UNIVERSIDADE FEDERAL FLUMINENSE

Leonardo Martins Reigoto

A platform for breast diagnosis from frontal thermographies

NITERÓI

2026

UNIVERSIDADE FEDERAL FLUMINENSE

Leonardo Martins Reigoto

A platform for breast diagnosis from frontal thermographies

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação.

Orientador:
Aura Conci

NITERÓI

2026

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

R347p Reigoto, Leonardo Martins
A platform for breast diagnosis from frontal thermographies / Leonardo Martins Reigoto. - 2026.
82 f.

Orientador: Aura Conci.
Dissertação (mestrado)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2026.

1. Inteligência artificial. 2. Aprendizado de máquina. 3. Neoplasia da mama. 4. Produção intelectual. I. Conci, Aura, orientadora. II. Universidade Federal Fluminense. Instituto de Computação. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Leonardo Martins Reigoto

A platform for breast diagnosis from frontal thermographies

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação.

Aprovada em Junho de 2026.

BANCA EXAMINADORA

Prof. Aura Conci - Orientador, UFF

Prof. Vanessa Braganholo, UFF

Prof. Leonardo Murta, UFF

Prof. Geraldo Braz Júnior, UFMA

Prof. João Dallyson Almeida, UFMA

Niterói

2026

Resumo

O câncer de mama continua sendo um desafio significativo para a saúde global, no qual a detecção precoce é primordial. Embora a mamografia seja o método padrão de rastreamento, suas limitações em custo, acessibilidade e eficácia em tecido denso têm motivado a busca por alternativas. A termografia médica, uma modalidade de imagem não invasiva e livre de radiação, apresenta-se como uma ferramenta complementar promissora. Contudo, a transição de modelos de aprendizado de máquina de protótipos de pesquisa para aplicações clínicas confiáveis é dificultada por desafios em reprodutibilidade, escalabilidade e rigor metodológico.

Esta tese enfrenta esses desafios através do desenvolvimento de uma plataforma robusta baseada em Operações de Aprendizado de Máquina (MLOps) e microserviços, que integra o ciclo de vida completo de modelos de Deep Learning, desde a ingestão e pré-processamento automatizado até à disponibilização de uma aplicação funcional de inferência. O sistema utiliza um conjunto de ferramentas modernas incluindo Apache Airflow, MLflow, Docker e FastAPI para automatizar e gerenciar todo o ciclo de vida do modelo. Uma contribuição central deste trabalho é sua estrita aderência a uma metodologia de separação de dados por paciente, que aborda o problema crítico do vazamento de dados (data leakage) prevalente na literatura anterior e garante uma avaliação válida da generalização do modelo.

Um modelo baseado na arquitetura EfficientNet-B0 foi desenvolvido e avaliado, utilizando uma estratégia de treinamento multiestágio com o descongelamento progressivo de camadas. Em um conjunto de dados de imagens de mamas segmentadas, o modelo alcançou um desempenho promissor, com uma acurácia de aproximadamente 92.29%, uma precisão de 100% e um recall de 85.83%.

Em última análise, esta tese fornece um plano abrangente para construir, validar e implantar sistemas de aprendizado de máquina robustos e confiáveis em um contexto médico. Ela demonstra um caminho claro da pesquisa acadêmica para um auxílio diagnóstico prático e acessível, contribuindo com um framework metodologicamente sólido e uma aplicação de prova de conceito (*Proof of Concept*) funcional para futuras inovações em aplicações computacionais de imagens médicas.

Palavras-chaves: Câncer de mama, imagens médicas, termografia, aprendizado de máquina, deep learning, MLOps, CNN (Convolutional Neural Network), pipeline automatizada, deploy de modelo.

Abstract

Breast cancer remains a significant global health challenge where early detection is paramount. While mammography is the standard screening method, its limitations in cost, accessibility, and efficacy in dense tissue have motivated the search for alternatives. Medical thermography, a non-invasive and radiation-free imaging modality, presents a promising complementary tool. However, the translation of machine learning models from research prototypes into reliable clinical applications is hindered by challenges in reproducibility, scalability, and methodological rigor.

This thesis confronts these challenges through the development of a robust platform based on Machine Learning Operations (MLOps) and microservices, integrating the complete Deep Learning model lifecycle, from automated ingestion and preprocessing to the deployment of a high-performance inference application. The system leverages a modern stack including Apache Airflow, MLflow, Docker, and FastAPI to automate and govern the entire model lifecycle. A core contribution of this work is its strict adherence to a patient-wise data separation methodology, which addresses the critical issue of data leakage prevalent in prior literature and ensures a valid assessment of model generalization.

A model based on the EfficientNet-B0 architecture was developed and evaluated, using a multi-stage training strategy with progressive layer unfreezing. On a segmented breast dataset, the model achieved promising performance, with an accuracy of approximately 92.29%, a precision of 100%, and a recall of 85.83%.

Ultimately, this thesis provides a comprehensive blueprint for building, validating, and deploying robust and reliable machine learning systems in a medical context. It demonstrates a clear pathway from academic research to a functional diagnostic Proof of Concept (PoC), contributing a methodologically sound framework for future innovation in computational medical imaging.

Keywords: Breast cancer, medical imaging, thermography, Machine Learning, Deep Learning, MLOps, CNN (Convolutional Neural Network), pipeline automation, model deployment.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Breast cancer stands as one of the most significant public health crises of our time, yet early and accessible detection remains a challenge. The current gold standard, mammography, is constrained by limitations in cost, patient discomfort, and reduced effectiveness in dense breast tissue. This dissertation confronts this issue by developing a practical diagnostic aid using medical thermography, a non-invasive, radiation-free alternative that detects physiological signs of malignancy. The core of this work is not simply another predictive model, but a complete, end-to-end Machine Learning Operations (MLOps) pipeline engineered to bridge the critical gap between academic research and a real-world, deployable application. This chapter establishes the primary research questions and outlines the key contributions of this methodologically rigorous approach.

1.1 The challenge in breast cancer screening

Breast cancer represents one of the most significant public health issues of our time. It stands as the most prevalent cancer among women globally, responsible for staggering rates of diagnosis and mortality annually. Decades of clinical research have unequivocally demonstrated that early detection [58] is the most critical factor in improving patient outcomes and survival rates. However, achieving early diagnosis remains a formidable challenge, particularly in lower-resource regions where access to advanced medical infrastructure is limited [58, 10].

The current gold standard for breast cancer screening is mammography. While it has been instrumental in detection programs worldwide, its effectiveness is constrained by several well-documented limitations. The procedure involves exposure to ionizing radiation, can be uncomfortable for the patient, and its diagnostic accuracy is notably

reduced in women with dense breast tissue, a characteristic common in younger patients. Furthermore, factors related to high cost and the need for specialized equipment and personnel limit its accessibility and widespread use. These challenges underscore the urgent need for alternative or complementary screening methods that are safer, effective and more accessible.

1.2 A non-invasive alternative: medical thermography

Medical thermography presents a promising, non-invasive imaging modality for breast cancer screening. Unlike mammography, which identifies structural anomalies, thermography detects physiological changes. It operates on the principle that cancerous tissues, due to their higher metabolic rate and associated angiogenesis (the formation of new blood vessels), present diverse heat transfer behavior than surrounding healthy tissue. This propagates from the unheat cells to the skin surface and can be captured by highly sensitive infrared cameras, creating a thermal map of the breast [37, 61].

The primary advantages of thermography stem from its fundamental difference from conventional imaging: it is a passive modality. Unlike X-rays or mammography, which are active techniques that transmit energy through the body to create an image, thermography functions strictly as a receiver. It captures the natural infrared radiation emitted by the skin surface, a thermodynamic characteristic inherent to all physical bodies with a temperature above absolute zero. Because it relies entirely on this endogenous heat emission rather than external radiation, the procedure is non-invasive and radiation-free. This safety profile makes it suitable for frequent monitoring and safe for patients of all demographics, including pregnant women. Despite its potential, the widespread clinical adoption of thermography has been historically hindered by the subjectivity and variability of manual thermogram interpretation. The recent convergence of advanced infrared imaging technology and the pattern-recognition power of machine learning (ML), however, has renewed interest in its potential to serve as a reliable, objective, and automated diagnostic aid [29, 61].

