

CHAPTER 3

METHODOLOGY

This section covers the details of the research's implementation, providing a comprehensive overview of the methodologies and techniques employed. It serves as a foundation for understanding how the various components integrate to achieve the study's objectives and highlights the systematic approach taken throughout the research process.

3.1 System Architecture

This chapter discusses the methodology used to develop an AI model for breast cancer classification. The methodology of this thesis is organized into three main segments: the implementation of the breast cancer classification models using CNN and ViT architectures, the application of XAI algorithms, and the evaluation process.

First, the thesis outlines the implementation of the classification model, detailing the data used, its selection and pre-processing, the model architecture, the training procedure, and the evaluation of the model.

Second, it discusses the implementation of the chosen XAI algorithms, highlighting both data and model explanation techniques. Lastly, the evaluation procedures for these implementations are described, emphasizing criteria such as runtime, simplicity, interpretability, and stability. These interconnected segments provide a comprehensive overview of the research process, each contributing to the overarching goal of enhancing the explanatory power of AI in breast cancer classification. The overall procedures are depicted in figure 3.1. Future chapters will discuss each component.

3.2 Dataset Description

Mammography is vital in breast cancer screening due to its ability to detect early-stage breast masses or suspicious regions. This study utilizes large-scale, routinely collected mammographic datasets to extract relevant information and identify explanatory features for various stages of breast cancer. Specifically, two publicly available datasets were used: CBIS-DDSM and RSNA Screening Mammography, chosen for their high-quality annotations and diversity in breast cancer cases.

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

   CBIS-DDSM is a well-established benchmark dataset in breast cancer diagnosis research. It is a curated subset of the broader DDSM dataset, providing high-resolution mammographic images annotated by trained mammographers. The

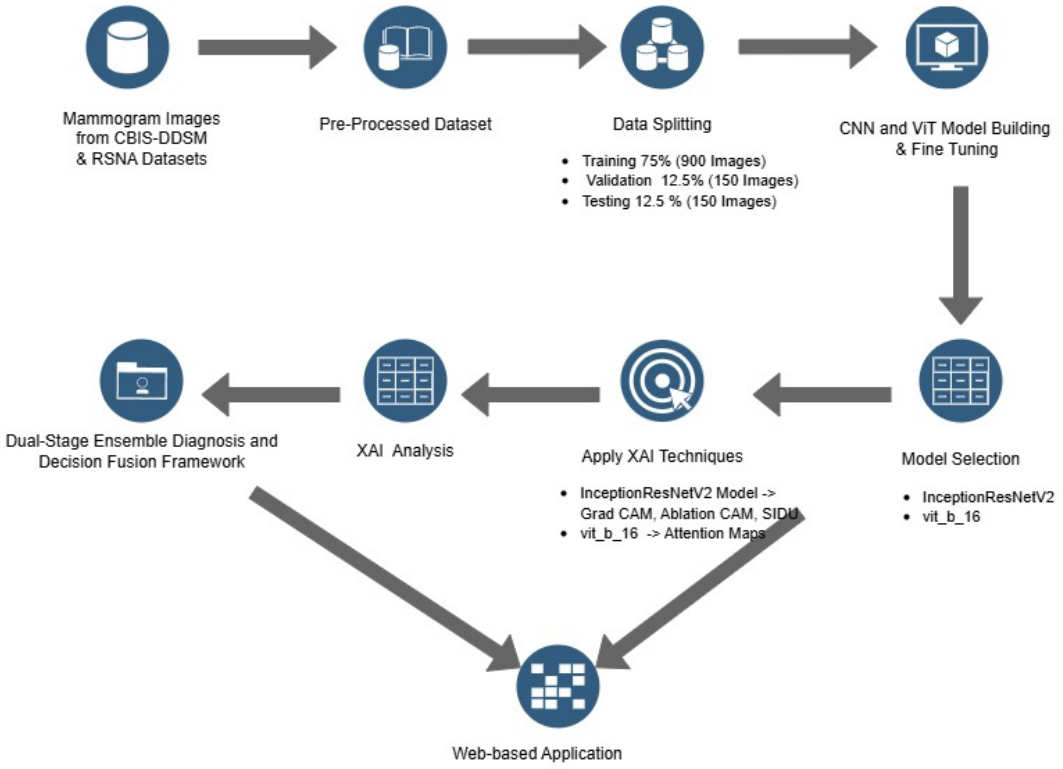

Fig. 3.1: System Architecture of the research

data set is available at the following link: CBIS-DDSM Dataset. A manuscript providing detailed instructions on how to use this dataset can be found at the following link Manuscript [87, 88]. CBIS-DDSM is a commonly utilized dataset in breast cancer diagnosis research and serves as a standard reference for assessing the effectiveness of computer-aided detection (CAD) systems. Its mammographic images are annotated with detailed information on breast abnormalities, including lesion type, size, and location [89].

Here is some information regarding this dataset:

- Number of images: 10,239
- Number of subjects: 6,671
- Total image size: 163.6 GB

The dataset includes:

- DICOM-formatted images (standard for medical imaging), including metadata like patient ID and image instance details.
- Binary masks highlighting regions of interest (ROIs) associated with abnormalities.

- ROI segmentations, bounding boxes, and detailed pathologic diagnoses (e.g., benign or malignant status, histological subtype, and tumor grade).
- High-quality annotations validated by medical experts to support robust CAD (computer-aided detection) system development.
- A user guide and supporting manuscript for the dataset are available at the official CBIS-DDSM portal.

2. RSNA Screening Mammography Breast Cancer Detection

   The RSNA dataset offers a large-scale collection of high-resolution screening mammograms with expert radiologist annotations. It enables the training of models across a variety of breast abnormalities and is particularly useful for binary classification tasks.

   Here is some information regarding this dataset:

   - Publicly available dataset with accurate labels for cancerous and non-cancerous cases at the following link: RSNA Dataset.
   - High interreader agreement and consistency due to expert curation.
   - Supports development of generalizable and clinically relevant breast cancer detection models.

The images collected from the datasets were initially categorized into two classes. The Breast Cancer class represents mammograms indicating the breast cancer, and the Non-Cancer class, includes mammograms showing no signs of breast cancer.With the datasets in place, the next step involves preparing the mammographic images to ensure they meet the quality and consistency requirements necessary for effective model training. The following section discusses the data pre-processing techniques employed to standardize the images, enhance relevant features, and eliminate irrelevant elements steps that are critical for improving the accuracy and interpretability of breast cancer detection models.

3.3 Data pre-Processing

Mammography images are critical in the early detection and diagnosis of breast cancer. Pre-processing is essential to enhance image quality, remove artifacts, and ensure that subsequent analysis algorithms can operate effectively. Proper pre-processing helps achieve more accurate diagnostic results and improved image analysis outcomes. Firstly, Noise reduction techniques such as Gaussian Blur and Non-local Means denoising were employed to improve the image clarity. These methods help minimize the random variations of brightness or color in the images that can affect the model's ability to identify important features.

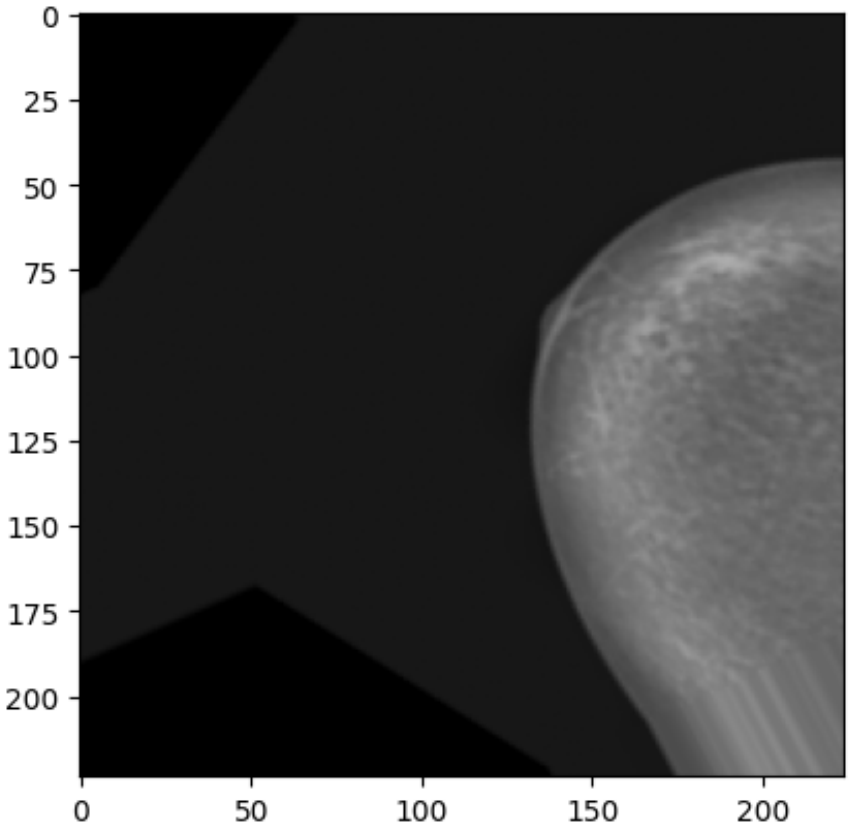

Fig. 3.2: A Mammogram image after applied transformations

Since most pre-trained models need consistent dimensions for optimal performance, the dataset's mammography images need to be resized to provide uniformity due to their height and width variations. Hence, the images were standardized in terms of size and pixel intensity. The dataset was organized and loaded using Keras's ImageDataGenerator, which facilitates automatic label assignment based on the directory structure. Each image was resized to a uniform target size of 229x229 pixels, ensuring consistency across the input dimensions required by the convolutional neural network. Image standardization ensures that the model is not biased towards images of specific dimensions or scales.

Moreover, improvement of the contrast of mammogram images was also performed as it is also an important step that emphasizes the structures within the mammogram, making it easier for deep-learning models to detect subtle signs of potential cancer. Furthermore, certain scans include white borders that may mislead the model into interpreting them as indicators of tumor presence also there can be markings on some images that must be eliminated, as they can introduce confusion regarding potential abnormalities. The lack of clear separation between the breast tissue and the background is another issue that requires solutions. To avoid these concerns, artifact removal was also performed to remove any artifacts from the images such as labels, markers, or

other non-tissue elements, which could potentially bias the model's learning process.

After the data cleaning stage, it consisted of 600 images per class, resulting in a total of 1200 images. This data set was organized using Python Split-folders library into separate directories for training, validation, and testing. The data was partitioned as follows:

1. Training Set(75%) - A total of 450 images per class(900 images in total) were allocated for training the model. This subset was used to teach the model to recognize the distinct features of breast cancer and non-cancer images, enabling it to learn the relationships between the input data and the target labels.

2. Validation Set(12.5%) - The next 75 images per class(150 images total) were allocated to the validation set. During the training process, this set was used to monitor the model's performance, helping adjust the hyperparameters and tune the model. It plays a key role in preventing overfitting, as it contains an unbiased evaluation of the model at various stages of training.

3. Testing Set(12.5%) - The remaining 75 images per class(150 images total) were reserved for the test set. This subset was not used during the training or validation process. This set is crucial for checking how well the trained model generalizes to new data.

Lastly, data augmentation techniques like rotation, flipping, and scaling were utilized to artificially expand the dataset. The ImageDataGenerator was utilized to perform advanced image augmentation and pre-processing. This is mainly useful in medical imaging, where the availability of labeled images is often limited. This technique helps in making deep learning models that are robust and can generalize well on new and unseen images. Augmentation techniques were applied to the training dataset to prevent overfitting and ensure robust model performance on unseen data.

- Rescaling: Normalizes pixel values to a uniform scale.

- Rotation: Introduces variability by rotating images to different angles.

- Shifting and Shear Transformation: Alters the alignment and angle to simulate variations in patient positioning.

- Zooming: Adjusts the scale of images to represent different levels of zoom found in clinical settings.

- Flipping and Brightness Adjustment: Reflects images and adjusts brightness to mimic variations in imaging techniques.

Following extensive data pre-processing to ensure clarity, consistency, and diversity in the input mammograms, the next phase focused on feature extraction and establishing a robust training framework. The upcoming section details the architecture, configuration, and strategies employed to extract features and train the model effectively.

3.4 Feature Extraction and Training Framework

3.4.1 Feature Extraction

The feature extraction phase is pivotal in the image analysis process, especially in the medical imaging domain. In recent years, the adoption of pre-trained CNNs has transformed how features are extracted from input image data. Unlike traditional methods that rely on handcrafted features such as texture, shape, and intensity, which demand extensive domain knowledge and often fail to capture complex patterns, pre-trained CNNs offer more robust and automated solutions from mammogram images.

The appearance of pre-trained convolutional neural networks (CNNs) has significantly changed the landscape of medical image analysis. These networks are initially trained on large, diverse datasets, enabling them to learn a wide range of features, from simple edges and textures to complex patterns that characterize different types of tissues. By employing transfer learning, these pre-trained models can adapt their learned features to specific tasks, such as mammogram analysis for breast cancer detection [90]. One of the key advantages of pre-trained CNNs is their hierarchical feature learning capability. These networks operate on a progressive basis, extracting low-level features in the initial layers and combining them into higher-level in subsequent layers. This allows the model to capture complex patterns that may indicate malignancy, which traditional methods might miss. Additionally, pre-trained CNNs exhibit robustness and generalization, making them less prone to overfitting, particularly when applied to limited or imbalanced medical datasets [91, 92].

3.4.2 Training Configuration Settings

Parameters like batch size, image size, and epochs play a crucial role in determining how effectively the model learns from the data and generalizes to unseen examples. An image size of 294x294 pixels was selected to optimize data processing through the network while minimizing the extra memory usage and computational time. The batch size of 16 is selected to balance training stability. Moreover, the training is run for 50 epochs, which means the entire dataset will pass through the model 50 times. Batch size and number of epochs were chosen to obtain the fullest learning outcome and prevent overfitting.

3.4.3 Model Setups

Efficient batch processing was enabled by loading data using generators to load data directly from the organized directories. This approach significantly optimizes memory usage, as it allows the model to load and process in smaller batches rather than loading the entire dataset into memory at once. In this study, various CNN architectures such as VGG, ResNet, DenseNet121, InceptionV3, and MobileNetV2 models were utilized for model building and these were chosen for their ability to effectively capture local dependencies and hierarchical features crucial for image-based classification. Each model has been slightly modified to suit the specific needs of binary classification. Here is a detailed overview of the architectures choses for this comparison

1. VGG16 Model

   VGG16 occupies a prominent position in image classification tasks. Originally VGG16 was introduced in the influential paper "Very Deep Convolutional Networks for Large-Scale Image Recognition" by Simonyan and Zisserman in 2014 [93]. It was developed for the ILSVRC 2014 challenge, by the Visual Geometry Group (VGG) at the University of Oxford and it achieved a top-5 accuracy of 92.7% on the ImageNet dataset. This remarkable performance highlights its effectiveness in handling complex image classification tasks. Despite its large size, which results in a slow training process, VGG16 is frequently utilized for transfer learning due to its flexibility and robust feature extraction capabilities. VGG16 is characterized by its architecture, which consists of sixteen convolutional layers followed by three fully connected layers.

   VGG16, recognized for its effectiveness in image classification, is a valuable asset for extracting relevant features from mammogram images. Its architectural features—such as consistent channel sizes, ReLU activation functions, and max-pooling layers—contribute to its ability to capture intricate image characteristics. Additionally, the progressive increase in the number of channels enhances its capacity to discern complex patterns. The standard architecture of VGG16 is illustrated in figure 3.3 [94].

   In summary, this VGG16 model consists of 16 layers, including 13 convolutional layers and 3 dense layers. For this, the top dense layers are excluded, and instead, a flattern layer is added to convert the 2D feature maps into a 1D vector. To suit the task of binary classification, a custom Dense layer with 2 output units is included at the end. This final dense layer is responsible for producing two output values, representing the two classes cancer or non-cancerous. This modified VGG16 architecture allows the model to effectively handle binary classification tasks while maintaining the essential features of the original VGG16 design.

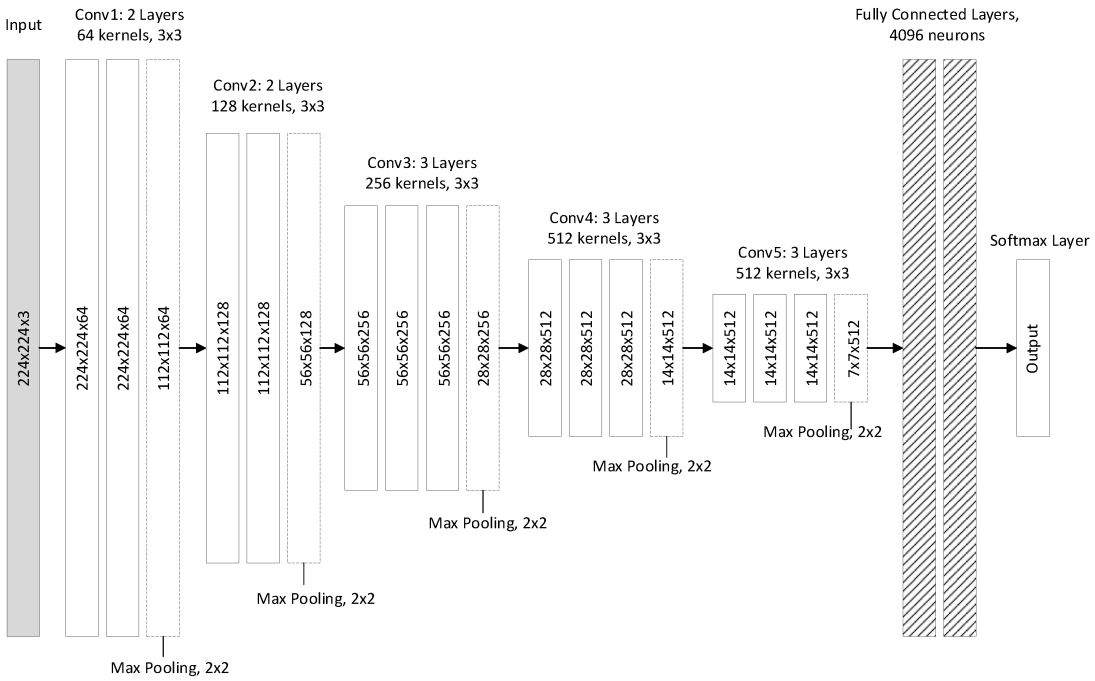

Fig. 3.3: General structure of the VGG16 network.

2. VGG19 Model

Originating from the Visual Geometry Group at the University of Oxford, the VGG19 model is a deep convolutional neural network comprising 19 layers including 16 convolutional layers and 3 dense layers. The architecture follows a straightforward and repetitive pattern, making it easier to understand and implement. The standard architecture of VGG19 is illustrated in figure 3.4 [95].

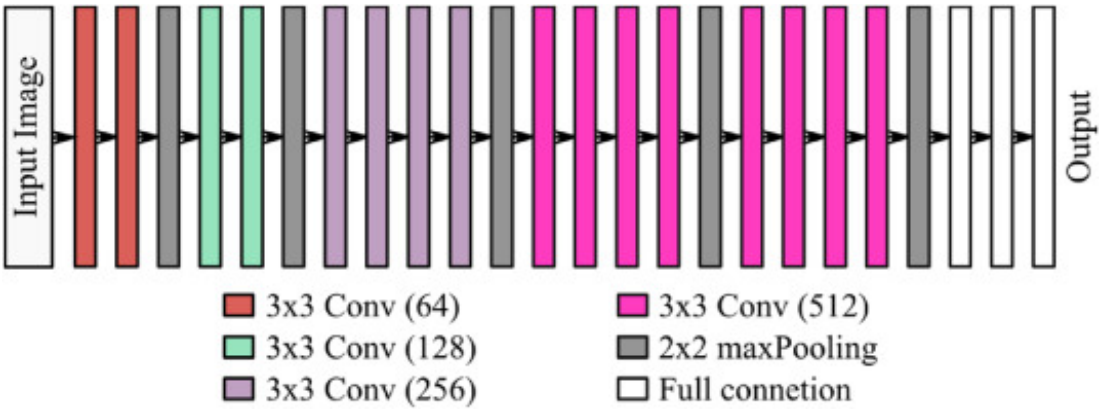

Fig. 3.4: General structure of the VGG19 network.

For our purposes, the top dense layers of the original VGG19 architecture are excluded and replaced with a flatten layer followed by a custom Dense layer with 2

output units. This modification is specifically designed for binary classification, where the model is tasked with classifying images into two categories, such as cancerous and non-cancerous. The Flatten layer converts the 2D feature maps into a 1D vector, making it suitable for input into the custom Dense layer. The final Dense layer outputs two values, each representing the probability of one of the two classes, with a softmax activation function typically applied to produce the final class probabilities. This adaptation allows the model to effectively perform binary classification while retaining the core strengths of the original VGG19 architecture.