1.3 An end-to-end MLOps pipeline

Translating a machine learning model from a research prototype into a robust, deployable application is a notoriously difficult task. This thesis directly confronts this challenge

by focusing not just on developing a predictive model, but on architecting a distributed system to support it.

The core of this work is the design and implementation of a scalable, reproducible, and accessible machine learning pipeline built upon a microservices architecture. This system is engineered a modern MLOps (Machine Learning Operations) stack to manage the complete model lifecycle, from automated data ingestion to production-grade deployment, including:

- Apache Airflow for orchestrating complex workflows.
- MLflow for tracking experiments and managing model lifecycles.
- Docker for ensuring environmental consistency and reproducibility.
- PyTorch Lightning for structuring and simplifying deep learning code.
- FastAPI for serving the model via a high-performance API.
- Gradio for creating an intuitive user interface for real-time interaction.

A basic principle of this research is a commitment to methodological rigor. Consequently, this work explicitly addresses the significant challenge of data leakage, which is the unintentional contamination of the training dataset with information from the test set, a potential limitation also addressed by recent studies (e.g., Çevik et al., 2024 [61]). We identified two primary sources of potential data leakage within the context of our study.

The first and most critical source stems from performing data splits at the image level rather than the patient level. Our dataset is characterized by high intra-patient image similarity, containing approximately 27 images for each subject. These include 5 static images, of which only one is a frontal view, and 22 dynamic images (which mixes frontal and lateral images). A simple split would inadvertently distribute these highly correlated images from the same patient across both training and validation/testing sets. The second potential source of leakage is the application of data augmentation techniques before partitioning the data. This sequence can cause synthetically generated variants of a training image to appear in the test set, compromising the integrity of the model evaluation.

To prevent these issues, we implemented a strict patient-wise separation for all data splits. This ensures that all images belonging to a single patient are confined to only one

set (i.e., training, validation, or testing). As a result, our model's performance is a more reliable reflection of its ability to generalize to unseen patients. This approach is crucial for avoiding overly optimistic performance estimates that can arise when models learn patient-specific, rather than truly generalizable, features.

1.4 Research questions

This work seeks to answer the following primary research questions:

- Can a modern MLOps pipeline orchestrated as a microservices ecosystem effectively address the common challenges of reproducibility, scalability, and asynchronous deployment in medical imaging tasks?
- How does an EfficientNet-B0 deep learning model perform for breast cancer classification using thermographic images when evaluated under a methodologically rigorous, patient-separated protocol?
- To what extent does a strict patient-wise data separation methodology impact model performance metrics, and how do these results compare to previously published studies where data leakage is a potential confounding factor?

1.5 Text structure

This dissertation is organized into eight chapters, each building upon the last to present a comprehensive account of the research.

- Chapter 2: Background and context, provides the scientific and technical context, reviewing medical imaging modalities, the physiological basis of thermography, foundational deep learning concepts, and the principles of MLOps, while critically examining the issue of data leakage in existing literature.
- Chapter 3: Literature Review presents a review of the latest developments in Computer-Aided Diagnosis (CAD) for breast thermography, focusing on the shift toward deep learning architectures and emerging challenges in clinical deployment.
- Chapter 4: System Architecture and MLOps Pipeline details the technical framework of the project. It provides a comprehensive overview of the end-to-end pipeline

and its microservices-based design, explaining the role of each component, including Docker for reproducibility, Apache Airflow for workflow orchestration, MLflow for experiment tracking, and FastAPI and Gradio for model deployment and user interaction.

- Chapter 5: Dataset and Experimental Methodology describes the data and methods used. It details the primary and secondary datasets, the ethical considerations, and the strict patient-wise partitioning strategy implemented to prevent data leakage. t also presents the EfficientNet-B0 model architecture, adapted for this task using a transfer learning and multi-step fine-tuning strategy, and outlines the preprocessing, training, and evaluation procedures.

- Chapter 6: Results and Analysis presents the empirical findings of the study. It provides a quantitative analysis of the model's performance on the test set, compares it against benchmarks, and presents a series of ablation studies to validate architectural and preprocessing choices. A visual explanation of the model decisions using Grad-CAM is also included within the functional diagnostic application.

- Chapter 7: Discussion interprets the significance of the results. It analyzes the clinical implications of the model's performance, revisits the importance of methodological rigo

- Chapter 8 - Conclusion and Future Work: The final chapter summarizes the research, restates the key contributions, and proposes a roadmap for future future research directions, including data enhancement, model improvements, and eventual clinical validation.

Chapter 2

Background and context

This chapter provides the scientific and technical context for the research presented in this thesis. It begins by establishing the clinical need for innovative and accessible breast cancer screening alternatives, examining the primary imaging modalities and their limitations. The chapter then examines into the principles of medical thermography, exploring its physiological basis and the evolution from subjective manual interpretation to objective, computer-aided diagnosis. Following this, a review of the foundational deep learning concepts. Then proceeds to a critical review of the state-of-the-art in automated thermographic analysis using deep learning is presented. Special attention is given to a prevalent and critical methodological flaw in the existing literature: data leakage, which systematically inflates reported performance and obscures the true capabilities of these models. Finally, the chapter introduces the principles of Machine Learning Operations (MLOps) as a necessary framework for bridging the significant gap between academic research prototypes and the reliable, reproducible, and deployable systems essential for translation into clinical practice. This comprehensive review serves to identify the key research gaps that this thesis aims to address, positioning its contributions at the intersection of methodological rigor and engineering for deployment.

2.1 Context of breast thermography

The development of advanced computational tools for medical diagnosis does not occur in a vacuum, it is driven by pressing clinical needs and the unique opportunities afforded by specific imaging technologies. This section focus on understanding the landscape of breast cancer screening, the limitations of current standards, and the scientific rationale for exploring thermography as a complementary modality.

2.1.1 Breast cancer screening

Breast cancer remains one of the most significant public health challenges globally, with early detection being the single most critical factor in improving patient outcomes and survival rates. For decades, mammography has been the undisputed gold standard for population-wide screening programs. This X-ray-based imaging technique has been instrumental in reducing mortality by detecting structural abnormalities such as microcalcifications and masses that are often indicative of early-stage tumors.

However, the efficacy of mammography is not universal, and its limitations are well-documented. A primary concern is its reduced diagnostic sensitivity in women with dense breast tissue, a characteristic common in younger patient populations [27]. Dense tissue can mask suspicious lesions on an X-ray, leading to a higher rate of false negatives. Furthermore, the procedure involves exposing the patient to low doses of ionizing radiation [38], which, while minimal, is a cumulative concern over a lifetime of screening. Patient discomfort due to the necessary breast compression is another significant factor that can affect compliance with screening recommendations. Finally, the high capital cost of mammography equipment and the need for specialized radiological personnel limit its accessibility, particularly in lower-resource regions where the need for effective screening is just as urgent.

Beyond these limitations, it is important to recognize that diagnostic accuracy improves when combining imaging modalities based on different physical principles. While mammography relies on X-ray attenuation (anatomical density), other methods leverage distinct physical properties—such as acoustic reflection in ultrasound or magnetic resonance response in MRI. Because these modalities provide complementary information, utilizing more than one type can significantly resolve uncertainties in difficult cases. These combined challenges, alongside the need for multimodal complementarity, create a clear and persistent clinical need for alternative or complementary screening methods that are safer, more comfortable, cost-effective, and equally or more effective across diverse patient populations.

2.1.2 Thermography as a functional imaging modality

Medical thermography, also known as infrared imaging, has emerged as a promising candidate to address many of the limitations of mammography. Unlike mammography, which provides a structural map of the breast, thermography is a functional imaging modal-

ity [45]. It is non-invasive, involves no ionizing radiation, is entirely painless, and is significantly more cost-effective, replacing expensive X-ray machinery with sensitive infrared cameras.

Its diagnostic potential is rooted in the unique physiology of malignant tumors, but is governed by the fundamental principles of bioheat transfer. An infrared camera captures the thermal radiation emitted from the skin surface [45]. This temperature is the result of three complex internal processes: heat generated by basal tissue metabolism (conduction), heat transported to and from the tissue by blood perfusion (convection), and heat exchanged with the environment (radiation).

The foundational model describing this process is the Pennes bioheat equation [57]. This model describes the energy balance in tissue, treating metabolic heat generation as a distributed source and, critically, modeling the convective heat transfer from blood perfusion as a primary factor in tissue temperature. Pennes work remains the cornerstone for understanding how internal physiological changes, such as the increased blood flow and metabolic rate associated with angiogenesis in tumors, manifest as detectable thermal anomalies on the skin surface. This thesis, therefore, uses deep learning to detect the complex patterns resulting from these fundamental heat transfer processes.

Thermographic protocols can be broadly categorized into static and dynamic approaches [11]. Static thermography involves capturing a thermal image after the patient has reached a state of thermal equilibrium with the ambient environment. Dynamic thermography involves applying a thermal stress (such as a cooling stimulus) and recording the temporal pattern of rewarming, as cancerous tissues may respond differently than healthy tissues. The work of Munguía-Siu et al. [37], which uses a hybrid CNN-RNN model, shows the promise of this approach for analyzing temporal thermal data.

While dynamic thermography is an active area of research exploring temporal data with hybrid CNN-RNN models, static thermography remains a foundational and widely studied approach. In this thesis, we focus on static thermography for our first experiments and a mix of segmented static and dinamic thermographies as an alternative approach.

2.1.3 From manual interpretation to computer-aided diagnosis (CAD)

Despite its theoretical advantages and approval by the FDA as an adjunctive diagnostic tool in 1982 [51], the widespread clinical adoption of thermography has been historically

hindered.

A primary barrier was the inherent subjectivity and high inter-observer variability associated with the manual interpretation of thermograms. The subtle and complex nature of thermal patterns made it difficult to establish standardized, reproducible diagnostic criteria, leading to inconsistent results.

The modern resurgence of interest in thermography is a direct result of the convergence of two technological advancements: the development of more sensitive, high-resolution infrared cameras and the revolutionary pattern-recognition capabilities of machine learning, particularly deep learning [49]. This has given rise to the field of Computer-Aided Diagnosis (CAD) for thermography, which aims to provide an objective, automated, and reliable analysis of thermal images. By training algorithms on large datasets of labeled thermograms, CAD systems can learn to identify the subtle, complex, and often non-intuitive thermal signatures associated with malignancy, overcoming the limitations of human interpretation and unlocking the full potential of thermography as a diagnostic aid [47]. This thesis is situated firmly within this modern paradigm, seeking not only to develop such an algorithm but to engineer the system required to make it a practical and reliable tool.

2.2 Deep learning in medical image analysis

The advent of deep learning has fundamentally transformed the field of medical image analysis, and its application to breast thermography is no exception [20, 30, 23]. These techniques have moved the field beyond traditional machine learning, which required manual, expert-driven feature engineering, to an end-to-end paradigm where relevant features are learned automatically from the data. A comprehensive understanding of the architectures and methods employed in the literature is essential to contextualize the technical choices made in this thesis.

2.2.1 Convolutional neural networks (CNNs)

Convolutional Neural Networks (CNNs) are a class of deep neural networks that have become the standard for image recognition and classification tasks. Their architecture is inspired by the human visual cortex and is designed to automatically and adaptively learn spatial hierarchies of features from images.

The fundamental building blocks of a CNN are:

- Convolutional Layers: These layers apply a set of learnable filters (or kernels) to the input image. Each filter is a small matrix of weights that slides over the input image, computing the dot product between the filter and the input at each position. This operation, called a convolution, produces a feature map that highlights specific features in the image, such as edges, corners, and textures.
- Activation Functions: After each convolution operation, a non-linear activation function, such as the Rectified Linear Unit (ReLU), is applied. ReLU introduces non-linearity to the model, allowing it to learn more complex patterns.
- Pooling Layers: These layers are used to downsample the feature maps, reducing their spatial dimensions. This helps to make the learned features more robust to small variations in the input image and reduces the computational complexity of the network. Common pooling operations include max pooling and average pooling.
- Fully Connected Layers: After several convolutional and pooling layers, the high-level features are flattened and fed into one or more fully connected layers. These layers are standard multi-layer perceptrons that perform the final classification based on the learned features

The application of CNNs to breast thermography has been extensively explored [31, 49]. Researchers have progressively adopted more sophisticated and powerful architectures as they have emerged from the broader computer vision community. Early studies often employed foundational models like VGG-16 and VGG-19, known for their simple and deep structure. Subsequently, architectures incorporating more complex concepts, such as the residual connections in ResNet variants (e.g., ResNet-50) and the feature reuse in DenseNet, were shown to achieve superior performance by enabling the training of deeper and more effective networks. This progression demonstrates a clear trend in the field: leveraging state-of-the-art, general-purpose vision models as powerful feature extractors for the specific task of thermogram classification.

2.2.2 The choice of EfficientNet

For this work, we chose the EfficientNet-B0 architecture, a CNN model that achieves high accuracy with a relatively small number of parameters. The EfficientNet architecture was introduced by Mingxing Tan and Quoc V. Le. [53] and represents a signifcant

advancement in the CNN models. This family of models are designed based on the principle of compound scaling, which systematically scales the depth, width, and resolution of the network in a balanced way to optimize performance. This approach was shown to yield models that achieve not only higher accuracy but also significantly better computational efficiency, requiring fewer parameters and floating-point operations (FLOPs) than previous architectures for a given level of performance

2.3 The critical role of preprocessing

The ultimate performance of a deep learning system for medical imaging is not determined by the model architecture alone. The methods used to prepare the data and train the model are equally, if not more, critical. A review of the literature reveals several auxiliary techniques that have become standard best practices.

First, image segmentation is widely recognized as a crucial preprocessing step. Raw frontal thermograms often include significant background regions, such as the neck, shoulders, chest wall, and arms, which are irrelevant to the diagnostic task and act as noise. By employing a segmentation model (such as a U-Net) to first isolate the breast regions, the subsequent classification model can focus its learning capacity exclusively on the relevant physiological information, leading to improved accuracy and robustness. The marked improvement in performance observed in this thesis when using the pre-segmented DMR-IR dataset corroborates this finding, highlighting that an effective segmentation module is a key component of a high-performing pipeline.

Second, transfer learning is an almost universally adopted strategy. Training a deep CNN from scratch requires an enormous dataset, often on the order of millions of images, which is rarely available in medical domains. Transfer learning circumvents this by starting with a model that has been pre-trained on a large, general-purpose dataset like ImageNet. The underlying principle is that the initial layers of this pre-trained model have already learned to recognize fundamental visual features (edges, textures, gradients) that are applicable to most image recognition tasks. By freezing these initial layers and fine-tuning the deeper, more task-specific layers (and a new classification head), this learned knowledge can be transferred to the new domain, leading to faster convergence and better generalization, even with a limited medical dataset.

The multi-step fine-tuning strategy with progressive unfreezing employed in this thesis is a sophisticated application of this core principle in a gradual and controlled manner.

This is achieved through a multi-step training process where different blocks of the network are sequentially unfrozen and trained. The later layers of the network learn more complex and abstract features, and by fine-tuning them first, we allow the model to adapt to the high-level characteristics of thermographic images. As we unfreeze earlier layers, the model can adjust the more fundamental feature extractors to better suit the nuances of our data.