3. ResNet101 Model

   A part of the Residual Network (ResNet) family developed by Microsoft, the ResNet101 model is a highly robust deep neural network comprising 101 layers. ResNet101 is known for its use of residual connections, which allow the model to bypass certain layers, helping to prevent the vanishing gradient problem and enabling the network to train deeper architectures effectively.

   For our purposes, the original top dense layer of ResNet101 was excluded and replaced with a flatten layer along with custom dense layer with two output units for binary classification.

4. ResNet50 Model

   ResNet, or Residual Network, is a convolutional neural network with 50 layers developed by He et al. for the ILSVRC 2015 competition, where it achieved first place [96]. It is widely regarded as one of the leading state-of-the-art models due to its innovative use of residual layers, which allow certain layers to be bypassed, effectively mitigating the vanishing gradient problem. This architecture consists of 50 layers, incorporating convolutional layers, pooling layers, fully connected layers, and shortcut connections that facilitate the bypassing of one or more layers. These shortcut connections are crucial for overcoming the vanishing gradient issue often encountered when training very deep networks.

   The design of ResNet50 is organized into four stages, each containing multiple residual blocks. Each block includes two or three convolutional layers, along with batch normalization and ReLU activation functions, complemented by shortcut connections that merge the input and output of the block. The first stage utilizes 64 filters, the second employs 128 filters, the third uses 256 filters, and the fourth stage has 512 filters. The output from the final residual block is directed to a global average pooling layer, which computes the average of the feature maps across the spatial dimensions. This produces a feature vector that is subsequently processed by a fully connected layer with SoftMax activation, generating the final class probabilities[97, 98].

In this implementation, the pre-trained ResNet50 model was utilized for a binary classification task involving breast cancer diagnosis. As a small variant of the ResNet101, it mirrors the modifications made in ResNet101, excluding the top layer and adding a flatten layer and a customer Dense layer with two output units for binary classification.

5. MobileNetV2 Model

   MobileNetV2 is a deep learning model architecture designed for efficient performance on mobile and edge devices. It is a lightweight, efficient convolutional neural network (CNN) model that uses depthwise separable convolutions to reduce the computational cost and number of parameters. The modification for this model are consistent with the others, featuring excluded top layers replaced by flattern layer and a custom dense layer with two putput units for binary classification.

6. InceptionV3 Model

   The InceptionV3 model developed by google, the model features an architecture with Inception modules, which allow the network to learn local features efficiently at multiple scales. InceptionV3 requires a larger input shape of (299, 299, 3) to accommodate the complexity of its network. This input shape ensures that the model can process high-resolution images, capturing fine details that are essential for accurate predictions. Similar to the other models, its top layer was replaced with a flatten layer and a custom Dense layer with 2 output units for binary classification.

7. DenseNet121 Model

   DenseNet is a deep learning architecture designed for image classification and object detection tasks. Its distinctive feature lies in the incorporation of outputs from all preceding layers as inputs to each subsequent layer. This dense connectivity allows for feature reuse and reduces the number of parameters required to learn various features by giving each layer direct access to the feature maps of all previous layers.DenseNet can serve as a feature extractor within prototypical network architectures. By utilizing a pre-trained DenseNet model, features can be extracted from images in both the support and query sets, producing a high-dimensional feature vector that effectively represents each image[99, 100]. The standard architecture of DenseNet is illustrated in figure 3.5 [99].

   In this study, pre-trained DenseNet121 model was utilized as a backbone for a binary classification task aimed at diagnosing breast cancer. DenseNet121 connects each layer to every other layer in a feed-forward fashion, totaling 121 layers. It also includes a configuration with excluded top layers, supplemented

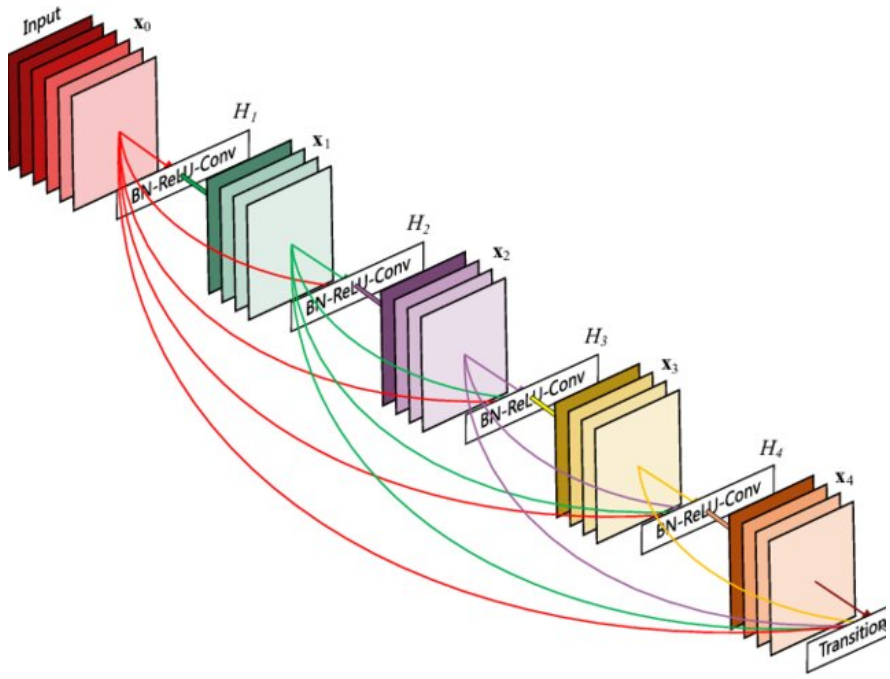

Fig. 3.5: General structure of the DenseNet Architecture

by a Flatten layer and a custom Dense layer with 2 output units for binary classification.

Each of these models were selected because of their ability to process complex image data effectively. The modifications made to each model are designed specifically to address the challenge of binary classification in mammogram images, to achieve high accuracy in distinguishing between cancerous and non-cancerous findings. This structured approach, involving multiple deep learning architectures, enables a thorough comparison of their performance, ensuring that the best-suited model was identified for this critical medical application.

After selecting and customizing a diverse range of CNN architectures for binary classification, the next logical step was to train and evaluate their performance on the prepared mammogram dataset. To determine the most effective architecture for this specific medical imaging task, a series of structured experiments were conducted. The following section outlines the training procedures, comparative analysis, and optimization strategies applied to select the best-performing model for breast cancer detection.

3.5 Experiment Training and Model Selection

Two main experiments were carried out initially to develop the base training models for classifying mammogram images into cancerous and non-cancerous classes. These experiments were designed to assess the effectiveness of different deep learning ar-

chitectures for binary classification and to establish a strong foundation for further optimization.

- Experiment 01

In experiment 01, initial training sessions were conducted using VGG16, VGG19, MobileNetV2, InceptionResNetV2, ResNet50, ResNet101, DenseNet121, and a custom CNN model. Each model was trained for 50 epochs on the prepared mammography dataset to evaluate their baseline performance in classifying images into cancerous and non-cancerous categories.

- Experiment 02

In Experiment 02, following the results of Experiment 01, the top-performing model was selected for further optimization and training. This step aimed to fine-tune the chosen model to improve its accuracy.

Among the models, the InceptionResNetV2 was selected as the best accurate model, and fine-tuning and optimization techniques were used to enhance its performance even more. The InceptionResNetV2 model was selected for its unique architecture that integrates both Inception and ResNet principles. The Inception module's multi-scale feature extraction combined with the ResNet's skip connections ensures improved gradient flow, allowing the model to learn deeper patterns without encountering vanishing gradient issues. These advantages make InceptionResNetV2 particularly effective in complex image data such as mammograms, where distinguishing between subtle cancerous anomalies and healthy tissue is crucial. Initial experiments demonstrated superior precision, recall, and overall accuracy with InceptionResNetV2, which further justified its selection for fine-tuning in Experiment 2.

Several optimizations were applied to the InceptionResNetV2 model:

1. The last 30 layers of the original InceptionResNetV2 were excluded to reduce model complexity and improve generalization.

2. A Dropout layer with a rate of 0.5 was added after the selected output layer from the deeper sections of the model to minimize overfitting by randomly deactivating neurons, which can help improve the model's ability to generalize to new, unseen data.

3. Learning rate adjustments and augmentation techniques such as rotation, flipping, and zooming were applied to improve performance across diverse image conditions.

To better understand the contribution of individual fine-tuning strategies applied to the InceptionResNetV2 model, a limited ablation study was conducted. Several experiments were performed where key modifications were selectively removed to evaluate their impact on model performance. Specifically, the study assessed the effects of (i) retaining all original layers (without removing the last 30 layers), (ii) omitting the dropout layer, and (iii) disabling data augmentation during training. Performance was measured using accuracy, precision, recall, and F1-score metrics on the validation set. Results indicated that removing the last 30 layers and adding a dropout layer significantly improved model generalization and reduced overfitting. Data augmentation further enhanced robustness to variations in mammogram images. The findings confirm that each fine-tuning step played a crucial role in optimizing the model for breast cancer detection from mammograms.

The model compilation, the Adam optimizer, was selected due to its efficiency in handling sparse gradients and adaptable learning rates. It was configured with a learning rate of 0.001 and an epsilon value of 0.1 to stabilize performance in the later stages of training. Categorical cross-entropy was chosen as the loss function, focusing on two classes. Accuracy was used as the primary performance metric to track how well the model classified the images. To prevent overfitting, an early stopping callback was implemented, which monitors the validation loss and stops training if no improvement is seen for 10 consecutive epochs. The model's best-performing weights are restored if training halts early.

In the training and validation process, the step sizes for training and validation were calculated based on the number of samples and batch size. The step size for training was determined by dividing the total number of training samples by the batch size, while the step size for validation was calculated similarly. The model was trained using the fit method, with the `train_generator` giving training data and the `valid_generator`' for supplying validation data. The model was set to train for up to 50 epochs, with early stopping ensuring that training would terminate when improvements stopped, further safeguarding against overfitting.

By implementing these strategies and improvements, the optimized InceptionResNetV2 model achieved improved stability, generalization, and accuracy in detecting breast cancer from mammogram images.

While the initial experiments focused on evaluating and optimizing CNN-based architectures for mammogram classification, recent advancements in deep learning have introduced alternative paradigms capable of capturing complex image relationships more effectively. To further enhance diagnostic performance and explore the potential of attention-based mechanisms, this study also investigated the use of ViT models. The following section outlines the implementation and training process of ViT-based architecture, offering a complementary approach to traditional convolutional models through their ability to model global dependencies within medical images.

3.6 Transformer Model-Building

The Vision Transformer (ViT) is a state-of-the-art deep learning architecture specifically designed for image classification tasks. By replacing conventional convolutional layers with self-attention layers, the ViT architecture allows for more flexible and efficient processing of image patches. For this study, pre-trained models from the DeiT family of ViT models were employed, consisting of a stack of transformer blocks featuring self-attention and feedforward layers. Data pre-processing for the `vit_b_16`

```
def create_model():
    vit_model = vit.vit_b16(
        image_size=IMAGE_SIZE,
        activation='softmax',
        pretrained=True,
        include_top=False,
        pretrained_top=False)

    model = tf.keras.Sequential([
        vit_model,
        Flatten(),
        Dense(2, activation='softmax')
    ])

    model.compile(optimizer=Adam(learning_rate=LEARNING_RATE),
                  loss=tf.keras.losses.CategoricalCrossentropy(),
                  metrics=[tf.keras.metrics.CategoricalAccuracy()])

    return model

model = create_model()
```

Fig. 3.6: `vit_b_16` Model

model involved resizing images to 224x224 pixels and converting them to tensors. The output layer was configured as a softmax activation function to handle binary classification. The top classification layer was excluded to allow customization of the output layer. A Flatten layer was added to convert the ViT output into a 1D array that feeds into a Dense layer. The final Dense layer with two units and a softmax activation function was used for binary classification. The model was trained using the `fit` method following `train_generator` and `validation_generator` to supply the training and validation data, respectively

During model training, the categorical cross-entropy loss function and the Adam optimizer were employed for further optimization. Adam is preferred for its efficiency in handling sparse gradients and its adaptive learning rate capabilities. The model was set to train for up to 50 epochs, although training could stop earlier if the early stopping condition was met, ensuring that the model doesn't overfit and optimizing

computational resources. Early stopping mechanism monitors the validation loss and halts training if there is no improvement for 10 consecutive epochs. It also restores the weights of the model to those of the epoch with the best performance.

The use of ViT in this context is part of an ongoing effort to enhance breast cancer detection through advanced AI techniques. The ViT model was integrated into this classification pipeline to enhance diagnostic performance using its attention-based mechanism.

Once the transformer-based ViT model and CNN models trained, it became essential to assess how well each architecture generalized to unseen data. While model training focuses on minimizing loss and improving prediction during learning, it is the validation and testing phase that provides a true measure of the model's effectiveness in real-world scenarios. The following section presents the performance metrics used to test and compare all trained models, ensuring a fair assessment of their diagnostic capabilities in breast cancer detection.

3.7 Validation and Testing

Validation plays a critical role in verifying that the model behaves as intended. This step involves evaluating the trained model on data it hasn't seen before to assess how well it performs. Common evaluation metrics include accuracy, precision, recall, F1 score, and AUC-ROC. Precision indicates how many of the model's positive predictions are actually correct, while recall reflects its effectiveness in identifying all true positive cases. The F1 score provides a balanced assessment by calculating the harmonic mean of precision and recall. Accuracy measures the proportion of correct predictions across all classes, making it a widely used metric for classification tasks. AUC-ROC evaluates the performance of a binary classifier by examining the trade-off between the true positive rate and false positive rate across various threshold settings.

- (Tp) = Number of times the model predicts the positive class correctly.
- (Tn) = Number of times the model predicts the negative class correctly.
- (Fp) = Number of times the model predicts the positive class incorrectly.
- (Fn) = Number of times the model predicts the negative class incorrectly.

Precision, recall, accuracy and F1 score can be expressed as follows.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.2}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

$$F\text{-Score} = \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

Through extensive experiments and comparison, this study evaluated CNN-based models and ViT-based models for image classification, using a variety of performance metrics, including accuracy, precision, recall, and F1-score. Based on the results of these evaluations, the best-performing model was identified, highlighting its strengths over the others. This analysis reveals the current state of the art in image classification and underscores the evolving landscape as newer architectures challenge traditional methods.

3.8 Applying XAI Technique

After selecting the best-performing model from both the CNN and ViT model validation, XAI techniques were applied to interpret the model's predictions, offering insights into how the models make their decisions. These XAI techniques were applied separately to the top-performing CNN model, InceptionResNetV2, and the textttvit_b_16 model, both of which demonstrated exceptional results.

3.8.1 Applying XAI Techniques for Vision Transformers (ViT) Model

Attention Maps in Vision Transformers provide a visualization of where the model is focusing within the image during the attention process. Each attention head in a transformer layer can attend to different parts of the image, making these maps useful for understanding which parts of the image are considered important for the model's predictions. This capability can be particularly insightful for tasks like image classification, object detection, and segmentation, where knowing 'why' a model made a certain decision can be as crucial as the decision itself.

Figure 3.7 depicts the process of the Attention Map generation for the ViT model. Firstly, the each images were converted to RGB format to maintain consistency in color representation, which is crucial for accurate image analysis. They were resized to a specific dimension 224x224 pixels, a requirement for the input layer of the Vision Transformer, ensuring that all images are treated equally by the model. Next, Vision Transformer Model Setup, the model processes the images by dividing them into smaller patches, much like how sentences are broken into words in natural language processing. This segmentation allows the ViT model to treat each part of the image individually, analogous to tokenization in NLP. After patch creation, each patch was embedded into a higher-dimensional space, and positional encodings were added.

These encodings provide crucial spatial context, helping the model understand the relative positions of patches within the image and preserving the spatial hierarchy of features.

Moving to next step, Transformer's Attention Mechanism, the self-attention mechanism within the transformer model calculates how each patch influences the representation of other patches. For each patch, a set of attention scores was computed to indicate its impact on every other patch, allowing the model to focus selectively on the most informative parts of the image. These attention scores were aggregated across multiple heads, enabling the model to capture diverse features from various positions, thus enriching the overall feature representation.

After the image passes through the transformer, the attention weights from the layers was extracted. These weights reflect the importance assigned to each patch in the decision-making process, providing insight into the model's focus areas. The extracted weights were then scaled back to the original image dimensions and converted into a heatmap. This heatmap visually represents regions of high and low focus, overlaying it on the original image to give a clear representation of the model's attention distribution.

Finally, the original image and its attention heatmap were displayed side-by-side. This comparative display allows for easy assessment of the model's interpretive accuracy, as users can directly compare the actual visual elements with the areas the model prioritized. By examining the focus areas, users can determine whether the model is attending to the correct features, which is particularly important in critical scenarios, such as medical diagnoses, where the model must focus on clinically relevant features rather than irrelevant background noise.

3.8.2 Applying XAI Techniques for CNN-based Model

Out of the many available XAI techniques, this research focused on three methods, Grad-CAM, Ablation-CAM, and SIDU method. These techniques were chosen because they are particularly good at highlighting the key areas of input data that impact the model's predictions. They were used to evaluate the performance of XAI on the top-performing CNN model, InceptionResNetV2, as they provide clear and accurate explanations of the model's decision-making process.

- **Applying Gradient-weighted Class Activation Mapping (Grad-CAM) Method**

This technique is particularly valuable in identifying which parts of the image the model focuses on when making predictions, thus improving the transparency of the model.The first step in using Grad-CAM is model and layer selection. In this case, the InceptionResNetV2 model, developed in Experiment 2, was used. Instead of using the final convolutional layer, the last batch normalization layer was selected for

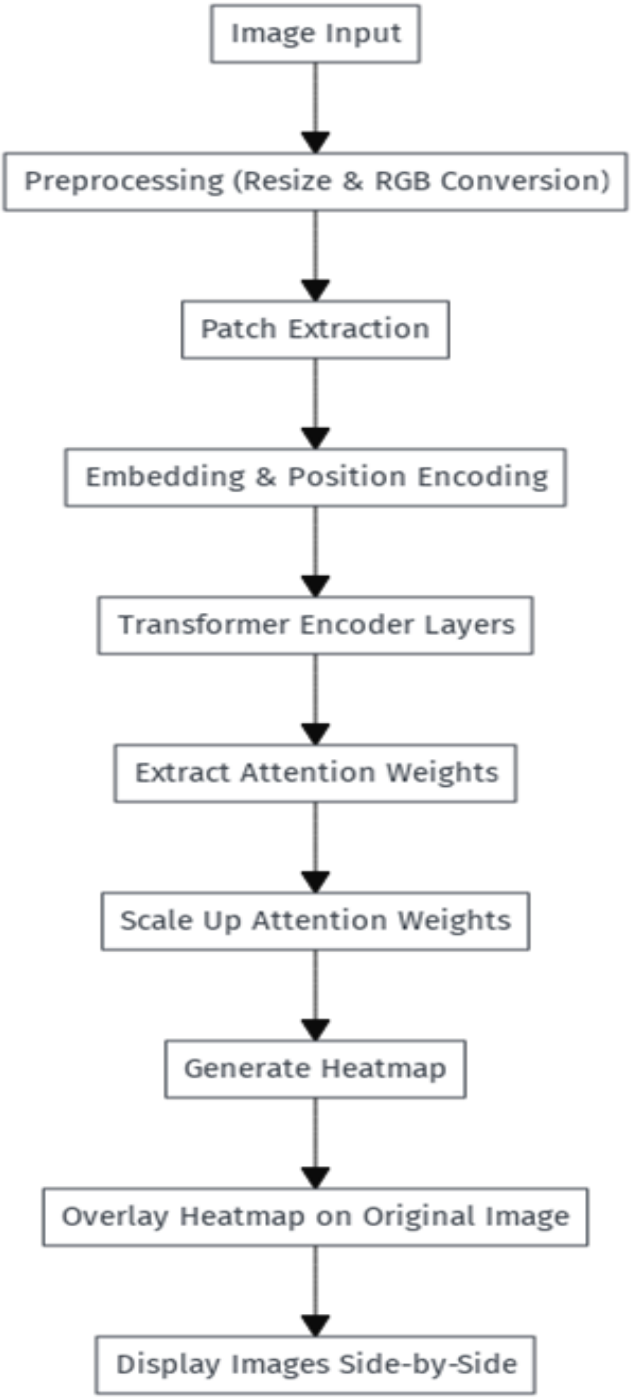

Fig. 3.7: Workflow of Attention Map Generation for ViT Model

analysis. Batch normalization layers were well-suited for this task as they normalized intermediate activations while preserving spatial information. This characteristic made them ideal for generating class activation maps, as they maintained the spatial context necessary for identifying important regions in the input image.