Finally, data augmentation is an essential technique for preventing overfitting and improving model generalization. By applying random, yet realistic, transformations to the training images—such as horizontal flips, rotations, and zooms—a wider variety of synthetic examples are created. This artificially expands the training set, making the model more robust to variations in patient positioning, orientation, and scale that it will encounter in real-world data. The ablation studies conducted in this thesis, which show a dramatic improvement in performance when augmentation is used, confirm its status as an indispensable component of the training methodology.

2.4 The methodological imperative: data leakage and the pursuit of valid generalization

While the selection of powerful architectures and the application of best-practice training techniques are important, they are rendered meaningless if the model's evaluation is not conducted with unimpeachable methodological rigor. A critical review of the medical imaging literature reveals a pervasive and fundamental flaw that calls into question a significant portion of reported results: data leakage. This section deconstructs this issue, analyzes its impact on the field, and establishes the methodological framework that underpins the credibility of this thesis.

2.4.1 Defining data leakage in medical imaging

Data leakage, in the context of machine learning, is the unintentional use of information from outside the training dataset to train a model [52]. It occurs when the model is exposed, directly or indirectly, to data from the test set during its training or selection phase. This violation of the separation between training and testing environments creates a model that appears to perform exceptionally well during evaluation but is incapable of generalizing to new, truly unseen data. When deployed in a real-world scenario, such a model will fail, often dramatically, because its high performance was an illusion created

by "cheating" on the test [12].

In medical imaging, and particularly in thermography, the most common and insidious form of data leakage arises from an incorrect data splitting strategy. A single patient's diagnostic session often yields multiple images (e.g., static and dynamic views, different angles). These images, all from the same individual, are highly correlated; they share the same underlying anatomy, breast shape, and unique vascular patterns. If the dataset is split into training and test sets randomly at the image level, it is virtually guaranteed that images from the same patient will be distributed across both sets.

The consequence of this flawed, image-wise split is severe. The model does not learn the generalizable thermal signatures of malignancy. Instead, it learns to identify patients. It memorizes the unique, person-specific anatomical features of Patient A from the images in the training set. When it then encounters other images of Patient A in the test set, it achieves a correct classification not by diagnosing a condition, but by simply recognizing the individual. This leads to dangerously optimistic performance metrics that have no bearing on the model's ability to diagnose a new, unseen patient. The only scientifically valid method for evaluating a diagnostic model is to enforce a strict patient-wise (or subject-wise) data separation, ensuring that all data from any given patient belongs exclusively to either the training set or the test set, but never both.

2.4.2 The illusion of performance

A systematic survey of the literature on deep learning for breast thermography reveals a troubling pattern: an abundance of studies reporting near-perfect or superhuman performance. It is not uncommon to find papers claiming classification accuracies of 95%, 98%, 99%, or even 100%. While seemingly impressive, these figures should be met with extreme skepticism, as they are often symptomatic of the methodological flaw of data leakage.

This phenomenon is not unique to thermography. Rigorous scoping reviews in other medical imaging domains, such as Alzheimer's disease diagnosis [12] from neuroimaging, have systematically investigated this issue and uncovered a stark inverse correlation between methodological rigor and reported accuracy. These reviews found that studies with a high risk of data leakage (e.g., those using image-wise splits or failing to specify their splitting strategy) were precisely the ones claiming 95-99% accuracy. In contrast, the small subset of studies that explicitly confirmed a proper, subject-wise data split reported accuracies in a much more realistic and clinically plausible range of 66-90%. One

study highlighted in these reviews directly demonstrated the effect: when evaluated with a leaky, image-wise split, their model achieved 94% accuracy; when re-evaluated correctly with a patient-wise split, the accuracy plummeted by 28 percentage points to 66%.

This evidence from parallel fields strongly suggests that the breast thermography literature is suffering from a similar "illusion of performance". Indeed, recent surveys of the thermography literature itself have begun to note this problem, cautioning that many reported high-accuracy results may be unreliable due to methodological flaws such as data leakage [47]. The proliferation of near-perfect results is likely not a sign that the problem of breast cancer classification from thermograms has been solved, but rather an indication that many studies are being evaluated using flawed methodologies that produce artificially inflated and misleading metrics. This not only hinders genuine scientific progress by creating unrealistic expectations but also poses a significant risk if these flawed models are considered for clinical translation.

2.5 Establishing a methodologically sound baseline

The antidote to this problem is a commitment to methodological rigor. This thesis explicitly adopts this commitment. By enforcing a strict patient-wise data separation for all training, validation, and test splits, it ensures that the reported performance metrics represent a valid and honest assessment of the model's ability to generalize to new, unseen patients.

The work of Çevik et al. (2024) [61], which is used as a key benchmark in this thesis, is a notable example of a study that also implements a patient-wise separation strategy to ensure methodological integrity. By positioning its results alongside this methodologically sound subset of the literature, this thesis helps to establish a more realistic, and trustworthy performance benchmark for the field.

2.6 Machine learning operations (MLOps): bridging research and practice

Achieving a high-performing model under rigorous validation is a necessary, but not sufficient, condition for clinical impact. A significant gap exists between a model trained in a research environment—often a static entity represented by a set of saved weights—and a robust, reliable, and scalable tool that can be safely integrated into a clinical workflow.

This thesis addresses this "research-to-practice" gap through the systematic application of Machine Learning Operations (MLOps), transitioning from a model-centric to a system-centric approach. By implementing a microservices-based architecture, the developed system ensures that experimental insights are automatically and reliably translated into functional clinical endpoints.

Machine Learning Operations (MLOps) is a discipline focused on closing this gap by applying DevOps principles to the machine learning lifecycle. This thesis embraces MLOps to ensure the developed system is scalable, reproducible, and accessible [28].

Key MLOps tools and principles used in this work include:

- Orchestration (Apache Airflow): Automates the entire workflow, from data processing and model training to evaluation and registration, using Directed Acyclic Graphs (DAGs). This ensures a reliable and repeatable process. In this work, the orchestration goes beyond training, incorporating an automated "Model Delivery Pipeline" that triggers upon model promotion in the registry, ensuring that the latest production-validated weights are seamlessly synchronized with the serving layer.

- Experiment Tracking and Governance (MLflow): Provides a central repository for logging all experiment parameters, metrics, and model artifacts, ensuring full reproducibility. It includes a feature called the Model Registry, which, in the context of Machine Learning Operations (MLOps), is a centralized system for managing the lifecycle of machine learning models [59]. The use of the MLflow Model Registry enables a governed transition between "Staging" and "Production" environments, a critical requirement for medical software where version control, auditability, and the tracking of model lineage are essential.

- Reproducibility (Docker): Utilizes containerization to package the application and its dependencies into a standardized unit. This guarantees that the training and serving environments are identical, eliminating inconsistencies across different machines.

- User Interaction and Feedback Loops (Gradio): This setup enables "stakeholder-in-the-loop validation", allowing medical professionals to interact with the model Explainable AI (XAI) output, Grad-CAM heatmaps, and enables them to send their feedback.

Chapter 3

Literature Review

3.1 Introduction

To contextualize the contributions of this thesis, this chapter presents a review of the latest developments in Computer-Aided Diagnosis (CAD) for breast analysis based on infrared imaging (IR) or termography. Concentrating on literature from 2024 and 2025, it examines the shift toward deep learning architectures and the emerging challenges in clinical deployment. This review is organized thematically to provide a comparative evaluation of current methodologies, contrasting them with established baselines and highlighting the critical need for reproducible, leakage-free MLOps pipelines in medical imaging.

3.2 Evolving approaches in computer-aided diagnosis for breast using infrared thermography

The application of machine learning to breast thermography, known as Computer-Aided Diagnosis (CAD), has evolved significantly. This evolution can be broadly categorized into two main eras: traditional machine learning with hand-crafted features and the modern end-to-end deep learning paradigm.

Early foundational work in thermographic CAD focused on traditional machine learning models. These approaches required a two-step process: first, domain experts would manually engineer and extract relevant features from the thermal images. Then, these features were fed into a classifier.As summarized in a recent survey by Ryan and Agaian (2025) [47], common features included statistical measures, asymmetry analysis, and advanced texture descriptors. Models such as Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN) were frequently employed to classify these hand-crafted fea-

tures. While these systems showed initial promise, their performance was fundamentally limited by the quality and completeness of the engineered features, which often failed to capture the full complexity of thermal patterns.

The shift to deep learning, revolutionized this field by enabling end-to-end feature learning. As this thesis has already noted, researchers began leveraging powerful, pre-trained CNNs. Architectures such as VGG-16, ResNet, and Inception-V3, all originally trained on ImageNet, were widely adopted as feature extractors. Recent studies continue to explore and refine these standard architectures, with Latha and Mahesh (2025) [55] successfully applying a VGG16 model and using Grad-CAM for explainability, and Al Husaini et al. (2022) [3] exploring models based on the Inception architecture.

Recent efforts (2024-2025) have begun to explore more specialized or hybrid architectures. For example, recognizing that thermography data can be captured over time (dynamic thermography), Munguía-Siu et al. (2024) [37] proposed a hybrid CNN-RNN model to analyze temporal thermal data. This body of work, recently surveyed by Mashekova et al. (2025) [32], shows a clear trend of architectural exploration to improve classification accuracy.

3.3 Modern Advances in Thermography and Deep Learning

The following subsections categorize significant recent contributions into distinct methodological streams. This organization facilitates a critical analysis of how specific subdomains from end-to-end classifiers to generative data augmentation, have evolved over the 2024–2025 period.

3.3.1 CNN-Based Advances and End-to-End Classifiers

3.3.1.1 Alzahrani et al. (2025)

Alzahrani and colleagues [4] introduced a CNN framework, optimized via Enhanced Particle Swarm Optimization (EPSO), for the automated classification of breast thermograms into specific abnormal conditions. Their model benefitted from an enhanced preprocessing pipeline incorporating Mamdani fuzzy logic and Contrast-Limited Adaptive Histogram Equalization (CLAHE) which improved edge detection and contrast rather than simply stabilizing variations. A particular strength of their work is its clinical alignment; the

authors utilized a dataset with documented acquisition conditions captured via a FLIR SC-620 camera.

However, despite these strengths, the study relied on a single-center dataset from the Database for Mastology Research, which may limit generalizability. Their model has not yet been validated across multi-institutional datasets or different camera types, leaving open questions about deployment robustness.

3.3.1.2 Attallah (2025)

This study [9] proposed "Thermo-CAD", a framework that integrates deep features from three pre-trained CNN architectures (Inception, ResNet-101, and InceptionResNet) to enhance diagnostic reliability.

Unlike prior works that rely on complex segmentation, this study obviates those steps, instead employing Non-Negative Matrix Factorization (NNMF) for feature transformation and Relief-F for feature selection to reduce dimensionality. The model achieved 100% accuracy on the DMR-IR dataset (normal vs. abnormal) but dropped to 79.3% accuracy when classifying benign versus malignant tumors on a second dataset. A major limitation identified was the model's diminished specificity (54.3%) in the benign-malignant task, leading to a high rate of false positives.

3.3.1.3 The production and deployment in recent literature

A common thread in the 2024-2025 literature, including the works of Alzahrani [4] and Attallah [9], is the focus on model-centric metrics while overlooking the system-centric challenges of deployment. Most studies present models as static artifacts trained in monolithic environments. There is a noticeable absence of discussion regarding how these multiconvolutional networks would scale in a hospital network or how model drift would be monitored. This thesis addresses this specific gap by transitioning from a single-model approach to a microservices-based ecosystem, ensuring that high-performance models (like those described by Attallah) can be reliably served and updated via automated pipelines.

3.3.2 Real-Time and Sensor-Driven Thermography

In Real-time thermography for breast cancer detection with deep learning, a study exploring real-time thermography (2024), Al Husaini et al. [2] developed a video-based pipeline

utilizing in situ cooling to enhance thermal contrast. Their modified Inception v4 model (Mv4) reached an accuracy of 99.748%, significantly outperforming standard baselines.

However, it is critical to note that these results were obtained in a controlled simulation using a silicone breast phantom with an internal lamp representing the tumor, rather than in a clinical setting. Consequently, while the dataset of 1,000 thermal frames demonstrates the theoretical potential of dynamic heat recovery analysis, the findings reflect performance on synthetic data rather than clinical screening capability

3.3.3 Explainable and Interpretable Methods

In Fully Interpretable Deep Learning Model Using IR Thermal Images for Possible Breast Cancer Cases (2024), Mirasbekov et al. [34] proposed a hybrid model integrating Convolutional Neural Networks (CNNs) and Bayesian Networks (BNs) to combine deep learning predictions with interpretable causal inference mechanisms. Their incorporation of clinical variables and Explainable AI (XAI) algorithms enhances model interpretability, creating decision pathways that clinicians can interrogate. "Model B" (CNN + BN) achieved 90.93% accuracy, a notable performance for an interpretable system. Distinct from studies limited to single-source data, their validation utilized a diverse dataset comprising images from a public database and local acquisitions captured via varying camera models (IRTIS 2000ME and FLUKE TiS60+) . Furthermore, rather than sacrificing efficiency, the authors reported that the Bayesian Network models demonstrated efficiency and robustness during K-fold cross-validation. Their work supports this thesis's emphasis on architecture transparency, noting that the primary limitation was the relatively small dataset size rather than the computational cost of explainability

3.3.4 Data Augmentation and GAN-Driven Approaches

In A hybrid GAN-based deep learning framework for thermogram-based breast cancer detection (2025) [56], Veerlapalli and Dutta used GANs to generate high-quality synthetic Regions of Interest (ROIs), addressing the well-known scarcity of breast thermography datasets. These synthetic samples improved classification accuracy and training stability, enabling models to learn diverse thermal patterns.

Still, GANs introduce risk: if synthetic images fail to represent realistic thermal signatures, models may overfit to GAN artifacts. Their paper lacks a thorough discussion of distributional fidelity metrics. For this reason, despite GANs usefulness, this thesis

opts for rigorous cross-validation and leakage prevention rather than relying on aggressive synthetic augmentation.

3.3.5 Reviews and Meta-Analyses (2024–2025)

3.3.5.1 Rahman et al. (2025)

In this review [42] Rahman et al, surveys ML and DL techniques across mammography, ultrasound, MRI, and thermography between 2022–2024. They identify trends toward hybrid models, mobile-ready architectures, and transformer-based pipelines. For thermography, the review identifies a need for interpretability and data augmentation.

Yet, because it is a broad survey, it lacks depth on thermography-specific operational issues such as calibration drift, protocol variance, and thermal camera heterogeneity.

3.3.5.2 Goñi-Arana et al (2024)

This meta-analysis [25] of 22 thermography studies (2001–2023) reports pooled sensitivity of 88.5% and rising specificity in recent years. The key insight is that while thermography has potential as an adjunct, heterogeneity in imaging protocols remains a critical barrier.

This finding strongly supports the methodology of this thesis, which places emphasis on strict acquisition consistency, temperature normalization, and leakage-free validation to minimize such heterogeneity.

3.3.6 The DMR-IR Database and UFF Contributions

Since the Database for Mastology Research with Infrared Image (DMR-IR) dataset was developed and made publicly available by the research group at Universidade Federal Fluminense (UFF)[15], a significant and continuous body of work has emerged directly from this institution. These contributions have established foundational baselines for thermogram analysis and driven the field forward through various master's and doctoral researches.