Once the model and layer were selected, the next step was the forward pass. The input image was passed through the trained InceptionResNetV2 model, which generated the output predictions. The target class was then chosen for analysis, which could either be the predicted class or a user-defined class. This step ensured that Grad-CAM focused on the relevant class, providing a clear understanding of the model's decision process.

After the forward pass, gradients were computed during the backward pass by calculating the derivatives of the target class score with respect to the activations in the chosen layer. These gradients indicated the significance of each activation for the target class, showing which features most strongly impacted the model's prediction.

Next, the feature maps were combined using weights derived from the gradients. These gradient-based weights reflected the relative importance of each feature map and were applied to the corresponding activations in the selected layer. The result was a class activation map that emphasized the areas of the input image most relevant to the model's prediction, offering a clear visual interpretation of the influential regions.

Following the creation of the class activation map, heatmap generation was performed. The class activation map was scaled to align with the size of the original input image. A colormap was applied to convert the activation values into a heatmap, making the regions of interest more visually interpretable. This heatmap served as a clear visual representation of the areas the model considered most important for its decision.

Finally, the overlay and visualization step was carried out. The heatmap was overlaid onto the original input image. This visual representation highlighted the areas the model concentrated on during its prediction, offering a clearer and more intuitive insight into the decision-making process and emphasizing the key regions that contributed to the outcome. Through this method, Grad-CAM provided an effective and interpretable visualization of the model's attention, enhancing the transparency of complex CNN models like InceptionResNetV2.

- **Ablation based Class Activation Mapping (Ablation CAM)**

Ablation-CAM is a refinement over traditional visualization techniques like Grad-CAM, designed to offer more accurate and focused interpretations of a neural network's predictions. Instead of relying only on gradient information, Ablation-CAM systematically removes parts of the model's feature maps during the computation to assess their contribution to the output. Similarly to Grad-CAM, Ablation-CAM was applied on the InceptionResNetV2 model that was developed in Experiment 2. Moreover, last batch normalization layer was selected to retain both spatial and feature in-

formation, which were considered essential for accurate localization in the generated visualizations.

Ablation-CAM operated by systematically ablating portions of the feature maps in the chosen layer. The effect of each ablation was then observed on the output score for the target class. The influence of each feature map on the final prediction was determined by systematically evaluating them one by one. Change in the target class score, caused by each ablation, was used to quantify the importance of the corresponding feature map. This step ensured that only the most critical activations were taken into account, thereby allowing for more precise localization of the relevant regions.

Following this, the calculated importance of each feature map was used as a weight to combine the maps. This weighted combination resulted in the creation of the ablation activation map, which highlighted the areas in the input image that had the strongest influence on the model's prediction. Subsequently, the ablation activation map was adjusted to align with the dimentions of the input image. A colormap was then applied to create a visually interpretable heatmap, similar to Grad-CAM, which allowed for better visual understanding of the areas of importance.

Finally, the overlay effectively highlighted the key regions influencing the model's predictions, offering a clearer and more interpretable view of its decision-making focus.

- **Applying Similarity Difference and Uniqueness (SIDU) Method**

SIDU method is an advanced approach in computer vision, designed to identify important pixels in an image using deep learning models. It works by analyzing the last convolutional layers of CNNs to provide a clear visual explanation of the model's predictions.The implementation of this method is categorized into three main steps: the generation of feature activation image masks, the computation of feature importance weights, and the comprehensive visual explanations.

1. Generation of Feature Activation Image Masks

   This step is the core of the SIDU method, where feature activation image masks are generated. It begins by extracting feature activations from the final convolution layer of the selected InceptionResNetV2 model. The figure 3.8 illustrates the steps involved in this process. This layer typically has dimensions n×n×N, where n represented the spatial size and N denoted the number of feature activations. Each activation map, representing the response of a specific feature across the input image, was transformed into a binary mask. The binary mask was then resized using bilinear interpolation to align with the original image's dimensions and combined with the input image via point-wise multiplication. This procedure not only illustrated the localized impact of each feature but also established the groundwork for deeper analysis in subsequent steps.

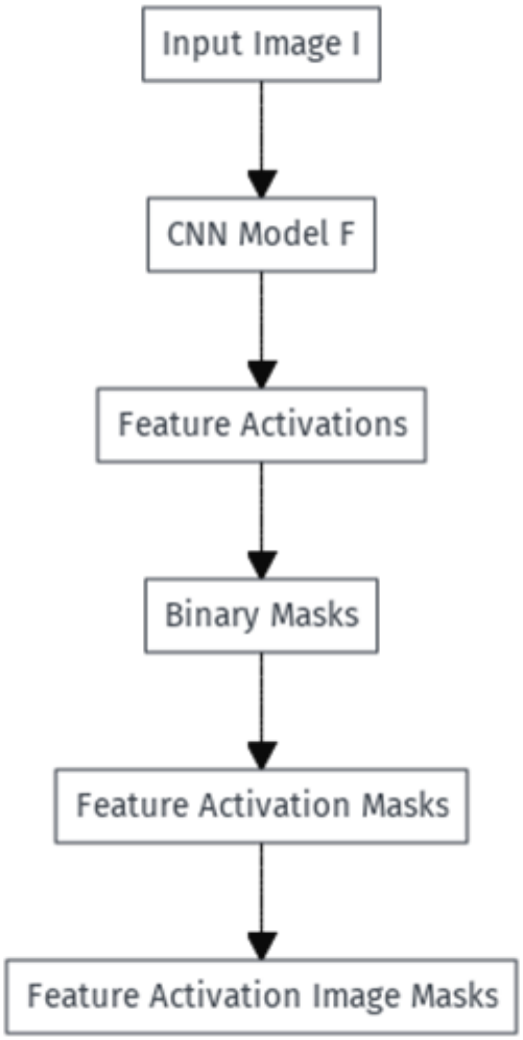

Fig. 3.8: Workflow of the Generation of Feature Activation Image Masks Step

2. Computing Feature Importance Weights

   The analytical core of the SIDU method involved computing the feature importance weights. This was achieved by first assessing the prediction scores assigned to each feature activation image mask, which were computed using the same InceptionResNetV2 model. The figure 3.9 illustrates the steps involved in this process. The most fundamental part of this step is the evaluation of each feature's contribution to the final prediction through a dual analysis of similarity differences and uniqueness. The similarity differences measured the deviation of each feature's prediction impact compared to the original image's overall prediction, highlighting features that significantly altered the prediction outcome. At the same time, the uniqueness measure identified features that distinctly affected the model's output, emphasizing their salient characteristics. The resultant feature importance weights were derived by integrating these two metrics, which

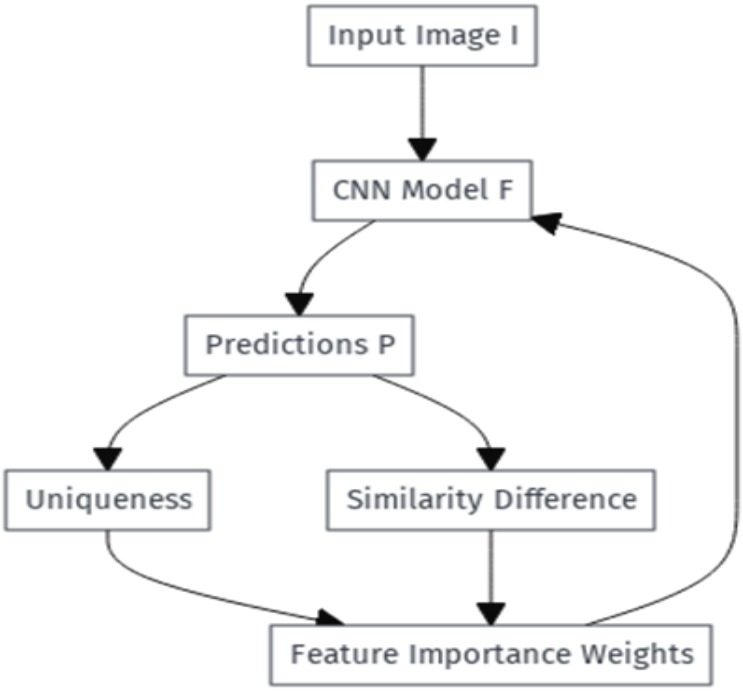

Fig. 3.9: Workflow of the Computing Feature Importance Weights Method

quantitatively reflected each feature's relevance and influence on the predicted result.

3. Visual Explanations for the Prediction

   The final step of the SIDU method is the generation of a visual explanation for the model's outcome. This is achieved by combining the weighted feature activation masks into a clear saliency map. In this process, each feature's weight, which shows its importance, adjusts the corresponding mask, leading to the creation of a heatmap. This heatmap highlights the features that most relevent for the model's decision-making. The saliency map not only helps in understanding the functionality of complex models but also improves the transparency and trustworthiness of automated decision-making systems, especially in critical applications.

Following the discussion of the three methods, Figure 3.10, presents a visual comparison of saliency maps generated by Grad-CAM, Ablation-CAM, and SIDU for cancerous images predicted by the InceptionResNetV2 CNN model highlights the differences in how each method emphasizes critical areas of the image. These output illustrates the different ways each technique highlights critical regions within the images, providing insights into the model's decision-making process and the strengths of each XAI approach in visualizing the most relevant features for prediction.

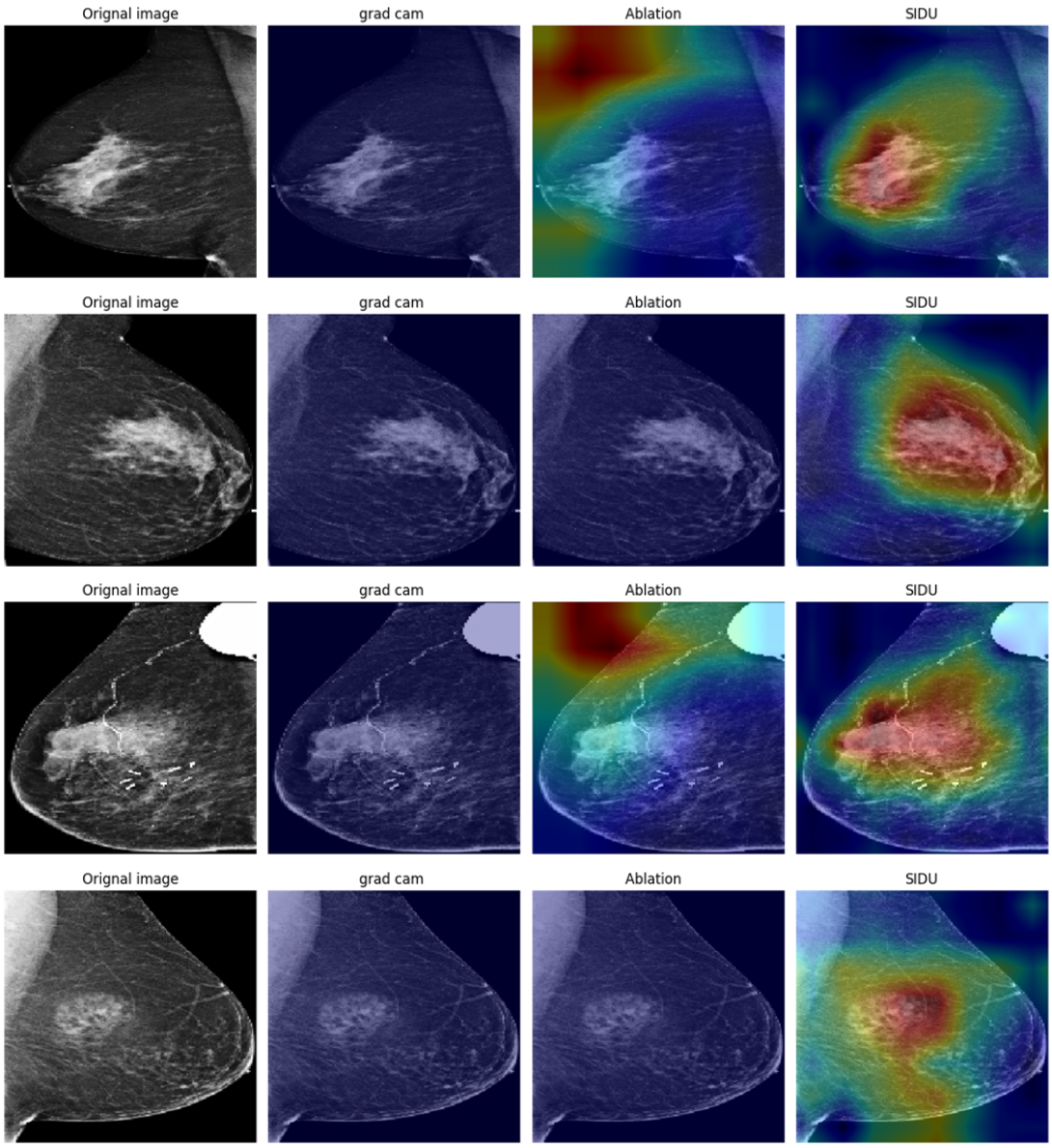

Fig. 3.10: Visual comparison of saliency maps generated from Grad-CAM, Ablation-CAM and SIDU for cancerous images on InceptionResNetV2 model.

3.9 Interpretability Analysis

Interpretability analysis in XAI is a main step in recognizing and assessing how well an AI model communicates its decision-making process. This analysis aims to ensure that the explanations provided by XAI methods are coherent, relevant, and effectively convey the underlying reasoning to healthcare professionals and end-users. To evaluate the effectiveness of XAI approaches, a robust IoU computation framework was used. The evaluation process was carried out through several key steps to ensure an accurate and comprehensive analysis of the saliency maps produced by different interpretability techniques.

The first step involved aligning the ground truth and prediction. The original image was annotated with a bounding box that indicated the region of interest (ROI). From this annotation, a corresponding ground truth mask was created, which represented the binary segmentation of the ROI. This ground truth mask served as the reference for evaluating the performance of the saliency methods in identifying the relevant areas. An independent review was conducted by an external expert in breast cancer detection using mammography to ensure the accuracy and clinical validity of the ground truth bounding boxes used in this study. The expert carefully redefined these bounding boxes to more precisely align with the actual lesion boundaries. The corrected annotations, which have been incorporated into the final dataset, provide a more accurate representation of the ROI.

All other ground truth annotations were confirmed to be satisfactory and did not require modification. The reviewing expert also recommended that for future studies, polygon-based ground truth annotations should be considered, as they would provide higher precision and clinical relevance, particularly in complex mammographic regions where lesion boundaries are irregular.

This validation process enhances the reliability of the dataset and ensures that the subsequent analysis and model evaluations are based on accurate and clinically acceptable annotations.

Next, heatmaps were generated using different feature selection techniques such as Grad-CAM, Ablation-CAM, and SIDU. These saliency maps were created by applying the respective methods to the original image, and each map highlighted different regions that were most influential in the model's decision-making process. To visualize these highlighted areas, the generated heatmaps were displayed onto the original image, showcasing the regions identified by each saliency method.

Following, predicted masks were generated from the saliency maps. Each heatmap was binarized by applying a threshold to emphasize the salient regions. This binarization converted the heatmaps into predicted masks, allowing for a direct comparison among the predicted and ground truth regions. The purpose of this step was to transform the saliency maps into a form that could be quantitatively analyzed.

The effectiveness of the saliency feature selection was then assessed using IoU calculation 3.5. The IoU was computed using the formula, which measured the overlap between the predicted mask and the ground truth mask. A higher IoU score indicated a better alignment between the predicted regions and the true regions of interest, reflecting the accuracy of each saliency method in highlighting the correct features.

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|Ground\,Truth\;Mask \cap Predicted\;Mask|}{|Ground\,Truth\;Mask \cup Predicted\;Mask|} \tag{3.5}$$

Finally, the entire process was visualized to provide a clear, intuitive understanding of the results. This visualization allowed for an easy interpretation of how well each saliency method performed in identifying the key regions that influenced the model's predictions.

Following the application of XAI techniques to both CNN-based and ViT-based model enhancing model transparency through saliency and attention maps it becomes essential to not only interpret individual model predictions but also to improve diagnostic robustness through model collaboration. While each model independently offers valuable insights, relying on a single architecture may introduce limitations in uncertain or borderline cases. To address this, the next section introduces a Dual-Stage Ensemble Diagnosis and Decision Fusion Framework, which strategically combines the outputs of both models to deliver a more reliable and clinically meaningful final diagnosis.

3.10 Dual-Stage Ensemble Diagnosis and Decision Framework

This section outlines the logic used to combine predictions from the best-performing deep learning models InceptionResNetV2 and ViT for a binary classification problem, where the objective was to determine whether cancer was present or not. Each model generates a class label along with a numeric confidence score. This ensemble fusion method improves diagnostic robustness and mitigates the risks associated with relying solely on one model's output. The logic accounts for both agreement and disagreement between models, incorporates confidence thresholds, and uses weighted score fusion to prioritize more reliable decisions. The function takes the following inputs:

- **Predicted class from each model M_i:** The categorical class output by the i-th model, either *"Cancer"* or *"No Cancer"*, where $i = 1, 2, ..., N$.
- **Confidence score from each model S_i:** A numerical score between 0 and 100 indicating the i-th model's confidence in its prediction.
- **Confidence threshold (T):** A user-defined minimum confidence level (typically set to 70%). Predictions below this threshold are considered less reliable.

- **Trust weight for each model W_i:** A numeric value representing the trust level assigned to the i-th model. All trust weights must sum to 1:

$$\sum_{i=1}^{N} W_i = 1$$

Weight Selection Strategy

The trust weights for each model were chosen based on their individual performance. The ViT model showed better accuracy, AUC, and F1-Score compared to the InceptionResNetV2 model. Therefore, the ViT model was given a weight of 0.6, and the InceptionResNetV2 model was given a weight of 0.4. This weighting gives more importance to the ViT model while still including the contribution of InceptionResNetV2. The selection was further validated through trial experiments that confirmed that this combination gave the best overall results.

To select the optimal trust weights for the ViT and InceptionResNetV2 models, several experiments were conducted using different weight combinations. The aim was to identify the weight configuration that achieved the highest classification accuracy on the validation set.

Experiment	ViT Weight	InceptionResNetV2 Weight	Ensemble Accuracy
1	0.80	0.20	96.00%
2	0.50	0.50	96.38%
3	0.60	0.40	97.12%

TABLE 3.1: Summary of weight selection experiments

In **Experiment 1**, the ensemble gave more importance to the ViT model by assigning it a weight of 0.80, while the InceptionResNetV2 model was given a smaller weight of 0.20. The achieved accuracy was 96%, which was almost the same as the ViT model's individual performance. This shows that the contribution of the InceptionResNetV2 model was very limited in this setup.

In **Experiment 2**, equal weights (0.50 each) were assigned to both models. This produced a slight improvement in accuracy, reaching 96.38%. This result indicates that combining the models, even with simple averaging, can provide some performance gain.

In **Experiment 3**, the ViT model was assigned a weight of 0.60, and the InceptionResNetV2 model was assigned a weight of 0.40. This combination achieved the best accuracy of 97.12%, outperforming all other tested configurations and both individual models.

The experimental results indicate that selecting appropriate trust weights helps to combine the strengths of both models, improve diagnostic accuracy, and reduce the risk

of overfitting by balancing their contributions. Based on these findings, the final trust weights chosen for the ensemble model were 0.60 for the ViT model and 0.40 for the InceptionResNetV2 model. This configuration was used in the final decision-making process of the ensemble system.

Decision-Making Process

The decision-making process is designed to handle predictions from either two models or *N* models. The logic is divided into two main cases: (1) when all models agree, and (2) when there is disagreement among the models. Each scenario is handled using carefully defined sub-rules that account for the models' confidence scores, a predefined confidence threshold, and trust-based weighted contributions.

Case 1: Agreement Between All Models

When all *N* models predict the same class (either *"Cancer"* or *"No Cancer"*), it shows a strong agreement between the models.

Decision Logic

- **Final Diagnosis:** Class predicted by all models.
- **Final Confidence Score:** $S_F = \max(S_1, S_2, ..., S_N)$

Example: Two-Model Scenario

- InceptionResNetV2 Model: "Cancer", $S_I = 85\%$
- ViT Model: "Cancer", $S_V = 91\%$
- Final Diagnosis: *Cancer*
- Final Confidence: *91%*

Case 2: Disagreement Between Models

When predictions from the *N* models are not the same, the system uses a two-step decision process.