The development of robust preprocessing pipelines and feature extraction has been a continuous focus for researchers utilizing the DMR-IR dataset. For instance, Baffa et al.[18] demonstrated the critical role of automated breast segmentation using adaptive refinement thresholding, achieving a segmentation accuracy of 93.00% and a sensitivity of 98.00% directly on this database. This emphasis on accurate region-of-interest isolation

aligns perfectly with the findings of this thesis, which demonstrates that evaluating models on pre-segmented datasets significantly improves downstream diagnostic classification. Similarly, Mendes et al.[33] explored Region of Interest (ROI) extraction in thermographic images employing Genetic Algorithms. More recently, the research group has expanded into spatial analysis, with Moura et al.[35] developing a tool for the 3D representation of 2D thermographic breast acquisitions, providing new structural perspectives on thermal distribution.

In the domain of classification and diagnosis, several studies have explored traditional and hybrid machine learning techniques. Principalli[41] investigated the impact of bilateral symmetry by combining Haralick descriptors and Uniform Local Binary Patterns (LBP) with Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN). Resmini et al.[43] advanced this approach by proposing hybrid methodologies, including combining Genetic Algorithms with SVM for breast cancer diagnosis, demonstrating how classical methods can be optimized for thermal feature classification. Furthermore, Silva et al.[16] introduced specific computational methods aimed at breast abnormality detection using static thermographs.

The UFF group has also made extensive contributions to the analysis of dynamic thermography and temperature time series. Silva et al.[17] proposed a hybrid analysis for indicating breast cancer using temperature time series, a concept later expanded into a comprehensive computational method to assist in breast disease diagnosis using dynamic thermography[50]. Building on the temporal analysis of thermal data, Araújo et al. have extensively researched the use of thermography for monitoring clinical interventions. Through a series of publications[5, 6, 7], they demonstrated how thermographic time series can be effectively utilized to monitor and assess the progression of breast cancer neoadjuvant treatments.

Finally, aligning closely with the objectives of this thesis, there has been a recent push within the group towards developing integrated computational systems for clinical aid. Santos et al.[48] proposed a comprehensive computing platform to analyze breast abnormalities using infrared images. This historical progression—from database creation, advanced preprocessing, and hybrid classification to treatment monitoring and the development of dedicated platforms—highlights a critical evolution in the field. It paves the way for the transition from manual feature extraction to the automated, end-to-end representation learning and MLOps deployment provided by architectures like the EfficientNet-B0 developed in this study.

Chapter 4

System Architecture

A successful machine learning model is more than just an algorithm, it is a component within a larger, well-engineered system. Moving a research prototype into a practical tool requires a robust and scalable framework. This chapter provides a detailed technical description of the end-to-end system designed and implemented for this thesis: an MLOps pipeline designed for reproducibility and accessibility. Moving beyond the theoretical model, it focuses on the engineering framework required to create a scalable, reproducible, and deployable machine learning application. The goal is to bridge the gap between a research prototype and a practical tool.

We will explore the high-level architecture before dissecting each component of the MLOps stack. This includes the containerization strategy for ensuring reproducibility, the orchestration engine that automates the workflow, the system for experiment tracking and model governance, the high-performance serving API, and the intuitive user interface. Each technological choice is justified by its role in building a robust and cohesive pipeline.

4.1 High-level architecture overview

The system is designed as a modular pipeline where specialized tools handle specific stages of the machine learning lifecycle. The overall architecture, depicted in Figure 4.1, can be conceptualized as two primary workflows: the MLOps Pipeline for model creation and management, and the Serving Pipeline for real-time inference.

The MLOps Pipeline, orchestrated by Apache Airflow, automates the process of data handling, preprocessing, model training, and evaluation. A successful new model is versioned and promoted within the MLflow Model Registry.

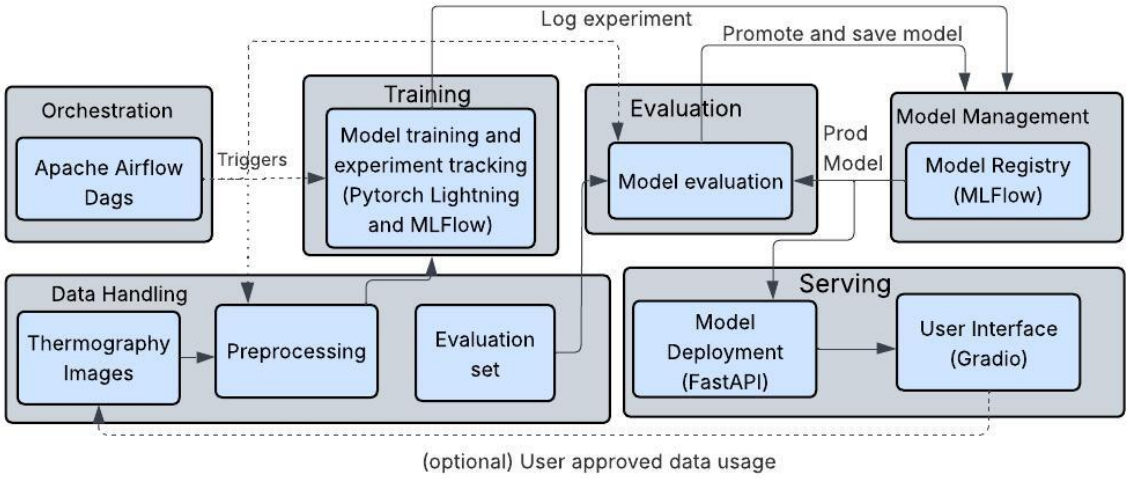

Figure 4.1: High-level system architecture diagram.

The Serving Pipeline begins when the FastAPI application loads the current "production" model from the registry. It exposes this model to the Gradio user interface, which allows users to upload new images and receive predictions.

4.2 Distributed Microservices Architecture

The technical core of this thesis is the transition from a monolithic script to a distributed microservices architecture. Each component operates as an independent service, orchestrated via Docker Compose, within a virtualized Docker network, ensuring isolation of concerns and specialized resource allocation. This design ensures that high-performance computing requirements for training deep neural networks do not hinder the accessibility and responsiveness of the inference layer.

The orchestration layer, managed by Apache Airflow, acts as the central coordinator, triggering tasks across different Docker containers and ensuring data consistency between the experimental phase and the production environment.

4.3 Containerization for reproducibility (Docker and Docker Compose)

Reproducibility is a foundational principle of this work. To eliminate inconsistencies between development, testing, and production environments, the entire system is con-

tainerized. Docker is used to package each component, including the training scripts, the MLflow server, the FastAPI application, and the Gradio interface, into lightweight, portable containers. Each container includes the application code along with all its libraries and dependencies, guaranteeing that it runs identically regardless of the host machine's configuration. Furthermore, Docker Compose is used to define and manage this multi-container application. It uses a single configuration file (docker-compose.yml) to orchestrate the services, enabling simplified deployment and scaling while managing the networking between containers so they can communicate seamlessly

4.4 Workflow orchestration (Apache Airflow)

Apache Airflow serves as the central orchestrator, automating the end-to-end machine learning workflow from model training to deployment. It uses Directed Acyclic Graphs (DAGs) to define, schedule, and monitor these workflows, ensuring that all tasks are executed in the correct order while managing their specific dependencies.

The system implements a Trigger-based Cross-DAG architecture to maintain a strict separation of concerns between model development and service delivery. The primary training DAG executes a sequence of data ingestion, preprocessing, and GPU-intensive training. Upon completion, a performance evaluation task acts as a Gatekeeper: it compares the metrics of the new candidate against the current production model using a held-out test set. If the performance is superior, the system programmatically registers and promotes the new artifact to the "Production" stage within the MLflow Model Registry. This promotion automatically updates the model served to users via the Gradio interface, the inference API utilizes a background polling mechanism that checks for model updates every ten minutes, ensuring seamless transitions without manual intervention.

If promotion occurs, a TriggerDagRunOperator is activated to launch a secondary, decoupled DAG. This second workflow, illustrated in Figure 4.2, handles model export and cloud synchronization to the demonstration environment. By decoupling these workflows, the system ensures that any transient failure in the deployment or synchronization phase does not necessitate a re-execution of the computationally expensive training stage, thereby enhancing the overall resilience and operational efficiency of the MLOps pipeline.