Step 1: All Models Have High Confidence ($\geq T$)

If all models have confidence scores greater than or equal to the confidence threshold (T), the system calculates the weighted ensemble confidence score using:

$$S_W = \sum_{i=1}^{N} (S_i \times W_i)$$

Sub-case A: Weighted Score $\geq T$

- Calculate the total weighted contribution for each class.
- Select the class with the highest total weighted support across all models.

Sub-case B: Weighted Score $<$ T

- If the combined score does not meet the threshold, the system outputs *"Uncertain"* to reflect low overall reliability.

Example: Two-Model Scenario

- InceptionResNetV2: "Cancer", $S_I = 78\%$
- ViT: "No Cancer", $S_V = 88\%$
- Threshold: $T = 70\%$
- Weights: $W_I = 0.4$, $W_V = 0.6$

$$S_W = (78 \times 0.4) + (88 \times 0.6) = 84.0$$

- Total Weighted Support for Cancer: $78 \times 0.4 = 31.2$
- Total Weighted Support for No Cancer: $88 \times 0.6 = 52.8$

Since $S_W = 84 \geq 70$ and "No Cancer" has higher weighted support, the final diagnosis is **"No Cancer"** with a confidence of **84%**.

Step 2: One or More Models Have Low Confidence ($< T$)

When one or more models have confidence scores below the threshold, the decision becomes more conservative:

- If at least one model's confidence $\geq T$, select the prediction of the most confident model that meets the threshold.
- If all models have confidence scores $< T$, the system outputs *"Uncertain"* and reports the highest individual confidence score.

Example A: Two-Model Scenario

- InceptionResNetV2: "Cancer", 66%
- ViT: "No Cancer", 82%
- Only ViT has confidence $\geq T$.
- Final Diagnosis: No Cancer with confidence: 82%

Example B: Two-Model Scenario

- InceptionResNetV2: "Cancer", 63%
- ViT: "No Cancer", 64%
- Both below threshold.
- Final Diagnosis: Uncertain with confidence: 64%

This conservative decision-making strategy ensures the system avoids overconfident predictions in uncertain cases, supporting reliable and safe clinical decision-making. Table 3.2 provides a summary of the decision logic for different scenarios based on model agreement and confidence levels.

TABLE 3.2: Summary of Decision Logic for Model Prediction Fusion

Condition	**Final Diagnosis**	**Confidence**	**Reason**
All models agree	Agreed class	Highest individual confidence	Strong consensus across models
All confident, but disagree	Class with highest total weighted support	Weighted ensemble score	Conflict resolved using model trust weights
One confident	Prediction from most confident model	That model's confidence	Decision prioritized to the reliable model
All uncertain	"Uncertain"	Highest individual confidence	No reliable signal from the ensemble

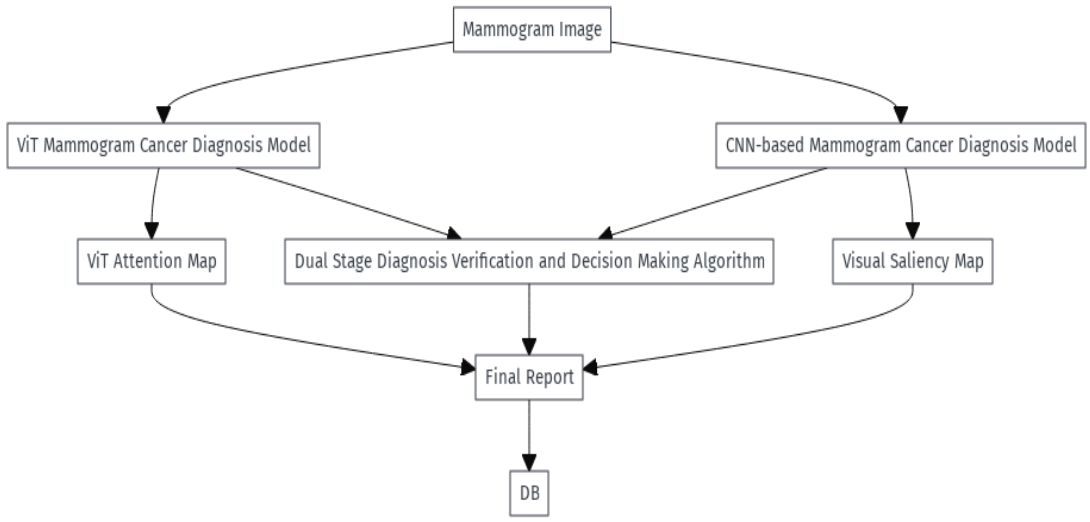

Fig. 3.11: Conceptual Work-flow Diagram for Dual-Stage Ensemble Diagnosis

Figure 3.11 illustrates the dual-stage breast cancer diagnosis pipeline. The system accepts a mammogram image as input and independently processes it through two models: a ViT-based model and a CNN-based model. Each model provides a diagnostic prediction and a corresponding interpretability artifact: the ViT model generates an attention map, while the CNN model produces a visual saliency map.

The predictions and confidence scores from both models are then evaluated using the Dual-Stage Diagnosis Verification and Decision-Making Algorithm, which applies the trust-weighted fusion strategy described above. The final diagnosis, along with the visual explanations, is compiled into a diagnostic report and stored in the database for future clinical reference and review. This system design supports both robust decision-making and interpretable outputs, which are critical for clinical use in breast cancer screening workflows.

The following section discusses how the developed framework was integrated into a web-based system, enabling healthcare professionals to interact with AI-powered predictions and apply them effectively within clinical workflows.

3.11 Integration with Medical System

In the final phase of this research, the developed ensemble-based deep learning model was integrated into a practical clinical tool named BreastAware. This web-based application was designed to assist healthcare professionals in the analysis and diagnosis of mammograms. By bridging the gap between AI-driven diagnostic predictions and real-world clinical workflows, BreastAware enables seamless interaction between radiologists and the deep learning models, enhancing the efficiency of the diagnostic process.

The backend of BreastAware was developed using Flask (v2.3.3), a lightweight Python web framework, to create a RESTful API interface for model inference. Flask-CORS (v4.0.0) was employed to enable secure communication between the frontend and backend across different domains. The backend handles key tasks such as uploading mammogram images, running inference with the trained ensemble models, and generating prediction results, confidence scores, and visual explanation maps like saliency maps and attention overlays.

Model serving was handled using TensorFlow (v2.8.0) and Vit-Keras (v0.1.2), responsible for loading and serving the trained ensemble models. TensorFlow Addons (v0.23.0) was used for implementing custom layers and loss functions when needed. The ensemble model prediction strategy combined output probabilities from both the CNN (InceptionResNetV2) and ViT models. The final classification decision was made using a weighted averaging technique to leverage the strengths of both architectures.

On the frontend, a client interface communicates with the Flask server through HTTP POST requests, sending mammogram images for processing. The server returns a JSON payload containing the classification label ("Cancer" or "No Cancer"), associated confidence scores, and visual explanations that highlight regions most influential in the AI's decision-making process.

For image pre-processing and post-processing, OpenCV (v4.10.0) and Pillow (v11.0.0) were utilized to resize, normalize, and overlay attention maps onto the original mammogram images. Matplotlib (v3.8.0) was used to create high-quality visualizations of the saliency maps, which were then converted into web-compatible formats (e.g., PNG) for easy display within the web application. These visual explanations help clinicians interpret the AI's decisions and validate predictions with greater confidence.

By offering both predicted labels and confidence scores, along with visual cues such as saliency and attention maps, the system fosters transparency and supports informed clinical judgment. Clinicians are better equipped to make quicker and more accurate assessments, allocate more time to direct patient care, and identify high-risk cases that may require immediate attention.

The integration of this AI system into clinical practice bridges the gap between ad-

vanced image analysis and real-world healthcare settings. It enhances time efficiency by streamlining the diagnostic workflow, reducing the manual burden of reviewing each scan, and allowing for better tracking of patient data over time. The result is a tool that empowers clinicians to make faster, more accurate diagnoses, ultimately improving patient care.

3.12 Tools and Technologies

The implementation of this project was carried out using Python due to its reputation as a high-level, interpreted programming language that excels in simplicity, readability, and user-friendliness, particularly in the domains of AI and ML. The following libraries and frameworks were utilized throughout the project:

- TensorFlow (Version 2.8.0): A powerful open-source framework used for building and training deep learning models. TensorFlow was integral in the model development and training phases.
- Flask (Version 2.3.3): A lightweight web framework for Python, used for deploying the application as a web service. Flask allows for easy creation of APIs to interact with the trained models.
- NumPy (Version 1.26.4): A core library for scientific computing in Python that offers support for large, multi-dimensional arrays and matrices, as well as a wide range of advanced mathematical functions to manipulate these structures.
- Matplotlib (Version 3.8.0): A library used for creating static, animated, and interactive visualizations in Python. It was utilized for plotting graphs, charts, and visualizing the model outputs.
- Pillow (Version 11.0.0): A Python Imaging Library used for opening, manipulating, and saving various image file formats, crucial for image preprocessing and handling in the project.
- OpenCV (Version 4.10.0): A widely used computer vision library that gives real-time image processing and computer vision tasks. It was used for tasks such as image reading, manipulation, and processing.
- PyTorch: An open-source machine learning library PyTorch was employed for certain aspects of model development and experimentation, providing flexibility in neural network building and training.
- PyDicom: A Python library for working with DICOM files, essential for handling medical imaging data, particularly for extracting metadata and pixel data from medical images.

- Vit-Keras (Version 0.1.2): A library for implementing Vision Transformer (ViT) models with Keras, facilitating the use of transformer-based models for image classification tasks.
- TensorFlow Addons (Version 0.23.0): A collection of additional functionality and custom operations for TensorFlow, used to extend the capabilities of TensorFlow during the model development process.
- Flask-Cors (Version 4.0.0): A Flask extension used to handle Cross-Origin Resource Sharing (CORS), enabling secure communication between the web server and frontend interfaces.

Together, these tools and technologies provide a comprehensive ecosystem for developing, deploying, and evaluating ML models, particularly in the context of computer vision and medical image analysis.

CHAPTER 4

RESULTS AND DISCUSSION

The content is structured into three main parts: Evaluating CNN Architecture, evaluating and comparing ViT against CNN models, and evaluating XAI Techniques.

4.1 Results and Discussion for CNN Architecture

In the Evaluating CNN Architecture section, two key experiments were conducted to assess the performance of different CNN models in classifying mammogram images. The first experiment evaluated the models' baseline performance, aiming to identify the best-performing CNN architecture. Based on the results, the second experiment involved applying further optimizations to the top-performing model to enhance its accuracy and generalization. This approach allowed for a more refined analysis of the models, ensuring the selected architecture was fine-tuned and its performance thoroughly tested under optimized conditions.

4.1.1 Results of Experiment 01

Table 4.1 presents the evaluation metrics (precision, recall, F1-score, and accuracy) for various deep-learning models used in the classification of mammogram images into cancerous and non-cancerous categories. The results show that the InceptionResNetV2

TABLE 4.1: Evaluation metrics of CNN pre-trained models

Method	Cancer Precision	Cancer Recall	Cancer F1-Score	No Cancer Precision	No Cancer Recall	No Cancer F1-Score	Overall Accuracy
ResNet50	0.38	0.21	0.27	0.48	0.67	0.55	44%
ResNet101	0.63	0.69	0.66	0.67	0.62	0.64	65%
VGG16	0.95	0.67	0.79	0.75	0.96	0.84	81%
VGG19	0.82	0.81	0.81	0.82	0.83	0.82	81%
MobileNetV2	0.91	0.89	0.90	0.90	0.92	0.91	90%
DenseNet121	0.95	0.80	0.86	0.84	0.95	0.88	87%
InceptionResNetV2	0.96	0.89	0.93	0.90	0.96	0.92	92%

model obtained the best overall performance, with precision, recall, and F1-score consistently high for both cancer and non-cancer classifications. Specifically, it achieved an impressive accuracy of 92%, surpassing all other models in terms of overall effectiveness. This high accuracy is a result of the model's ability to balance precision and recall across both classes, indicating that it successfully minimized both false positives and false negatives.

MobileNetV2, which received an accuracy of 90%, also delivered excellent results, with high precision and recall across both cancer and non-cancer. This model stands out due to its lightweight architecture, which suggests that it could be a good candidate for deployment in resource-constrained environments like mobile applications or hospitals with limited computing power. However, MobileNetV2's slightly lower performance compared to InceptionResNetV2 may indicate that it is less capable of capturing complex features in mammogram images, which could be critical for high-risk cancer detection.

The VGG models, VGG16 and VGG19, also exhibited robust performance, particularly in cancer detection. VGG16 achieved a high cancer precision of 0.95, making it highly accurate for identifying cancerous cases. Both models demonstrated a high recall for non-cancerous images, resulting in an overall accuracy of 81%. While these models performed admirably in detecting cancer, they were relatively less effective in distinguishing non-cancerous cases, as seen in their lower recall and F1-scores for non-cancer classifications. This points to a potential issue in their ability to generalize to diverse mammogram patterns, which could be addressed by further data augmentation or fine-tuning of the model's hyperparameters.

DenseNet121 scored highly in cancer precision and achieved a good balance across all metrics, with an overall accuracy of 87%. This model performed better than the ResNet models, particularly in identifying cancerous cases. While it exhibited a more balanced performance than the VGG models, there is still chance for improvement, particularly in terms of recall for non-cancerous images. The overall results suggest that DenseNet121 could be a reliable model for general use, but further optimization is necessary for fine-tuning its ability to identify differences in benign and malignant cases.

The ResNet models, specifically ResNet101, demonstrated moderate performance, with ResNet101 outperforming ResNet50, particularly in handling cancer cases. ResNet50 showed relatively poor precision and recall for cancer detection, resulting in an overall accuracy of just 44%. In contrast, ResNet101 achieved a more balanced classification with an accuracy of 65%. While ResNet models are generally robust in handling deep feature extraction, their performance in this experiment suggests that these models require more tailored adjustments, such as customized loss functions or additional regularization, to perform optimally in mammogram classification tasks.

These results highlight the effectiveness of using advanced convolutional neural networks for the task of mammogram image classification, with InceptionResNetV2 and MobileNetV2 standing out as particularly promising models for further development and optimization in this application.

Classification Report:

	precision	recall	f1-score	support
No Cancer	0.91	1.00	0.96	75
cancer	1.00	0.91	0.95	75
accuracy			0.95	150
macro avg	0.96	0.95	0.95	150
weighted avg	0.96	0.95	0.95	150

Fig. 4.1: Classification Report for Experiment 02 InceptionResNetV2 Model: Demonstrates the model's precision, recall, and F1-scores across both classes, highlighting strengths and weaknesses in cancer detection.

4.1.2 Results of Experiment 02

The enhanced version of the InceptionResNetV2 model demonstrates impressive results 4.1 in mammogram classification, achieving an overall accuracy of 95%. The model exhibits high precision for both non-cancerous (91%) and cancerous (100%) cases, which highlights its exceptional ability to identify true positives in cancer detection. The 100% precision for cancer detection is especially noteworthy, as it means the model correctly classified all cancerous cases without any false positives. This is a key feature in medical diagnostics, where false positives can result in unnecessary, and sometimes invasive, procedures for patients.

The recall scores are high as well, suggesting that the model is good at identifying most of the cancerous and non-cancerous cases. Recall, in medical settings, is often more important than precision because it reflects the model's ability to capture as many true positives as possible. The balanced F1-score reinforces this by indicating that the model performs well across both classes, without a significant bias toward one over the other. This is a positive outcome, as it suggests the model is reliable for both detecting cancer and distinguishing between cancerous and non-cancerous cases.

However, despite these strengths, the confusion matrix 4.2 reveals some issues that warrant attention. Specifically, the model made 7 false negative predictions and 5 false positives. False negatives are a critical concern in medical diagnostics, as they indicate missed cancerous cases that could result in delayed treatment and potentially worsen patient outcomes. This underscores the importance of improving the model's sensitivity to cancerous cases. One potential solution is to incorporate more diverse and challenging mammogram data during training, particularly cases where the cancer is less obvious or early-stage, as these can be more difficult to detect. Additionally, using techniques such as class balancing, oversampling, or weighted loss functions could help to mitigate the occurrence of false negatives by making the model more

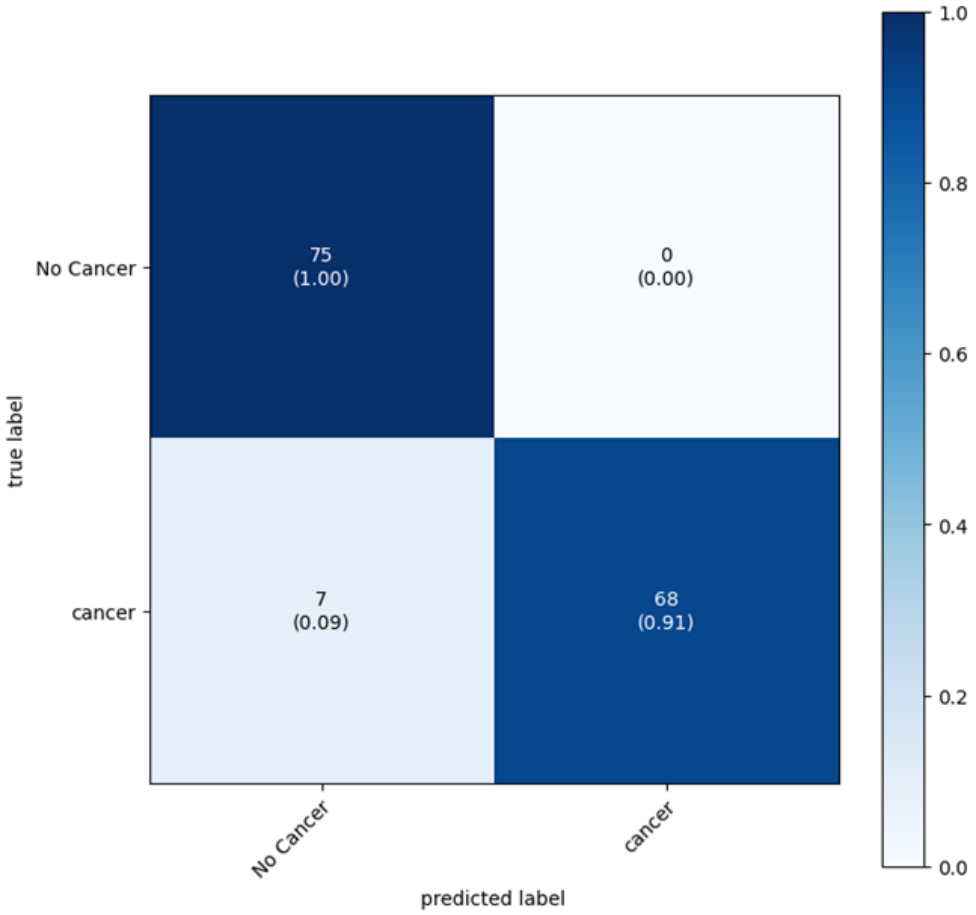

Fig. 4.2: Confusion Matrix for Experiment 02 InceptionResNetV2 Model: displays the distribution of true positives, true negatives, false positives, and false negatives, emphasizing areas for improvement in misclassification reduction.

sensitive to rare cancerous cases.

The training loss plot 4.3 suggests that the model learned effectively, with the loss decreasing steadily toward the final epochs. This indicates that the model was able to optimize its parameters during training without overfitting, which is a positive sign. However, slight fluctuations in the validation loss suggest minor instability, possibly due to variability in the validation data. Techniques such as additional regularization or improved learning rate schedules could enhance this stability.Furthermore, The training accuracy plot in Figure 4.4 shows a sharp increase in accuracy during early epochs before stabilizing near 95%. While the training and validation accuracies align closely, the small gap suggests marginal overfitting. This can be mitigated by incorporating additional augmentation techniques or refining early stopping strategies to ensure optimal generalization.

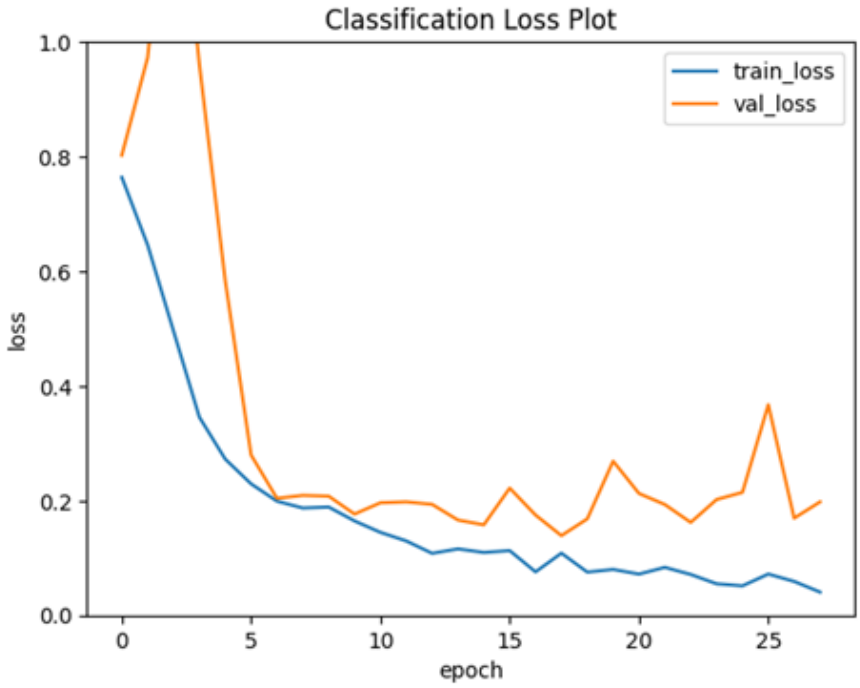

Fig. 4.3: Classification Loss Plot for Experiment 02 InceptionResNetV2 Model: Illustrates the reduction in model loss during training and validation, showing stable learning progression with minor fluctuations

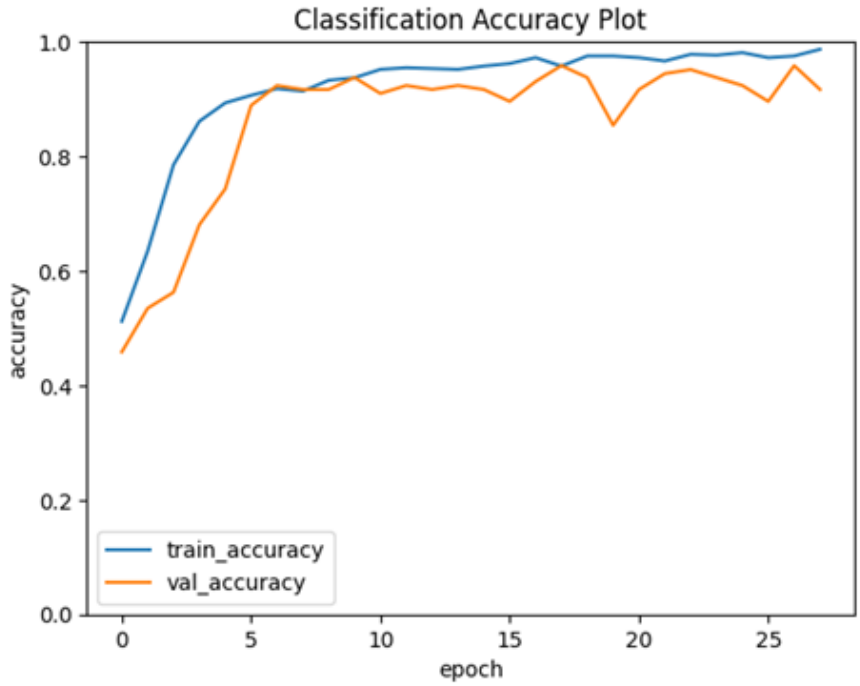

Fig. 4.4: Classification Accuracy Plot for Experiment 02 InceptionResNetV2 Model: Shows the progression of training and validation accuracy, indicating improved performance with minimal risk of overfitting.

4.2 Comparison of Experiment 01 & 02

TABLE 4.2: Comparison of Experiment 01 and Experiement 02 InceptionResNetV2 model

Metric	Experiement 02	Experiment 02	Improvement in Experiement 02
Cancer Precision	0.90	1.00	Yes
Cancer Recall	0.89	0.91	Yes
Cancer F1-Score	0.93	0.95	Yes
No Cancer Precision	0.90	0.91	Yes
No Cancer Recall	0.96	1.00	Yes
No Cancer F1-Score	0.92	0.96	Yes
Overall Accuracy	92%	95%	Yes

Table 4.2 highlights the comparison between Experiment 01 and Experiment 02, both utilizing the InceptionResNetV2 model for breast cancer classification in mammogram images. The improvements observed in Experiment 02 across all metrics are highly significant and demonstrate the positive impact of the adjustments made in this phase, including fine-tuning model parameters, refining data augmentation techniques, and optimizing the network architecture.

The most notable improvement is in the cancer precision, which rose from 0.90 in Experiment 01 to a perfect 1.00 in Experiment 02. Achieving 100% precision for cancer detection is a remarkable result, as it means that the model identified every cancerous case correctly without mistakenly labeling any non-cancerous case as cancerous (i.e., no false positives). In medical diagnostics, this precision is crucial as it directly impacts the avoidance of unnecessary interventions. False positives could result in unwarranted biopsies, leading to patient anxiety, unnecessary procedures, and healthcare costs. However, it is essential to approach this result with some caution. While a precision of 1.00 is ideal in terms of avoiding false positives, this metric must be balanced with recall to ensure that true positives (actual cancer cases) are not missed. Given that the precision metric doesn't account for false negatives, it is essential to examine the recall and F1-score for a more comprehensive evaluation.

In addition, cancer recall improvement from 0.89 to 0.91 indicates a modest increase in the model's ability to detect cancer, reducing false negatives but still highlighting the importance of minimizing missed diagnoses. False negatives remain a concern in breast cancer detection due to their potential impact on patient outcomes.

The cancer F1-score improvement from 0.93 to 0.95 reflects a better balance between precision and recall, ensuring the model is neither too cautious nor too aggressive in its predictions. This makes the model more reliable and consistent, enhancing its clinical applicability for cancer detection.

The non-cancer precision improved slightly from 0.90 to 0.91, reflecting a modest increase in the model's ability to avoid misclassifying non-cancerous cases as cancerous. This improvement is beneficial as it reduces the risk of patients being falsely diagnosed with cancer. However, the precision increase is marginal, and the more striking improvement is observed in non-cancer recall, which improved from 0.96 to a perfect 1.00. A recall of 1.00 means the model correctly identified every non-cancerous case, leaving no room for false negatives in this category. In real-world clinical settings, non-cancer recall is vital because misdiagnosing a non-cancerous case as cancerous can lead to unnecessary treatments and procedures.

With this improvement, the non-cancer F1-score saw a significant rise from 0.92 to 0.96, indicating stronger overall model reliability in identifying non-cancerous cases. This enhancement ensures that the model is not only precise in detecting cancer but also highly reliable in ruling out non-cancerous cases, thus reducing both false positives and false negatives.

The overall accuracy of the model increased from 92% in Experiment 01 to 95% in Experiment 02. While accuracy is an essential metric, it must be viewed with caution in highly imbalanced datasets like medical image classification, where certain classes (e.g., cancerous cases) may be underrepresented. In such contexts, a high overall accuracy does not always imply good model performance across all classes, especially in terms of detecting rare conditions. Despite this, the 3% improvement in accuracy, coupled with the marked improvements in precision, recall, and F1-scores, highlights the model's enhanced ability to distinguish between cancerous and non-cancerous cases. This increased accuracy is promising for real-world clinical settings where quick and accurate diagnoses are essential.

The improvements in Experiment 02 underscore the importance of fine-tuning model parameters and optimizing the network structure. These adjustments have contributed significantly to the model's increased ability to identify both cancerous and non-cancerous cases accurately. For instance, the marginal increase in precision for non-cancer cases and the substantial improvement in recall and F1-score for both cancerous and non-cancerous cases suggest that the model's ability to generalize has improved.

Data augmentation likely played a role in these enhancements. By generating more varied training data, augmentation techniques such as rotation, flipping, and scaling can help the model learn more generalized features, improving its performance on unseen data. The increase in both cancer and non-cancer recall points to an improved model sensitivity, which could be a result of the enriched dataset used during training.

In conclusion, as the model is refined further, its potential to assist healthcare professionals in early breast cancer detection becomes even more promising. However, attention should be given to reducing false negatives, improving generalization across diverse datasets, and maintaining a balance between precision and recall to ensure that the model remains a trustworthy tool in medical diagnostics.

4.3 Results and Discussion for ViT Architecture

Implementing the ViT model in the classification of mammogram images marked a significant phase in this project. The ViT model demonstrated excellent classification accuracy and consistency across metrics, as in the classification report in figure 4.5 for the test data comprising 150 images.

The model achieved an overall accuracy of 96%, a robust indicator of its ability to classify both conditions effectively. Both classes provide high precision, recall, and F1-score rates at 0.96. These metrics underscore the model's balanced capability in identifying true positives while minimizing false negatives and false positives, which is essential for reliable medical diagnostics. Moreover, the model training was halted at the 15th epoch which elaborates the efficiency of the ViT model, facilitated by the early stopping. The high precision, recall, and F1 scores across both cancer and no cancer categories indicate that the ViT architecture is highly effective in medical image analysis. This is largely attributed to its ability to focus on various parts of the image and understand the contextual relationships, which is vital in detecting nuanced features in mammograms.

Moreover, the confusion matrix in figure 4.6 also displays the classification accuracy of the model for the 'No Cancer' and 'Cancer' classes. Out of 75 cases in each class, the model correctly identified 72 cases in both categories. This translates to a TPR and TNR of 96% each. The model's generation of three false positives and three false negatives points to specific areas for optimization. The low false positive and false negative rates (4% each) suggest that the model is highly reliable in distinguishing between cancerous and non-cancerous images. However, in a clinical setting, even a 4% error rate requires close monitoring to mitigate potential risks

The classification accuracy plot in figure 4.7 illustrates that both training and validation accuracies start high and remain stable throughout the training process, converging close to a 96% accuracy rate. The high accuracy, along with the small gap between training and validation performance, suggests strong generalization and minimal overfitting. The classification loss plot in figure 4.8 illustrates a sharp decrease in training loss in the initial epochs, stabilizing around a low value as epochs progress. The validation loss also decreases and remains relatively flat, with slight fluctuations towards the later epochs. The convergence of training and validation loss with minimal divergence further supports the model's robustness and generalization capabilities.

Classification Report:

	precision	recall	f1-score	support
No Cancer	0.96	0.96	0.96	75
cancer	0.96	0.96	0.96	75
accuracy			0.96	150
macro avg	0.96	0.96	0.96	150
weighted avg	0.96	0.96	0.96	150

Fig. 4.5: Classification Report for `vit_b_16` Model: Highlights ViT's balanced precision, recall, and F1-scores, demonstrating consistent performance in classifying cancerous and non-cancerous cases.

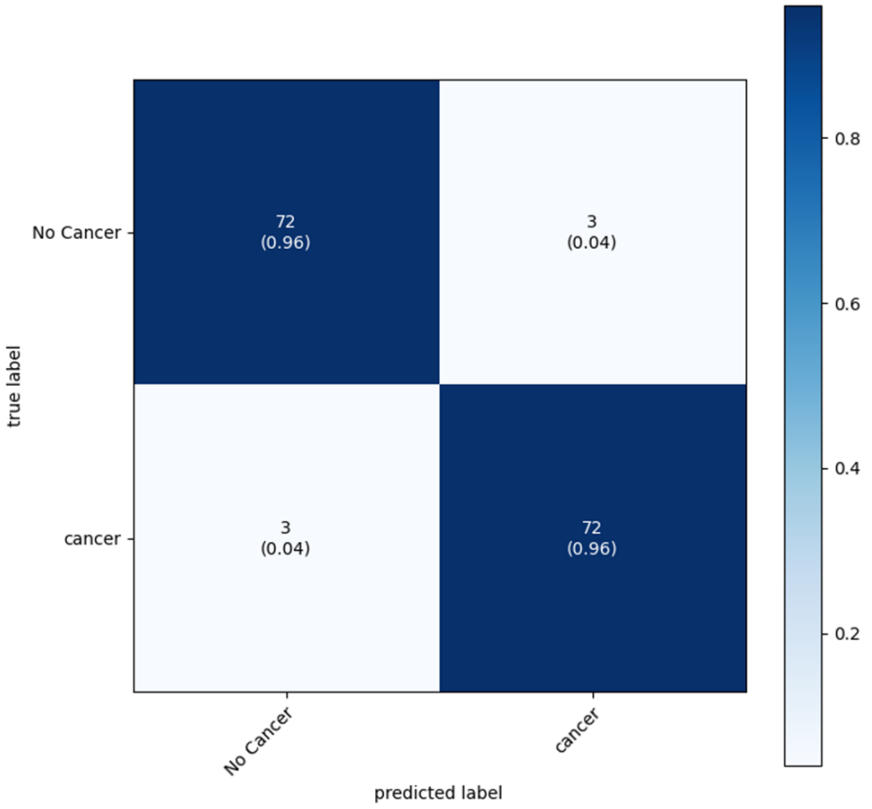

Fig. 4.6: Confusion Matrix for `vit_b_16` Model: Visualizes the ViT model's ability to minimize false positives and negatives, showcasing accurate classification

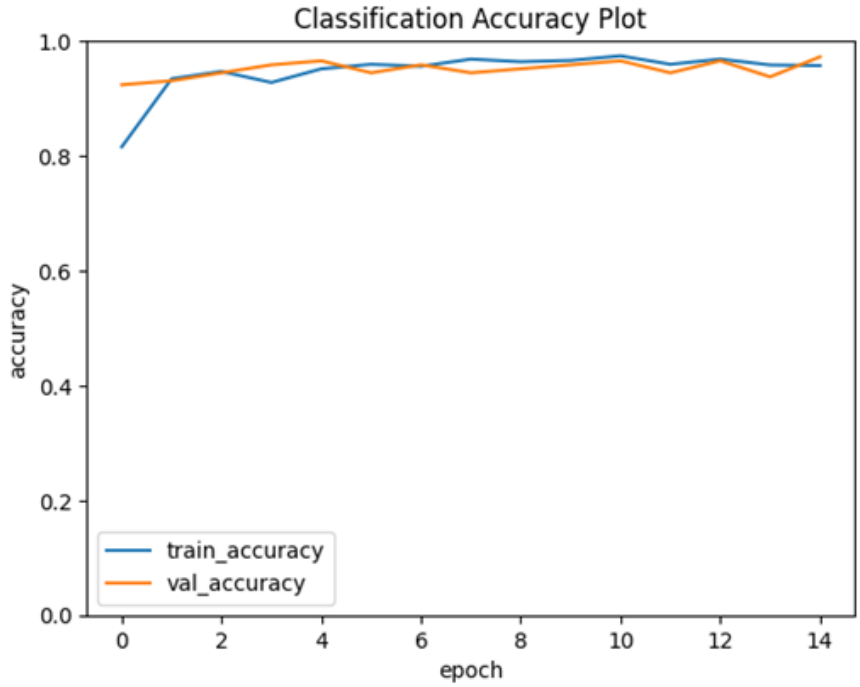

Fig. 4.7: Classification Accuracy Plot `vit_b_16` Model: Depicts consistent training and validation accuracy trends, reinforcing the model's stability.

The minimal gap between training and validation accuracy indicates that the model has achieved good generalization, avoiding overfitting. The stability in both accuracy and loss plots supports the model's robustness in real-world diagnostic applications.

Furthermore to determine whether the ViT model's improved performance over the InceptionResNetV2 model was statistically significant, a paired t-test was conducted for key evaluation metrics: precision, recall, and F1-score. The results indicated no statistically significant difference in precision ($t = -0.14$, $p = 0.90$) or recall ($t = -0.41$, $p = 0.71$), suggesting that the observed differences in these metrics could be attributed to random variation. However, the F1-score comparison yielded a marginally significant result ($t = -3.00$, $p = 0.057$), which is close to the conventional significance threshold of 0.05. This suggests that while the ViT model's improved F1-score may indicate a meaningful advantage in balancing precision and recall.Overall, the ViT model demonstrated consistently strong performance across all metrics, with a borderline significant advantage in maintaining a well-balanced precision-recall tradeoff.

While the ViT model performed well, it is more demanding in terms of computational resources compared to CNN models like InceptionResNetV2. ViT's self-attention mechanism processes multiple image patches at once, requiring more memory and increasing processing time. This can slow down training and make deployment more challenging in resource-limited environments.

In contrast, CNN models like InceptionResNetV2 use efficient convolutional layers

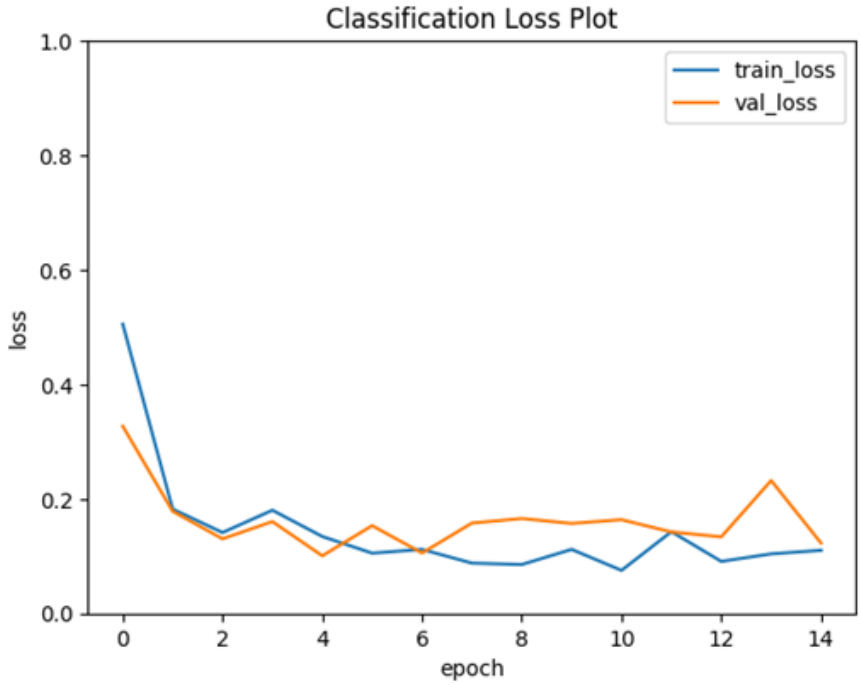

Fig. 4.8: Classification Loss Plot `vit_b_16` Model: Displays stable loss reduction during training, confirming ViT's effective convergence with minimal overfitting.

that require less memory and computing power. This makes CNN models faster and more practical for real-world deployment, particularly in clinics or mobile solutions with limited resources. While ViT may offer slight performance improvements, CNN models are often a better choice for faster predictions and efficiency. For applications where deeper insights and interpretability are critical, ViT models may still be the preferred option despite their higher resource needs.

However, given the comparable performance of both models, ViT and Inception-ResNetV2 can be considered well-suited models for further exploration with advanced XAI techniques. Their consistent and reliable performance makes them strong candidates for interpretability methods that provide deeper insights into model predictions. The next sections 4.4 and 4.5 discuss the application of these XAI techniques for both models, exploring how they enhance model transparency and decision-making.