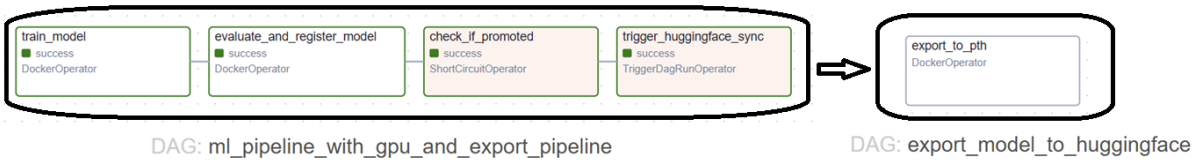

Figure 4.2: Example Airflow DAG for training and deployment.

4.5 Experiment tracking and model management (MLflow)

MLflow is the central piece of our system's governance and reproducibility, functioning as both a detailed digital lab notebook and a formal model library.

4.5.1 Experiment tracking

Integrated directly with PyTorch Lightning, MLflow Tracking automatically logs crucial information for every single training run executed by Airflow. This systematic logging ensures that every experiment is reproducible and comparable. Logged data includes parameters, metrics, and artifacts. The parameters consist of the hyperparameters used for the run, such as learning rate, batch size, and optimizer choice. The metrics include performance metrics calculated on the validation set at each epoch, including loss, accuracy, precision, and recall. Artifacts refer to any output file, most importantly the trained model itself, but also including configuration files, code versions, and visualization plots.

4.5.2 Model registry

The MLflow Model Registry provides a centralized repository for managing the lifecycle of trained models. It provides a formal structure for versioning and transitioning models between stages such as "Staging" and "Production". After the evaluation task in our Airflow DAG, a new model is registered. If its performance on the patient-separated test set is superior to the existing production model, it is programmatically promoted to the "Production" stage. This controlled process ensures that only validated, high-performing models are served to end-users.

4.6 Model serving (FastAPI)

To make the best-trained model accessible for predictions, a high-performance prediction API is built using FastAPI. FastAPI was chosen for its asynchronous nature, which allows

it to handle a high volume of concurrent requests with low latency, a critical feature for an interactive application.

Upon startup, the application queries the MLflow Model Registry and loads the latest version of the model marked as "Production". The API exposes two primary inference endpoints:

- */predict*: A lightweight endpoint that returns the classification score and diagnosis in JSON format, optimized for rapid numerical analysis.
- */predict_with_gradcam*: A specialized endpoint that returns both the prediction and the processed Grad-CAM heatmap. This endpoint is essential for Explainable AI (XAI), allowing the user to visualize which regions of the thermogram influenced the model's decision.

Both endpoints are secured via API key authentication. For observation and management, the API also provides a */health* endpoint for monitoring and an */info* endpoint to retrieve metadata about the currently served model. To ensure the API always serves the best available model, a background task periodically checks the MLflow Registry for a newly promoted production model. If a new version is found, the application automatically reloads it without requiring a restart.

4.7 User interface (Gradio)

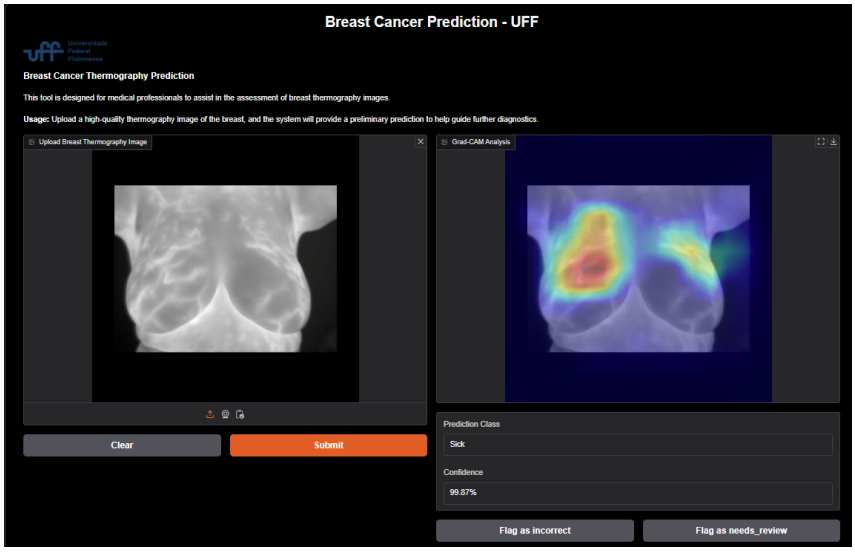

Figure 4.3: Gradio user interface for prediction.

To facilitate straightforward user interaction and testing, an intuitive web interface was rapidly developed using Gradio. This interface serves as a simple front-end that communicates with the FastAPI backend (see Figure 4.3). The user workflow begins when the user visits the web page and is presented with an image upload widget. After uploading an image and clicking "Submit", the Gradio application base64 encodes the image and sends it via a POST request to the */predict_with_gradcam* endpoint of the FastAPI server, which includes the necessary authentication key. The API then returns the predicted label, prediction confidence, and the XAI (Grad-CAM) heatmap. These results are clearly displayed, providing the user with the model's classification, its confidence level, and a visual explanation of the regions that most influenced the decision. This component demonstrates how a complex backend can be made accessible, enabling feedback collection and model testing without requiring front-end development expertise.

4.8 Decoupled demonstration environment (HuggingFace Spaces)

In addition to the local serving infrastructure, the system includes a decoupled demonstration environment hosted on HuggingFace Spaces. The primary intent of this component is to provide an easily accessible, "zero-setup" web entry point for stakeholders, such as thesis advisors and medical collaborators, to interact with the model during the development lifecycle.

This environment is intentionally isolated from the main MLOps pipeline to serve as a pure demonstration layer. When a new model is promoted and the export DAG is triggered, the model artifact, encapsulated with its preprocessing logic via the ModelWithPipeline wrapper, is pushed to the HuggingFace repository. This setup allows for rapid feedback collection and clinical testing without requiring the stakeholders to access the internal Docker network or manage local dependencies. Furthermore, by serving the demo on an external cloud infrastructure, the system demonstrates its infrastructure-agnostic nature, proving that the diagnostic tool can maintain its operational integrity across different hosting tiers.

4.9 Chapter Summary

The architecture presented in this chapter moves beyond traditional research scripts by establishing a distributed MLOps ecosystem designed for clinical reliability. By decoupling the system into specialized microservices, this work addresses the three primary technical barriers identified in the literature review:

1. Environmental Inconsistency: Through the use of Docker and multi-stage containerization, we eliminated the "it works on my machine" problem. The separation of training (CUDA/GPU) and inference (Python-slim/CPU) environments ensures that the diagnostic tool remains portable and stable across different infrastructure tiers. Furthermore, this decoupled approach enables the scaling of inference resources to meet higher user demand without unnecessary overhead or creating resource bottlenecks between specialized services.

2. Lack of Traceability: The integration of MLflow as a centralized governance server provides a complete audit trail for every model. This ensures that every prediction made in the final application can be traced back to its specific training parameters, dataset version, and validation metrics, satisfying the high standards required for medical software.

3. Deployment Friction: The automated delivery pipeline, orchestrated by Apache Airflow, ensures that methodological rigor is maintained even during updates. By automating the transition from the local registry to the serving layer, we ensure that the end-user (the clinician) always interacts with the most accurate, production-validated version of the system.

In summary, this architecture serves as the functional backbone of the thesis, transforming a deep learning model into a transparent, reproducible, and accessible diagnostic aid. This technical foundation provides the necessary framework for the experimental results and performance analysis that will be presented in the subsequent chapters.