4.4 Results and Discussion for ViT Based XAI Applicaton

As discussed in the 4.3, ViT model achieved stable performance with 96% accuracy across both training and validation phases, demonstrating strong generalization without signs of overfitting. This reliability underscores its potential for real-world application. The model's explainability can be further illustrated using the pair of Figures 4.9, which demonstrates the attention maps for cancerous images generated by the ViT

model. The model's attention mechanism and highlights the areas it focuses on when making predictions. In medical imaging, this can correlate closely with areas of diagnostic interest, such as tumors or other abnormalities. **First Pair of Figures**

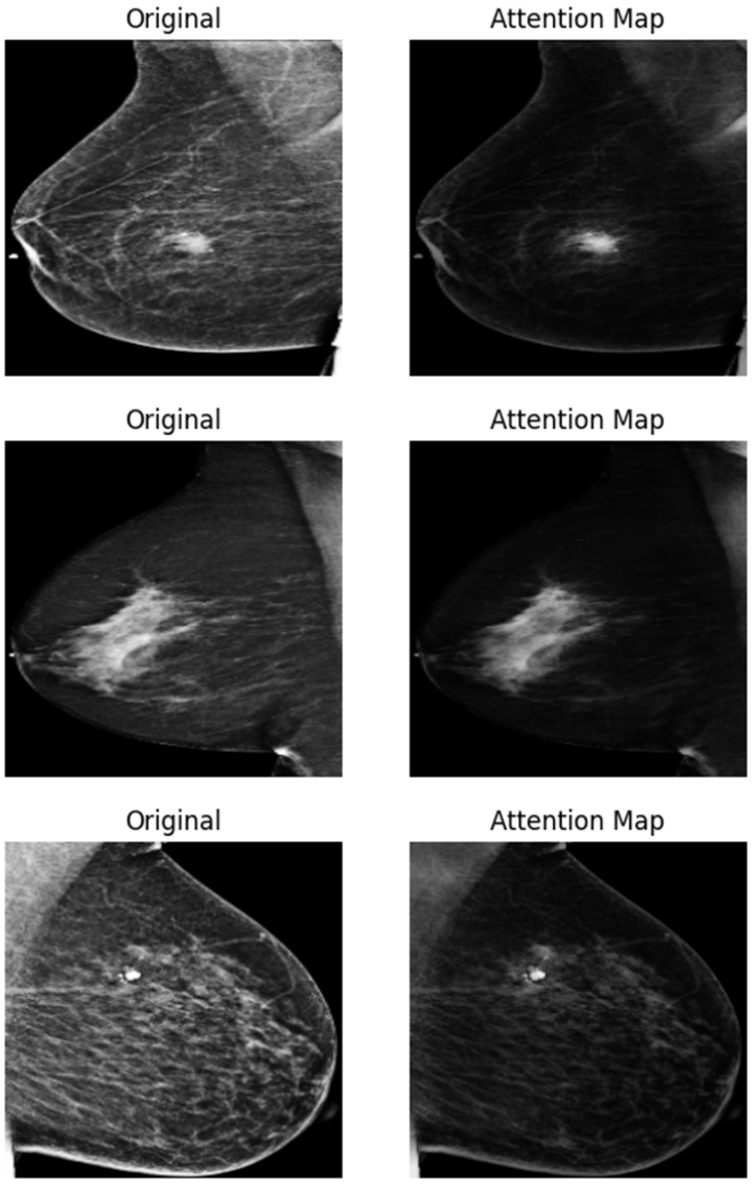

Fig. 4.9: ViT model-generated attention maps, showing focused regions in cancerous mammogram images for improved model interpretability

- Original Image: This image is labeled as malignant cancer from the test set, indicating the presence of a potential abnormality. This displays a typical mammogram with visibility of the breast tissue. A notable dense area is observed which could be a region of interest for diagnostic purposes.
- Attention Map: Highlights brighter areas where the model's attention was par-

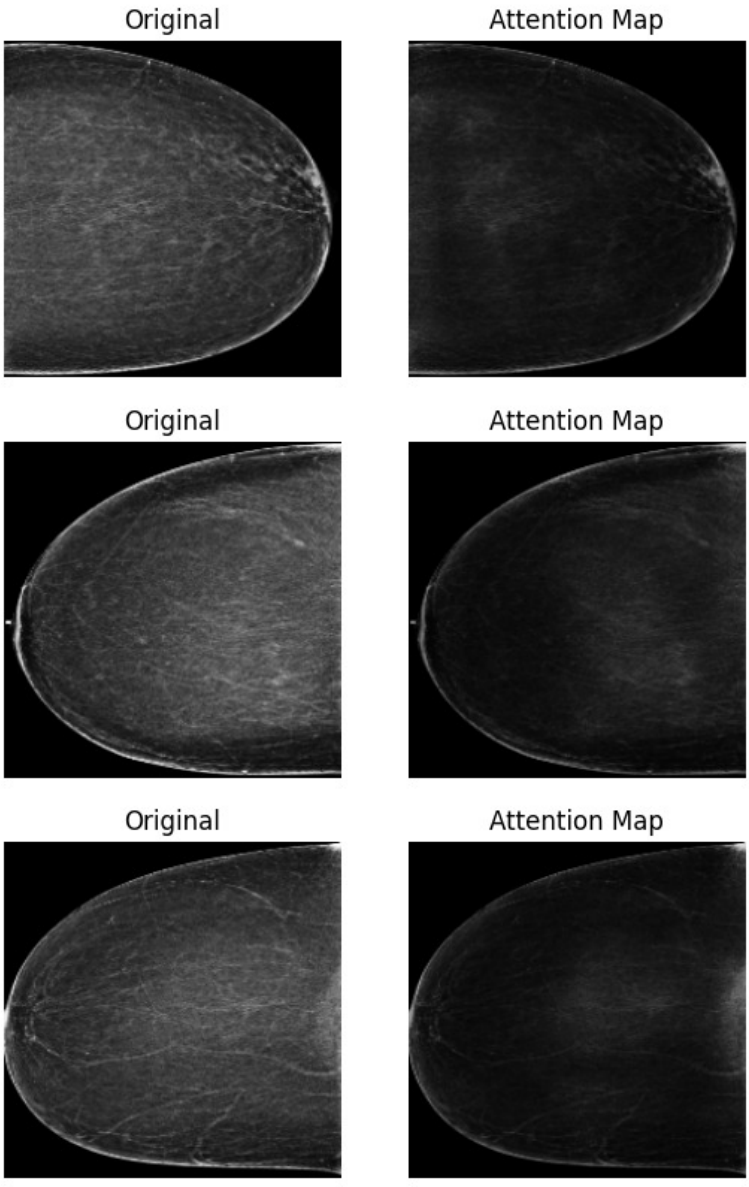

Fig. 4.10: ViT model-generated attention maps overlaid on non-cancerous mammogram images, highlighting key non-cancer regions.

ticularly focused. In this case, the model concentrates around the dense region, possibly indicating an area that the model predicts as having higher likelihood of abnormality. The attention is not uniformly spread but is localized, suggesting that the model is effectively focusing on potentially significant anomalies.

Second Pair of Images

- Original Image: This is another mammogram with clear fibroglandular densities. Such features in mammograms are often scrutinized for any irregular patterns that might suggest malignancy.
- Attention Map: The attention map here again illuminates these dense areas more brightly. This focused attention indicates that the model recognizes the complexity of tissue patterns here, which could be crucial for identifying pathological changes.

Third Pair of Images

- Original Image: This features a mammogram where there's a conspicuous bright spot, which might be calcifications or another diagnostic marker significant in breast cancer screening.
- Attention Map: The map shows intense focus around this bright spot, reinforcing its potential importance. The specificity of attention here could help in confirming the relevance of these findings, supporting radiologists in their assessment.

The explainability of the ViT model is illustrated through the pair of Figures 4.10, which show the attention maps for non-cancerous images generated by the ViT model. As these images are predicted as non-cancerous from the Vit base model, the attention maps do not highlight any specific focus areas, reflecting the absence of prominent features. The lack of highlighted areas in the attention maps for non-cancerous images supports the model's prediction. Since the ViT model classifies the images as non-cancerous and the attention maps don't focus on any specific regions, it shows that the model did not find any abnormalities. This consistency between the model's prediction and the attention maps confirms the model's decision, highlighting the reliability of its prediction and the usefulness of XAI techniques in explaining the model's behavior.

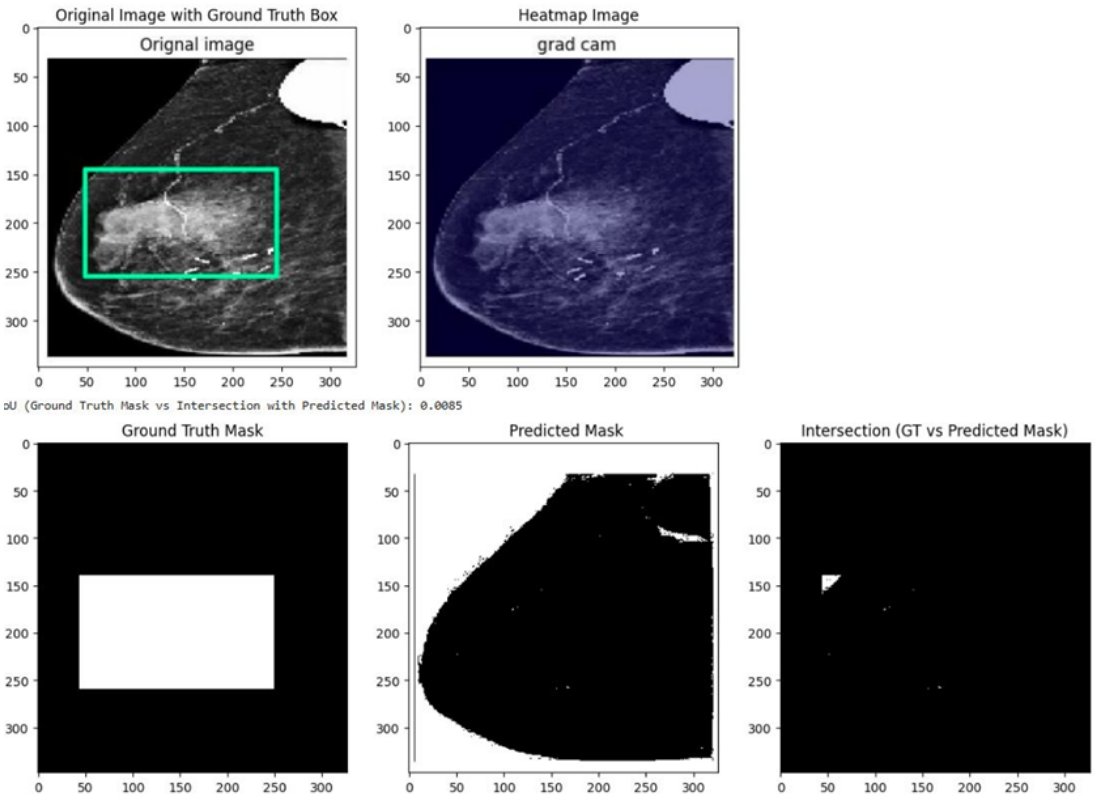

Fig. 4.11: Grad-CAM Results — Illustrates the Grad-CAM-generated heatmaps over mammogram images, showing important regions the InceptionResNetV2 model focused on for both cancerous and non-cancerous predictions.

4.5 Results and Discussion for CNN Based XAI Applicaton

This section covers the results of the implementation of the XAI techniques on the selected best-performing InceptionResNetV2 model. The study focuses on the XAI methods outlined in Section 3.8.2, which were applied to evaluate the model's performance. The results demonstrate how each technique interprets the decision-making process and interpretability of the InceptionResNetV2 model, highlighting key areas of the image that influence its predictions.

- **Grad-CAM**

To evaluate the effectiveness of the saliency maps, IoU metric was applied, which quantitatively measures the similarity between the ground truth mask and the predicted mask generated by the XAI techniques. The evaluation process was visualized by overlaying heatmaps on the original image, displaying binary masks for both ground truth and predicted saliency regions, and presenting an intersection mask to show the overlap between the two. This approach allowed for a clear and comprehensive comparison of how well each XAI method highlighted the relevant areas in the images, setting the

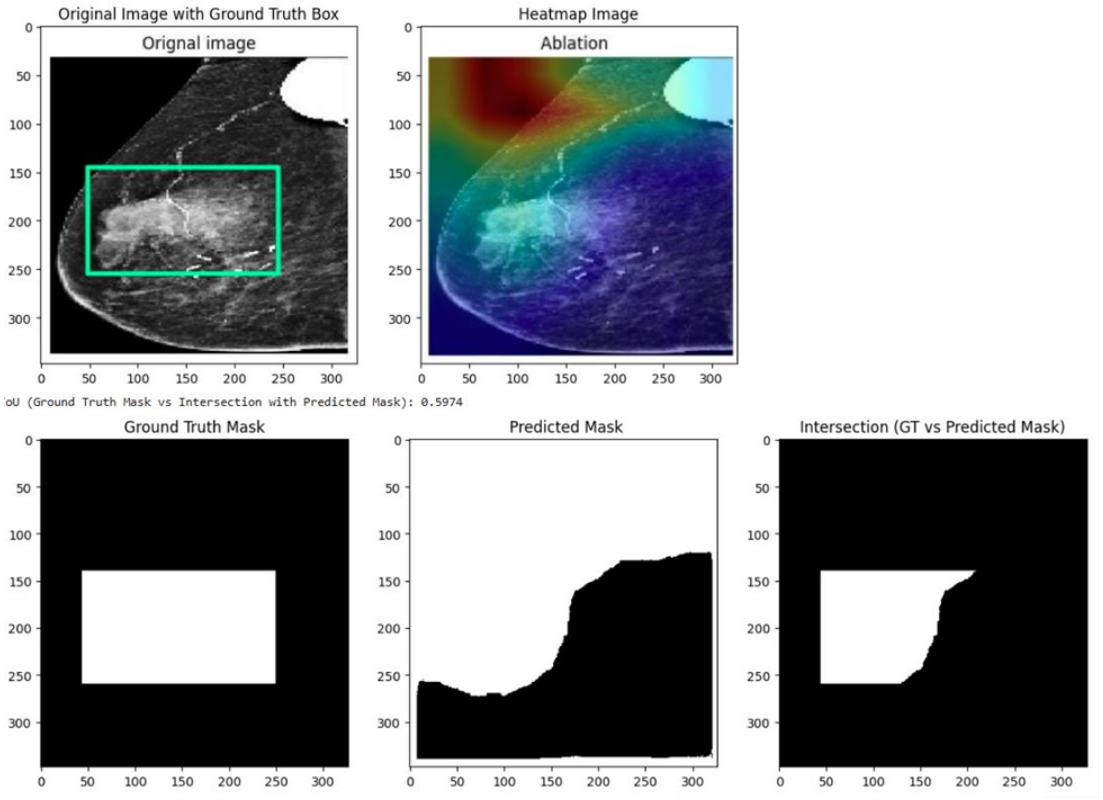

Fig. 4.12: Ablation-CAM Results — Displays Ablation-CAM-generated heatmaps that emphasize key visual patterns influencing the InceptionResNetV2 model's predictions.

stage for further medical discussion of the results. These visualizations and metrics provided valuable insights into the strengths and limitations of each XAI technique, helping to guide future improvements.

The Figure 4.11 illustrates the evaluation of the Grad-CAM approach for cancerous saliency feature selections, with an IoU score of 0.0085. The original image shows the ground truth bounding box, representing the region of interest. The heatmap generated by Grad-CAM highlights certain regions, but it fails to align effectively with the ground truth ROI. The ground truth mask and predicted mask exhibit minimal overlap, as evident in the intersection visualization, where the overlap area is negligible. This low IoU score indicates that the Grad-CAM method struggled to accurately localize the features corresponding to the ROI in this instance.

In non-cancer clasification tasks,the Grad-CAM output with an IoU score of 0.2123, reflecting limited alignment with the ROI. Although the heatmap activated some relevant areas, there was considerable deviation from the actual ROI. The minimal intersection indicates Grad-CAM's difficulty in precisely localizing non-cancer-related features, suggesting limited interpretability in this context.

- **Ablation-CAM**

The Figure 4.12 illustrates the evaluation of the Ablation-CAM for approach for cancerous saliency feature selection, achieving an IoU score of 0.5974. The original image includes the ground truth bounding box, representing the region of interest. The heatmap generated by Ablation-CAM partially aligns with the ROI, as evident from the overlap between the ground truth mask and the predicted mask. The predicted mask captures some relevant regions but also includes significant unnecessary areas, as seen in the intersection visualization. This moderate IoU score indicates that Ablation-CAM provides a reasonable approximation of the ROI but lacks the precision required for highly accurate feature localization.

In non-cancer clasification tasks,the Ablation-CAM output with an IoU score of 0.6542. The heatmaps revealed partial alignment with the ground truth regions, and the intersection masks showed a moderate level of overlap. While some irrelevant areas were highlighted, Ablation-CAM provided a better approximation of the ROI than Grad-CAM.

- **SIDU**

The Figure 4.13 illustrates the evaluation of the SIDU approach for approach for cancerous saliency feature selection, achieving a high score of 0.9805. The original image with the ground truth bounding box outlines the region of interest, while the heatmap generated by SIDU accurately highlights this area with minimal deviation. The ground truth mask and the predicted mask showcase significant overlap, as visualized in the intersection panel, reflecting excellent alignment between the predicted saliency and the ground truth. This high IoU score emphasizes SIDU's precision and reliability in identifying key features within the ROI.

As in Figure 4.14, non-cancer clasification tasks,the SIDU output with an IoU score of 0.9587. The generated heatmaps strongly aligned with the ground truth non-cancer regions, capturing the critical features with high fidelity. The intersection masks showed a strong overlap, confirming this method's ability to provide precise and reliable explanations. This result indicates that Similarity Difference is particularly effective in non-cancer classification by accurately highlighting decision-relevant areas.

TABLE 4.3: Comparison of SIDU, Grad-CAM, and Ablation-CAM Performance Using the Intersection over Union (IoU) Metric

Method	SIDU	Grad-CAM	Ablation-CAM
IoU Score	0.9805	0.0085	0.5974

In summary, the performance of the SIDU, Grad-CAM, and Ablation-CAM saliency methods was evaluated using the IoU metric, which measures the overlap between the

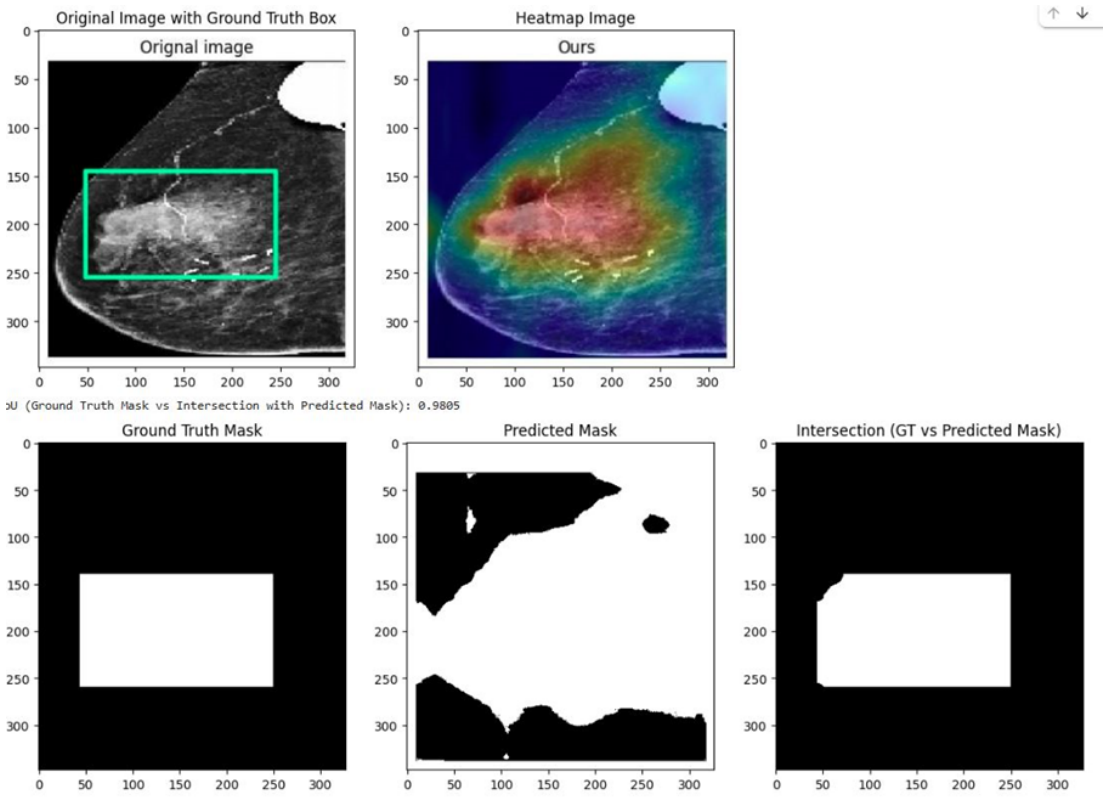

Fig. 4.13: SIDU Results — Demonstrates SIDU-generated saliency maps, emphasizing high-accuracy feature selection both cancerous images for improved transparency in AI predictions.

predicted mask and the ground truth mask. As results shown in Table 4.3 that SIDU achieved the highest IoU score of 0.9805, demonstrating its superior ability to accurately identify the regions of interest in the images. In contrast, Grad-CAM performed poorly with a very low IoU score of 0.0085, indicating its struggle in correctly highlighting the relevant areas. Ablation-CAM, with an IoU score of 0.5974, showed moderate performance, partially aligning with the ground truth but still not as accurate as SIDU.

TABLE 4.4: Average IoU Scores for SIDU, Grad-CAM, and Ablation-CAM across 50 Cancerours Images

Method	SIDU	Grad-CAM	Ablation-CAM
IoU Score	0.9404	0.2545	0.7554

To further validate the effectiveness of the saliency methods, additional experiments were conducted using both cancerous and non-cancerous images from the test set. A total of 50 randomly selected images were evaluated using three saliency techniques: Similarity Difference and Uniqueness (SIDU), Grad-CAM, and Ablation-CAM. The average Intersection over Union (IoU) scores for each method were cal-

TABLE 4.5: Average IoU Scores for SIDU, Grad-CAM, and Ablation-CAM across 50 Non-Cancerours Images

Method	SIDU	Grad-CAM	Ablation-CAM
IoU Score	0.9424	0.3387	0.7028

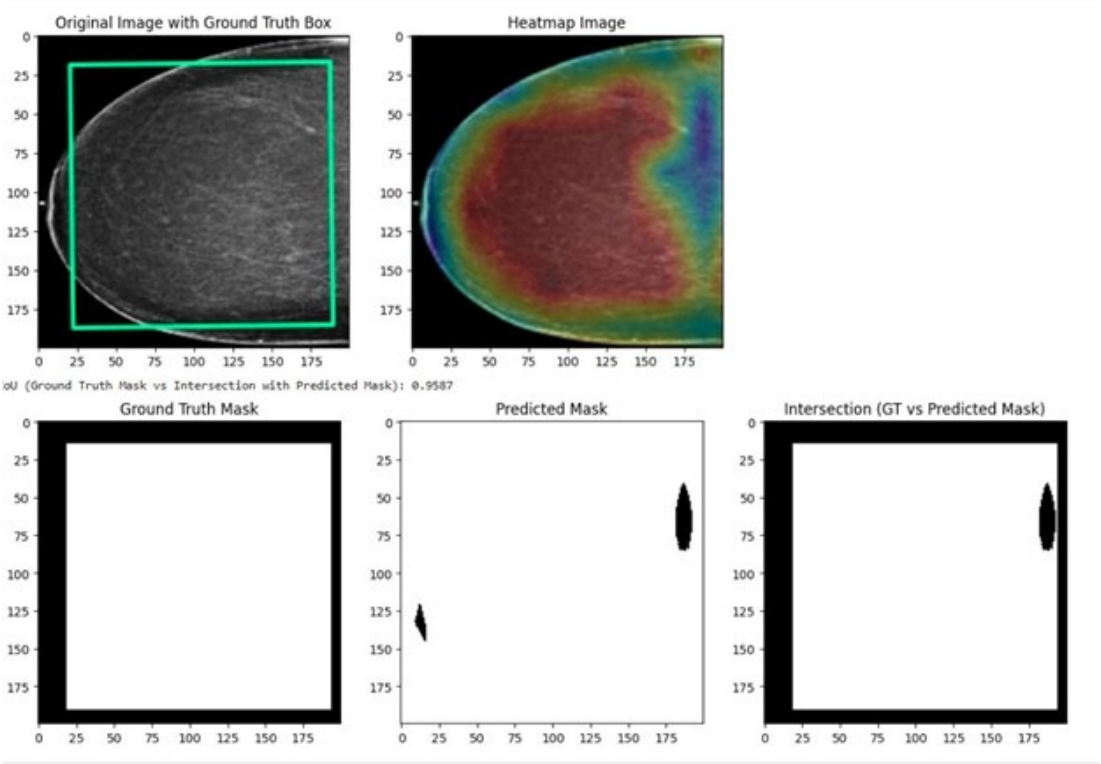

Fig. 4.14: SIDU Results — Demonstrates SIDU-generated saliency maps, emphasizing high-accuracy feature selection in non-cancerous cases for improved transparency in AI predictions.

culated to assess their ability to correctly identify regions of interest. As shown in Table 4.4, SIDU achieved the highest average IoU score of 0.9404, confirming its superior accuracy in selecting relevant saliency features. In contrast, Grad-CAM produced a significantly lower average IoU score of 0.2545, indicating poor performance in localizing critical areas, while Ablation-CAM achieved a moderate score of 0.7554.

To further test the methods in non-cancerous contexts, an additional evaluation was performed using 50 non-cancer test images. The average IoU scores from this experiment are summarized in Table 4.5.Again, SIDU outperformed the other methods with an average IoU of 0.9424. Ablation-CAM maintained reasonable performance with an IoU of 0.7028, while Grad-CAM continued to show weak localization ability, achieving only 0.3387.

These combined results clearly demonstrate that SIDU consistently provides the most accurate and clinically meaningful saliency maps in both cancerous and non-

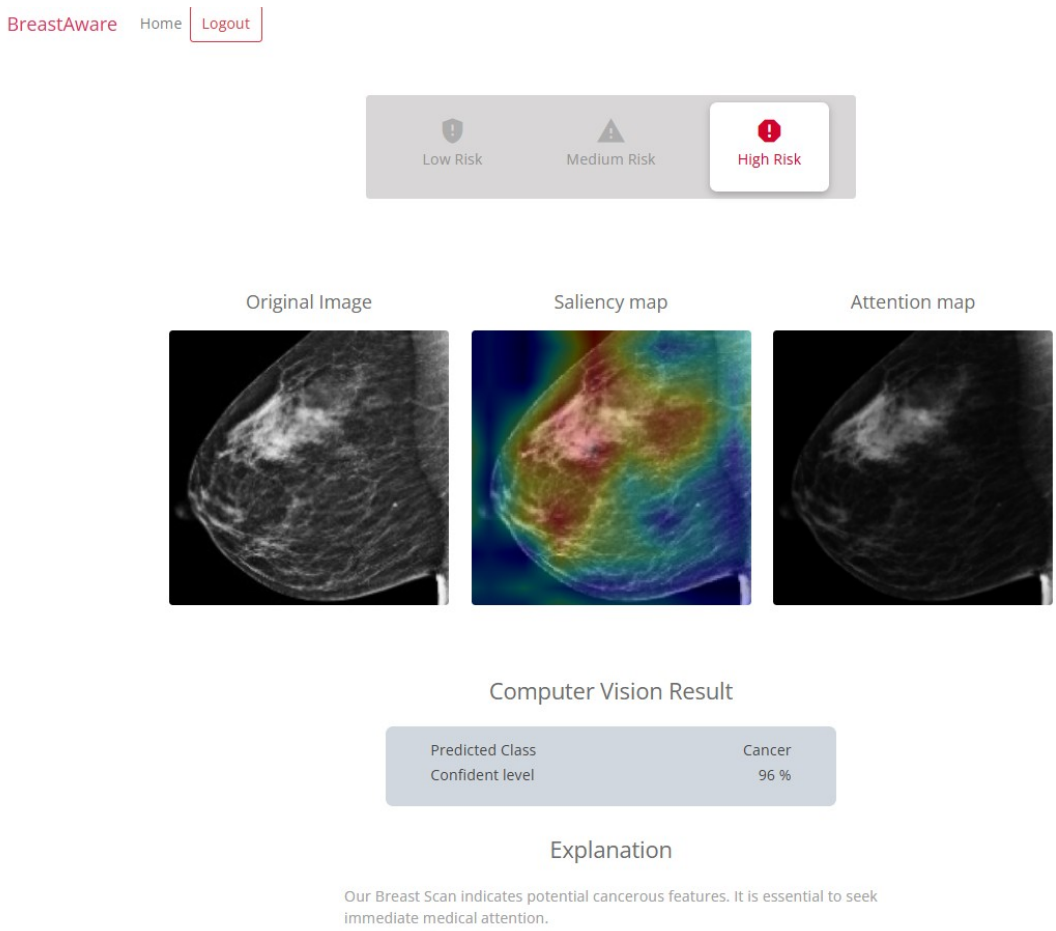

Fig. 4.15: Clinical Validation Output: High-Risk Case

cancerous cases. The findings also highlight the limitations of Grad-CAM, which, despite its popularity, showed inadequate performance in correctly identifying relevant regions, particularly in non-cancerous images.

4.6 Discussion of the Integration with Medical System

Another significant objective of this research project was to develop a clinically viable web-based decision support tool BreastAware, to help medical professionals in the detection of breast cancer by mammogram analysis. The application enables users to upload mammogram images and receive automated diagnostic results powered by deep learning. Upon image upload, BreastAware performs classification using a dual-model ensemble architecture explained in section 3.10 combining a CNN-based InceptionResNetV2 model and a transformer-based ViT model. The system generates interpretable outputs including class prediction (Cancer or No Cancer), confidence level, and visualization overlays such as saliency and attention maps to enhance decision transparency. Additionally, the platform categorizes the overall case risk (Low,

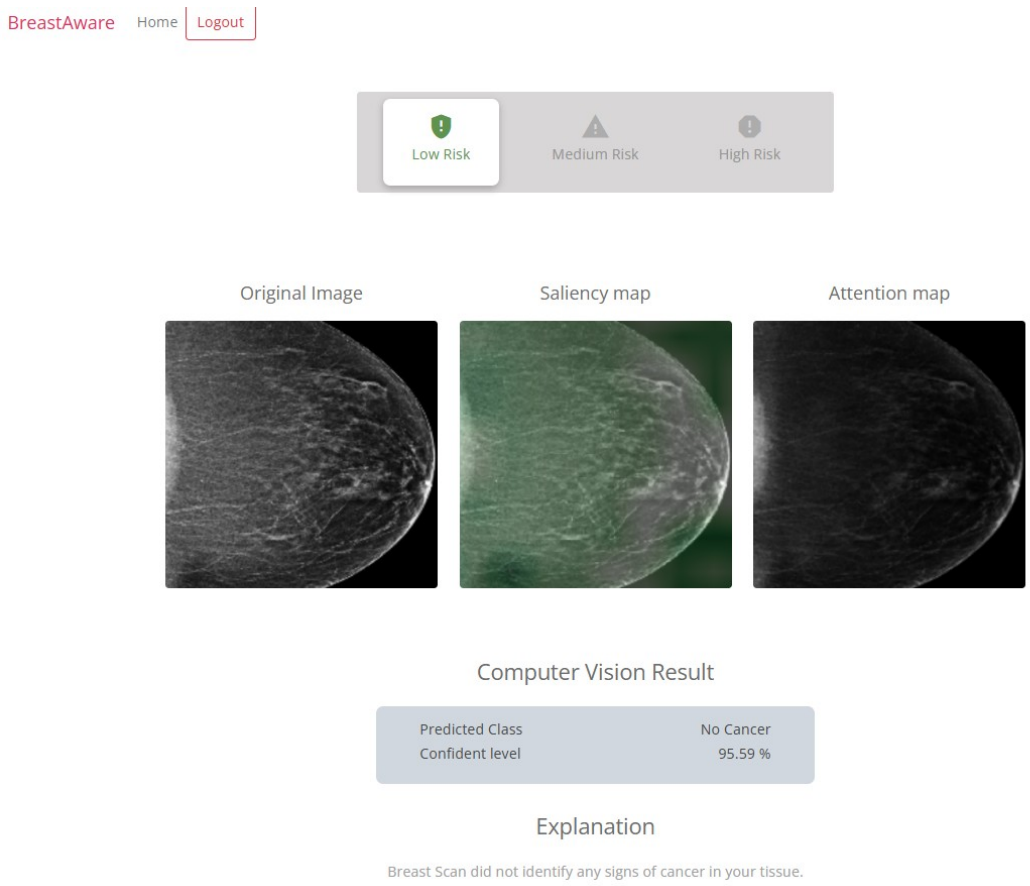

Fig. 4.16: Clinical Validation Output: Low-Risk Case

Medium, or High), providing clinicians with quick triage insights. The final decision, including visual explanations and the system's confidence, is presented to the user within the application interface and stored in a backend database for further review, audits, or clinical discussion. As demonstrated in the figures 4.15, 4.16 this end-to-end pipeline provides not not just high accuracy, but also interpretable and clinically actionable insights.

Combining predictions from a ViT and a CNN-based Inception-ResNet-v2 leverages the unique strengths of each model family. The fusion logic described in the previous sections introduces a structured and adaptive method to resolve disagreement, handle uncertainty, and improve overall performance. The key advantages of this ensemble approach can be categorized into four dimensions: architectural complementarity, performance reliability, uncertainty management, and clinical applicability.

1.Combining Different Model Strengths

CNNs and ViTs learn from images in very different ways. Combining them helps the system benefit from both detailed local features and overall image context.

- **CNN Strengths (InceptionResNetV2):**
  - Excels at detecting edges, textures, and small patterns.
  - Focuses on specific regions like lesions or nodules.
- **ViT Strengths:**
  - Understands relationships across the entire image using attention mechanisms.
  - Useful for detecting scattered or complex patterns that cover multiple areas.
  - Performs best when trained on large datasets.

By combining CNNs and ViTs, the system can detect both fine details and larger image patterns, which is especially important for catching early-stage cancers or small abnormalities.

2. Improving Diagnostic Confidence with Model Fusion

The use of weighted decision fusion significantly enhances diagnostic reliability:

- **Weighted Voting Based on Trust:**
  - Each model's prediction is given a weight based on its performance.
  - More trusted models have a stronger influence on the final decision.
- **Handling Model Disagreements:**
  - When models disagree, the system uses confidence levels and the assigned weights to choose the best answer.
  - This method is especially helpful in difficult or borderline cases.l.
- **Avoids Overconfident Errors:**
  - The system prevents overconfidence from any single model.
  - This reduces the chance of missing cancers, which is critical in medical diagnosis.

3. Managing Uncertainty in Predictions

The system is designed to openly handle uncertainty, which increases safety.

- **Threshold-Based Filtering:**
  - MOnly predictions that meet a certain confidence level are considered in the final decision.
  - Low-confidence predictions are automatically reduced in importance.
- **"Uncertain" Classification Mode:**
  - If the models give unclear or conflicting results, the system can output "Uncertain" instead of forcing a decision.
  - This encourages clinician review and supports follow-up decisions.
- **Focus on Patient Safety**
  - Unlike some AI systems that always provide a final answer, this system is built to admit uncertainty when needed.
  - This safety feature aligns with clinical best practices.

4. Enhanced Clinical Applicability

The ensemble framework has been tailored for real-world deployment in medical imaging and diagnostics:

- **Interpretable Output:**
  - The system provides both the final prediction and the confidence scores.
  - Results can be reviewed, logged, and explained to support medical decision-making.
- **Customizability for Clinical Use-Cases:**
  - Thresholds and model weights can be adapted to specific contexts, including:
    - Prioritize detecting all possible cancers in screening (high sensitivity).
    - Focus on reducing false positives in confirmatory tests (high specificity).
    - Focus on reducing false positives in confirmatory tests (high specificity).
- **Reduced Bias and Model Failure Risks:**
  - CNNs and ViTs fail under different conditions (e.g., CNNs with image blur; ViTs with rare patterns).
  - By combining them, the system becomes more stable and works better across different types of patients and imaging quality.

TABLE 4.6: Comparison Between Single Model and Combined Model Approach

Aspect	**Single Model (CNN or ViT)**	**Combined Model (CNN + ViT)**
Image Understanding	Focuses on either small or large features	Learns both small and large features
Reliability	Can make overconfident errors	More balanced and reliable decisions
Uncertainty Handling	Usually not addressed	Actively managed with thresholds and "Uncertain" option
Borderline Cases	Performance drops when confidence is low	Fusion improves decision-making in difficult cases
Result Transparency	Only provides confidence scores	Provides confidence and model agreement/disagreement information
Clinical Safety	Can miss critical cases with high confidence	Offers a safe "Uncertain" option to alert doctors for review

In summary, as shown in Table 4.6, the combined model using both CNN and ViT with a weighted decision method provides a more reliable and practical solution than using a single model. This approach captures both detailed and overall image patterns, manages uncertainty clearly, and offers easy-to-understand results. The system improves diagnostic safety, handles difficult cases better, and fits well into real clinical settings. Its combination of different model types, clear explanations, and flexible setup makes it a strong option for use in AI-supported radiology, pathology, dermatology, and other medical imaging areas.

CHAPTER 5

CONCLUSION

This section presents a concise overview of the main findings and contributions outlined in the thesis. The main objective of this study was to create an interpretable classification model for breast cancer detection using mammogram images. To achieve this, the study explored and applied CNN and ViT techniques for medical image analysis, specifically for detecting breast cancer. Following the identification of the most effective model, XAI techniques were applied to ensure transparency, interpretability, and trustworthiness in the model's predictions. The research also led to the development of an XAI-enhanced tool aimed at supporting medical professionals in the correct identification of breast cancer using mammogram scans.

In summary, Experiment 1 explored the effectiveness of several pre-trained CNN models, with the InceptionResNetV2 model emerging as the highest-performing model, achieving 92% overall accuracy. MobileNetV2 also performed well with 90% accuracy. VGG16 succeeded in cancer detection, with a precision of 0.95, while DenseNet121 and ResNet models showed good results but with slightly lower accuracy. These findings emphasize the importance of selecting the right CNN architecture, with InceptionResNetV2 and MobileNetV2 proving to be the most effective models for further optimization. Experiment 2 involved selecting the InceptionResNetV2 model for further optimization. After fine-tuning, the model demonstrated exceptional precision of 91% in detecting non-cancerous cases and 100% precision for cancer detection. In Experiment 2, the InceptionResNetV2 model was further optimized, resulting in 91% precision for non-cancerous cases and 100% precision for cancer detection. The model's overall accuracy improved to 95%, showcasing its strong performance and ability to adapt effectively to unseen data, making it a strong candidate for clinical use.

Moreover, ViT technique was chosen for this research due to its ability to capture long-range dependencies in images using an attention mechanism. The ViT model was used to improve breast cancer detection by focusing on important areas in mammogram images. Its attention-based mechanism helps enhance diagnostic accuracy, supporting more reliable and efficient decision-making in clinical settings. The ViT model achieved excellent performance in mammogram classification, with an overall accuracy of 96%. It demonstrated high precision, recall, and F1-scores for both cancer and non-cancer classes, underscoring its reliability for medical diagnostics.

In the final phase of the research, XAI techniques were applied to both the `vit_b_16` and InceptionResNetV2 models to improve interpretability.The ViT model utilizes an attention mechanism to identify and capture on specific regions in the input image when creating predictions. The attention maps visually demonstrate how the model identifies critical areas in the images, showing where the model directs its focus during

classification. For instance, when the model correctly identifies cancerous regions, the attention map highlights these areas, offering a clear understanding of what the model is focusing on during its prediction. This process not only improves the interpretability of the ViT model but also provide valuable descisions, demonstrating its effectiveness in medical image analysis.

For the InceptionResNetV2 model, XAI techniques such as Grad-CAM, Ablation-CAM, and SIDU were used to visualize the model's decision-making process. The performance of SIDU, Grad-CAM, and Ablation-CAM saliency techniques was evaluated using the IoU metric. SIDU outperforms the other methods with a 0.9805 IoU score, demonstrating near-perfect alignment with the ground truth and minimal false positives and negatives. The SIDU approach effectively isolates the key features, minimizing irrelevant areas, making it highly suitable for precise saliency tasks. Eventhough, Grad-Cam is simple and and widely used, it performs poorly, with no overlap with the ground truth and many errors, making it unreliable for accurate saliency detection. Ablation-CAM performs better than Grad-CAM but still includes unnecessary areas and it does not match the precision and reliability of SIDU. Hence, based on the IoU scores and visual results, the SIDU method is the best-performing saliency feature selection technique.

In summary, the research highlights the significance of selecting and optimizing the right CNN models, as well as incorporating newer methods like ViT, for effective mammogram classification. The integration of XAI techniques further enhances model interpretability, ensuring that the predictions made by these models are not only accurate but also understandable, which is crucial for their clinical adoption. This study's findings offer valuable advancements in medical image analysis, supporting the development of more effective and interpretable AI-driven diagnostic tools in healthcare. By combining both traditional and trending AI methods, such as CNNs and ViTs, with interpretability techniques, this provides a comprehensive way to improving diagnostic accuracy and trust in AI-powered healthcare applications.

To improve prediction accuracy and make the system more useful in clinical practice, a combined decision-making model was used by bringing together the results from InceptionResNetV2 and ViT. This approach used the different strengths of both models, leading to more stable and reliable diagnostic outcomes. The developed web application using ensemble model enables both medical professionals and patients to upload and analyze mammogram images with AI-generated predictions. For clinicians, it offers visualizations and interpretability tools that support more informed decision-making. Patients can also use the platform to receive understandable results about their scans. Its user-friendly design fosters better communication between medical professionals and patients, enhancing transparency and building trust in AI-powered diagnostic tools. In addition to the technical evaluations, the developed system was reviewed by medical experts to assess its clinical relevance and practical usefulness.

The initial expert feedback suggests that the system is considered a promising tool for supporting clinical decision-making. However, the responses also indicate the importance of conducting broader validation studies with more radiologists to fully establish its reliability and clinical applicability.

5.1 Limitations and Future Work

Although the results are encouraging for the XAI-assisted digital breast cancer detection platform, several limitations exists that need to be addressed in future work. First, the model is specifically developed for mammogram images, limiting its applicability to other medical test scans, such as CT, MRIs, or ultrasound images. Expanding the model to handle a wider range of imaging modalities could expand its use in diverse clinical settings. A future solution to this limitation could involve integrating DINO v2, a self-supervised learning model capable of learning features from large amounts of unlabeled data. By leveraging DINO v2, the platform can be expanded to handle various medical imaging modalities, improving its generalizability across different clinical settings.

Additionally, while the model performs well with the date used, its capability to handle new, unseen inputs or different populations remains a concern. Ensuring that the model works effectively across a variety of patient demographics and imaging conditions is crucial for its real-world application.

An important area to improve is the accuracy of the ground truth annotations used in training and testing the model. This study used rectangular boxes to mark tumor areas, but these boxes sometimes covered extra, unnecessary tissue. In the future, using polygon-based annotations would be better, as they can more precisely follow the actual shape of the tumor. This would help the model learn more accurately and make the XAI visualizations clearer and more useful for doctors. Using polygons is likely to make the system's results more reliable and acceptable in real clinical practice

In future work, a key area for improvement would be enhancing the ability of XAI techniques to precisely identify and visualize the tumor regions in medical images. While methods like Grad-CAM, SIDU, and Ablation-CAM provide valuable saliency maps, they often lack the ability to pinpoint the exact location of cancerous tissue or the tumor area with high accuracy. Medical-SAM, an interactive segmentation model, can be incorporated to improve the precision of tumor identification by enabling clinicians to quickly and accurately segment abnormal regions with minimal input, allowing for clearer and more actionable insights. Future XAI techniques could be developed or optimized to focus specifically on highlighting the tumor or abnormal regions with greater precision, providing more actionable insights to clinicians.

Furthermore, future work should aim to enhance the sensitivity of these models to detect early stage cancer, where the tumor is not easily visible. Improving model

robustness in detecting these small or low-contrast lesions would significantly improve the platform's clinical value. To enhance sensitivity, the integration of DINO v2 and TransUNet could help the platform detect subtle abnormalities in early-stage cancer by improving feature extraction and segmentation in low-contrast areas, enabling more reliable detection of small lesions.

For future development, this study proposes the integration of an advanced false-negative safeguarding system by combining XAI techniques with a GPT-4o vision-language reasoning agent. This approach aims to provide an additional safety layer to ensure that non-cancer cases are not misclassified and overlooked in clinical workflows.

In conclusion, while the current XAI-assisted platform holds great promise for breast cancer detection, addressing the above limitations and incorporating these advanced techniques and future directions will make it a more robust, accurate, and reliable tool for healthcare professionals, leading to improved patient outcomes.

REFERENCES

[1] C. Wild, E. Weiderpass, and B. W. Stewart, *World cancer report: cancer research for cancer prevention*. International Agency for Research on Cancer, 2020.

[2] W. H. Organization *et al.*, *World Cancer Report: Cancer Research for Cancer Development*. IARC, 2020.

[3] G. Muneeswaran, M. Pandiaraj, S. Kartheeswaran, M. Sankaralingam, K. Muthukumar, and C. Karunakaran, "Molecular dynamics simulation approach to explore atomistic molecular mechanism of peroxidase activity of apoptotic cytochrome c mutants," *Informatics in Medicine Unlocked*, vol. 11, pp. 51–60, Jan 2018.

[4] M. Ataollahi, J. Sharifi, M. Paknahad, and A. Paknahad, "Breast cancer and associated factors: a review," *Journal of medicine and life*, vol. 8, no. Spec Iss 4, p. 6, 2015.

[5] American Cancer Society, "Types of breast cancer: About breast cancer," Available at: https://www.cancer.org/cancer/types/breast-cancer/about/types-of-breast-cancer.html, accessed: 24 July 2024.

[6] WebMD Editorial Contributors, "Understanding breast cancer symptoms," https://www.webmd.com/breast-cancer/understanding-breast-cancer-symptoms, accessed: 2024-07-24.

[7] National Institute of Biomedical Imaging and Bioengineering, "Mammography," https://www.nibib.nih.gov/science-education/science-topics/mammography, accessed: 2024-07-24.

[8] M. M. Alshammari, A. Almuhanna, and J. Alhiyafi, "Mammography image-based diagnosis of breast cancer using machine learning: A pilot study," *Sensors*, vol. 22, 1 2022.

[9] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: A review," *Clinical Imaging*, vol. 37, pp. 420–426, 5 2013.

[10] Z. Guo, J. Xie, Y. Wan, M. Zhang, L. Qiao, J. Yu, S. Chen, B. Li, and Y. Yao, "A review of the current state of the computer-aided diagnosis (cad) systems for breast cancer diagnosis," *Open Life Sciences*, vol. 17, pp. 1600–1611, 1 2022.

[11] M. Epimack, M. He, H. Li, F. Kulwa, and J. Li, "Breast cancer segmentation methods: Current status and future potentials," *BioMed Research International*, vol. 2021, p. 9962109, 2021.

[12] M. P. Sampat, M. K. Markey, and A. C. Bovik, "Computer-aided detection and diagnosis in mammography," in *Handbook of Image and Video Processing, Second Edition*, A. C. Bovik, Ed. Elsevier, 2005, pp. 1195–1217.

[13] L. Sun, H. Sun, J. Wang, S. Wu, Y. Zhao, and Y. Xu, "Breast mass detection in mammography based on image template matching and cnn," *Sensors*, vol. 21, 4 2021.

[14] N. M. Hassan, S. Hamad, and K. Mahar, "Mammogram breast cancer cad systems for mass detection and classification: a review," *Multimedia Tools and Applications*, vol. 81, pp. 20 043–20 075, 6 2022.

[15] M. M. Eltoukhy, I. Faye, and B. B. Samir, "A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation," *Computers in Biology and Medicine*, vol. 42, pp. 123–128, 1 2012.

[16] K. Loizidou, G. Skouroumouni, C. Nikolaou, and C. Pitris, "Automatic breast mass segmentation and classification using subtraction of temporally sequential digital mammograms," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, 2022.

[17] R. Rashmi, K. Prasad, and C. B. K. Udupa, "Breast histopathological image analysis using image processing techniques for diagnostic puposes: A methodological review," *Journal of Medical Systems*, vol. 46, 1 2022.

[18] S. A. Kumar and S. Sasikala, "Review on deep learning-based cad systems for breast cancer diagnosis," *Technology in Cancer Research and Treatment*, vol. 22, 1 2023.

[19] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, pp. 5455–5516, 12 2020.

[20] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific Reports*, vol. 9, 12 2019.

[21] G. Baselli, M. Codari, and F. Sardanelli, "Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way?" *European Radiology Experimental*, vol. 4, 12 2020.

[22] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103

[23] L. Wilkinson and T. Gathani, "Understanding breast cancer as a global health concern," *The British journal of radiology*, vol. 95, no. 1130, p. 20211033, 2022.

[24] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, p. 18, 12 2020.

[25] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, 9 2021.

[26] Q. Zhang and S. Zhu, "Visual interpretability for deep learning: a survey," *Frontiers of Information Technology and Electronic Engineering*, vol. 19, 02 2018.

[27] L. Balkenende, J. Teuwen, and R. M. Mann, "Application of deep learning in breast cancer imaging," *Seminars in Nuclear Medicine*, vol. 52, pp. 584–596, 9 2022.

[28] M. Nasser and U. K. Yusof, "Deep learning based methods for breast cancer diagnosis: a systematic review and future direction," *Diagnostics*, vol. 13, no. 1, p. 161, 2023.

[29] G. Boesch, "Vision transformers (vit) in image recognition–2022 guide," *Viso AI*, 2022.

[30] S. M. Shah, R. A. Khan, S. Arif, and U. Sajid, "Artificial intelligence for breast cancer detection: trends & directions," *arXiv preprint arXiv:2110.00942*, 2021.

[31] E. Luczyńska, S. Heinze, A. Adamczyk, J. Rys, J. W. Mitus, and E. Hendrick, "Comparison of the mammography, contrast-enhanced spectral mammography and ultrasonography in a group of 116 patients," *Anticancer research*, vol. 36, no. 8, p. 4359—4366, August 2016.

[32] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review," *Clinical imaging*, vol. 37, no. 3, pp. 420–426, 2013.

[33] A. Sharma, S. Kulshrestha, and S. Daniel, "Machine learning approaches for breast cancer diagnosis and prognosis," in *2017 International conference on soft computing and its engineering applications (icSoftComp)*. IEEE, 2017, pp. 1–5.

[34] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine learning with applications in breast cancer diagnosis and prognosis," *Designs*, vol. 2, no. 2, p. 13, 2018.

[35] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine learning classification techniques for breast cancer diagnosis," in *IOP conference series: materials science and engineering*, vol. 495. IOP Publishing, 2019, p. 012033.

[36] M. Nasser and U. K. Yusof, "Deep learning based methods for breast cancer diagnosis: a systematic review and future direction," *Diagnostics*, vol. 13, no. 1, p. 161, 2023.

[37] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PloS one*, vol. 12, no. 6, p. e0177544, 2017.

[38] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*. Springer, 2018, pp. 737–744.

[39] C. Zhu, F. Song, Y. Wang, H. Dong, Y. Guo, and J. Liu, "Breast cancer histopathology image classification through assembling multiple compact cnns," *BMC medical informatics and decision making*, vol. 19, pp. 1–17, 2019.

[40] J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, and J. Fan, "Deep learning assisted efficient adaboost algorithm for breast cancer detection and early diagnosis," *IEEE Access*, vol. 8, pp. 96 946–96 954, 2020.

[41] L. Shen, "End-to-end training for whole image breast cancer diagnosis using an all convolutional design," *arXiv preprint arXiv:1711.05775*, 2017.

[42] Y. J. Suh, J. Jung, and B.-J. Cho, "Automated breast cancer detection in digital mammograms of various densities via deep learning," *Journal of personalized medicine*, vol. 10, no. 4, p. 211, 2020.

[43] A. Mohiyuddin, A. Basharat, U. Ghani, V. Peter, S. Abbas, O. B. Naeem, and M. Rizwan, "[retracted] breast tumor detection and classification in mammogram images using modified yolov5 network," *Computational and mathematical methods in medicine*, vol. 2022, no. 1, p. 1359019, 2022.

[44] S. Arooj, M. Zubair, M. F. Khan, K. Alissa, M. A. Khan, and A. Mosavi, "Breast cancer detection and classification empowered with transfer learning," *Frontiers in Public Health*, vol. 10, p. 924432, 2022.

[45] H. N. Khan, A. R. Shahid, B. Raza, A. H. Dar, and H. Alquhayz, "Multi-view feature fusion based four views model for mammogram classification using convolutional neural network," *IEEE Access*, vol. 7, pp. 165 724–165 733, 2019.

[46] P. Xi, C. Shu, and R. Goubran, "Abnormality detection in mammography using deep convolutional neural networks," in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2018, pp. 1–6.

[47] H. Chougrad, H. Zouaki, and O. Alheyane, "Convolutional neural networks for breast cancer screening: transfer learning with exponential decay," *arXiv preprint arXiv:1711.10752*, 2017.

[48] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, "A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique," *IEEe Access*, vol. 9, pp. 71 194–71 209, 2021.

[49] M. A. Al-Antari, M. A. Al-Masni, and T.-S. Kim, "Deep learning computer-aided diagnosis for breast lesion in digital mammogram," *Deep Learning in Medical Image Analysis: Challenges and Applications*, pp. 59–72, 2020.

[50] A. Sahu, P. K. Das, and S. Meher, "Recent advancements in machine learning and deep learning-based breast cancer detection using mammograms," *Physica Medica*, vol. 114, p. 103138, 2023.

[51] H. N. Huynh, N. A. D. Nguyen, A. T. Tran, V. C. Nguyen, and T. N. Tran, "Classification of breast cancer using radiological society of north america data by efficientnet," *Engineering Proceedings*, vol. 55, no. 1, p. 6, 2023.

[52] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman, and O. Al-Kadi, "Vision transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases," *Journal of Imaging Informatics in Medicine*, pp. 1–44, 2025.

[53] Y. Huo, K. Jin, J. Cai, H. Xiong, and J. Pang, "Vision transformer (vit)-based applications in image classification," in *2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2023, pp. 135–140.

[54] M. Cantone, C. Marrocco, F. Tortorella, and A. Bria, "Convolutional networks and transformers for mammography classification: an experimental study," *Sensors*, vol. 23, no. 3, p. 1229, 2023.

[55] M. L. Abimouloud, K. Bensid, M. Elleuch, O. Aiadi, and M. Kherallah, "Vision transformer-convolution for breast cancer classification using mammography images: A comparative study," *International Journal of Hybrid Intelligent Systems*, vol. 20, no. 2, pp. 67–83, 2024.

[56] G. Ayana, K. Dese, Y. Dereje, Y. Kebede, H. Barki, D. Amdissa, N. Husen, F. Mulugeta, B. Habtamu, and S.-W. Choe, "Vision-transformer-based transfer learning for mammogram classification," *Diagnostics*, vol. 13, no. 2, p. 178, 2023.

[57] M. L. Abimouloud, K. Bensid, M. Elleuch, O. Aiadi, and M. Kherallah, "Vision transformer-convolution for breast cancer classification using mammography images: A comparative study," *International Journal of Hybrid Intelligent Systems*, vol. 20, no. 2, pp. 67–83, 2024.

[58] B. Gheflati and H. Rivaz, "Vision transformers for classification of breast ultrasound images," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 480–483.

[59] A. Alotaibi, T. Alafif, F. Alkhilaiwi, Y. Alatawi, H. Althobaiti, A. Alrefaei, Y. Hawsawi, and T. Nguyen, "Vit-deit: An ensemble model for breast cancer histopathological images classification," in *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*. IEEE, 2023, pp. 1–6.

[60] S. Tummala, J. Kim, and S. Kadry, "Breast-net: Multi-class classification of breast cancer from histopathological images using ensemble of swin transformers. mathematics, 10 (21), 4109, 2022," 2022.

[61] S. S. Boudouh and M. Bouakkaz, "Advancing precision in breast cancer detection: a fusion of vision transformers and cnns for calcification mammography classification," *Applied Intelligence*, vol. 54, no. 17, pp. 8170–8183, 2024.

[62] O. Tanimola, O. Shobayo, O. Popoola, and O. Okoyeigbo, "Breast cancer classification using fine-tuned swin transformer model on mammographic images," *Analytics*, vol. 3, no. 4, pp. 461–475, 2024.

[63] S. Paavankumar, R. Karthik, G. Idayachandiran, P. P. D. Sri, and T. Illakiya, "Classification of benign and malignant breast lesions in mammograms using

dense-unified multiscale attention network and data-efficient image transformers," *The European Physical Journal Special Topics*, pp. 1–19, 2025.

[64] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[65] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[66] Y. Wan, Y. Yang, H. Guo, Y. Yan, T. Liu, W. Liu, Y. Wang, W. Wang, and H. Dang, "D-transunet: A breast tumor ultrasound image segmentation model based on deep feature fusion," *Journal home: http*, vol. 5, no. 1-2, pp. 01–08, 2024.

[67] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[68] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Incorporating explainable artificial intelligence (xai) to aid the understanding of machine learning in the healthcare domain." in *Aics*. Seattle, WA, USA, 2020, pp. 169–180.

[69] E. Dağlarli, "Explainable artificial intelligence (xai) approaches and deep meta-learning models," in *Advances and applications in deep learning*. IntechOpen, 2020.

[70] I. Shivhare, V. Jogani, J. Purohit, and S. C. Shrawne, "Analysis of explainable artificial intelligence methods on medical image classification," in *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2023, pp. 1–5.

[71] 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.

[72] M. I. Hossain, G. Zamzmi, P. R. Mouton, M. S. Salekin, Y. Sun, and D. Goldgof, "Explainable ai for medical data: Current methods, limitations, and future directions," *ACM Computing Surveys*, 2023.

[73] P. Gohel, P. Singh, and M. Mohanty, "Explainable ai: current status and future directions," *arXiv preprint arXiv:2107.07045*, 2021.

[74] K. Raghavan, "Attention guided grad-cam: an improved explainable artificial intelligence model for infrared breast cancer detection," *Multimedia Tools and Applications*, vol. 83, no. 19, pp. 57 551–57 578, 2024.

[75] S. Desai and H. G. Ramaswamy, "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 972–980.

[76] V. Petsiuk, "Rise: Randomized input sampling for explanation of black-box models," *arXiv preprint arXiv:1806.07421*, 2018.

[77] S. M. Muddamsetty, N. J. Mohammad, and T. B. Moeslund, "Sidu: Similarity difference and uniqueness method for explainable ai," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3269–3273.

[78] S. Pertuz, D. Ortega, É. Suarez, W. Cancino, G. Africano, I. Rinta-Kiikka, O. Arponen, S. Paris, and A. Lozano, "Saliency of breast lesions in breast cancer detection using artificial intelligence," *Scientific Reports*, vol. 13, no. 1, p. 20545, 2023.

[79] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMlR, 2017, pp. 3145–3153.

[80] F. A. Imouokhome, O. G. Ehimiyein, and F. O. Chete, "Diagnosis and interpretation of breast cancer using explainable artificial intelligence," *NIPES-Journal of Science and Technology Research*, vol. 5, no. 2, 2023.

[81] R. Kashefi, L. Barekatain, M. Sabokrou, and F. Aghaeipoor, "Explainability of vision transformers: A comprehensive review and new perspectives," *arXiv preprint arXiv:2311.06786*, 2023.

[82] F. M. Talaat, S. A. Gamel, R. M. El-Balka, M. Shehata, and H. ZainEldin, "Grad-cam enabled breast cancer classification with a 3d inception-resnet v2: Empowering radiologists with explainable insights," *Cancers*, vol. 16, no. 21, p. 3668, 2024.

[83] H. Mankodiya, D. Jadav, R. Gupta, S. Tanwar, W.-C. Hong, and R. Sharma, "Od-xai: Explainable ai-based semantic object detection for autonomous vehicles," *Applied Sciences*, vol. 12, no. 11, p. 5310, 2022.

[84] Y. S. Lin, W. C. Lee, and Z. B. Celik, "What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 1027–1035.

[85] M. A. Talukder, "An improved xai-based densenet model for breast cancer detection using reconstruction and fine-tuning," *Results in Engineering*, p. 104802, 2025.

[86] J. Hou, S. Liu, Y. Bie, H. Wang, A. Tan, L. Luo, and H. Chen, "Self-explainable ai for medical image analysis: A survey and new outlooks," *arXiv preprint arXiv:2410.02331*, 2024.

[87] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "Data descriptor: A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, 12 2017.

[88] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (tcia): Maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, pp. 1045–1057, 12 2013.

[89] S. Gengtian, B. Bing, and Z. Guoyou, "Efficientnet-based deep learning approach for breast cancer detection with mammography images," in *2023 8th International Conference on Computer and Communication Systems (ICCCS)*, 2023, pp. 972–977.

[90] A. W. Salehi, S. Khan, G. Gupta, B. I. Alabduallah, A. Almjally, H. Alsolai, T. Siddiqui, and A. Mellit, "A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope," *Sustainability*, vol. 15, no. 7, p. 5930, 2023.

[91] M. Z. Hanane and M. Mejdeded, "Utilization of pre-trained models of cnn in mammograms processing for the diagnosis of breast cancer," in *2022 7th International Conference on Image and Signal Processing and their Applications (ISPA)*, 2022, pp. 1–5.

[92] R. Agarwal, O. Diaz, X. Lladó, and R. Martí, "Mass detection in mammograms using pre-trained deep learning models," in *14th International workshop on breast imaging (IWBI 2018)*, vol. 10718. SPIE, 2018, pp. 376–381.

[93] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[94] L. Tsochatzidis, L. Costaridou, and I. Pratikakis, "Deep learning for breast cancer diagnosis from mammograms—a comparative study," *Journal of Imaging*, vol. 5, no. 3, p. 37, 2019.

[95] A. Anaya-Isaza, L. Mera-Jiménez, and M. Zequera-Diaz, "An overview of deep learning in medical imaging," *Informatics in medicine unlocked*, vol. 26, p. 100723, 2021.

[96] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[97] Q. A. Al-Haija and A. Adebanjo, "Breast cancer diagnosis in histopathological images using resnet-50 convolutional neural network," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2020, pp. 1–7.

[98] Y. Chen, Q. Zhang, Y. Wu, B. Liu, M. Wang, and Y. Lin, "Fine-tuning resnet for breast cancer classification from mammography," in *Proceedings of the 2nd International Conference on Healthcare Science and Engineering 2nd*. Springer, 2019, pp. 83–96.

[99] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[100] X. Yuan, L. Zhang, and S. Zhao, "Densenet convolutional neural network for breast cancer diagnosis," in *Proceedings of the 2022 3rd International Conference on Artificial Intelligence and Education (IC-ICAIE 2022)*, vol. 9. Springer Nature, 2023, p. 197.