Chapter 5

Dataset and Experiments

A robust and reliable machine learning model is determined not only on its architecture but also on the quality of the data it is trained on and the rigor of the methodology used to evaluate it.

This chapter is divided into three main parts. First, it provides a comprehensive description of the thermography dataset, including its source, ethical considerations, cleaning protocol, and the strict partitioning strategy used to prevent data leakage. Second, it presents an in-depth view of the model architecture. Third, it outlines the complete experimental methodology, detailing the image preprocessing pipeline, the training procedure (and its hyperparameters), and the specific metrics used to evaluate the model's performance.

5.1 Dataset

The dataset is the cornerstone of any supervised learning task. The data used in this work was acquired through formal partnerships with medical institutions, handled with strict ethical oversight, and curated to ensure quality.

5.1.1 Source and ethical considerations

The dataset was acquired at the UFF University Hospital (Hospital Universitario Antonio Pedro - HUAP) and the "Hospital Federal dos Servidores do Estado do Rio de Janeiro" (HFSE-RJ). All data collection was conducted under projects approved by the Brazilian Ministry of Health, governed by the "Certificado de Apresentação para Apreciação Ética" (CAAE) numbers 01042812.0.0000.5243 and 04134918.7.0000.5252 [8]. All participants

were volunteers who provided informed consent by signing the "Termo de Consentimento Livre e Esclarecido" (TCLE), granting permission for their anonymized images to be used in this research.

5.1.2 Data cleaning protocol

The raw dataset was carefully curated to ensure a high-quality sample for model training and evaluation. The original set of images was cleaned by removing images with improper capture quality, such as those affected by motion blur or low contrast, as well as images with ambiguous or missing labels. Figure 5.1 shows examples of an adequate image suitable for analysis and an inappropriate image that would be discarded during this cleaning phase.

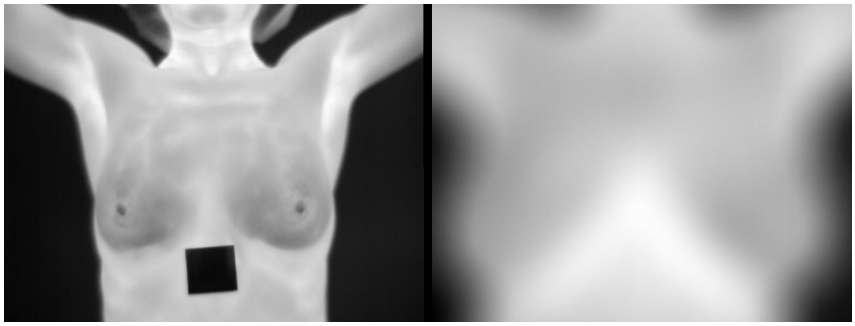

Figure 5.1: Example of adequate and an inappropriate breast thermography.

5.1.3 Data partitioning and leakage prevention

To obtain a realistic assessment of the model's ability to generalize to new, unseen patients, a strict patient-wise data separation strategy was implemented. This methodology is critical to prevent the optimistic performance bias caused by data leakage, a common flaw identified in related work (see Chapter 2).

The dataset was partitioned into training, validation, and test sets. This partitioning ensures that all images belonging to a single patient reside exclusively in one of these sets. This prevents the model from learning patient-specific anatomical features, forcing it instead to learn the generalizable thermal patterns indicative of pathology.

5.1.4 First dataset: original images

The dataset, which was cleaned and partitioned as previously described, was constructed to ensure class balance. For the healthy class, one frontal static image was selected per patient. To balance this, the sick class was composed of one frontal static image and three additional frontal dynamic images for each patient. The precise distribution of patients and images across the different sets is detailed in Table 5.1

Table 5.1: Distribution of images across the training and testing sets after cleaning and patient-wise partitioning.

Dataset Split	**Number of Images**	**Healthy Class (Images)**	**Cancer Class (Images)**
Training	252	129	123
Testing	82	42	40
Total	**334**	**171**	**163**

5.1.5 Secondary dataset: segmented thermograms

To further explore the model's capabilities and the impact of input data quality, a second experiment was conducted using a publicly available collection of pre-segmented thermograms from the DMR-IR (Database for Mastology Research with Infrared Image) dataset [1]. This dataset adds both static and dynamic images.

The segmentation pre-processing step allows the model to focus exclusively on the relevant physiological information within the breast tissue, potentially reducing noise and simplifying the classification task. The distribution for this dataset is detailed in Table 5.2.

Table 5.2: Distribution of patients and images across the training and test sets using the segmented dataset from DMR-IR [1].

Dataset Split	**Number of Images**	**Healthy Class (Images)**	**Cancer Class (Images)**
Training	1280	640	640
Testing	240	120	120
Total	**1520**	**760**	**760**

5.2 Model architecture and training strategy

For this work, a Convolutional Neural Network (CNN) based on the EfficientNet-B0 architecture was developed. The model's design and training were tailored to effectively learn from the thermographic images while maintaining computational efficiency.

5.2.1 Core architecture: EfficientNet-B0

The backbone of the model is an EfficientNet-B0, recognized for its excellent balance between high accuracy and computational efficiency. We utilize a model pre-trained on the ImageNet dataset, leveraging the power of transfer learning.

This pre-trained network serves as a powerful feature extractor. The initial layers are adept at identifying general, low-level features such as edges, textures, and color gradients from the input thermogram. These features are then progressively combined into more complex representations in the deeper layers, ultimately forming a rich feature map that encodes intricate local patterns and textures relevant to the diagnostic task.

5.2.2 Classification head

The standard EfficientNet-B0 classifier, originally designed for 1000 classes, was replaced to suit our binary classification task (malignant vs. benign). A new classification head was added, consisting of a Global Average Pooling layer to reduce the dimensionality of the feature map from the CNN backbone, followed by a Dense (fully connected) layer and a final Sigmoid activation function. This final stage maps the high-dimensional feature representation to a single scalar value between 0 and 1, representing the model's predicted probability of malignancy.

5.2.3 Multi-step fine-tuning strategy

To effectively adapt the pre-trained model to the specific domain of medical thermography, a multi-step fine-tuning strategy was employed. This gradual "unfreezing" of layers allows the network to learn domain-specific features without forgetting the valuable knowledge learned from ImageNet. The training was conducted in stages, starting with classifier training, where all layers of the EfficientNet-B0 backbone were initially frozen and only the newly added classification head was trained. This step allowed the classifier to interpret the general-purpose features extracted by the backbone for our specific task. Subsequently,

progressive unfreezing was applied, where blocks of layers within the backbone were made trainable—starting from those closest to the output and moving backward.

This methodical approach allows the network to first adapt its high-level, more abstract feature representations to the nuances of thermographic images. Subsequently, as earlier layers are unfrozen, the model can fine-tune the more fundamental feature detectors. This strategy ensures a stable learning process and helps the model achieve better generalization on the patient-separated test set. Furthermore, the training procedure is integrated with an experiment tracking server (MLflow), where each fine-tuning stage is logged as a specific run, ensuring that the transition from a frozen backbone to a fully optimized model is fully auditable and reproducible.

5.3 Training, validation, and evaluation methodology

A standardized and rigorous methodology was used for all experiments to ensure consistency and reliability.

5.3.1 Image preprocessing and data augmentation

Before being fed to the model, all images undergo a consistent sequence of preprocessing and augmentation steps to prepare them for training and to increase the model's robustness.

5.3.1.1 Preprocessing steps (for both training and test sets)

These steps are applied to both the training and test sets. First, images are resized while maintaining their original aspect ratio. Padding is then added to the borders to prevent the loss of information, which is particularly important for this dataset as the breast frequently touches the image's edge. After padding, the image is center-cropped to the fixed input size required by the model. This is used to prevent image distortions caused by resizing operations that might alter the aspect ratio. Finally, the image is converted into a PyTorch tensor and normalized using the mean and standard deviation statistics from the ImageNet dataset, as is standard practice for using pre-trained models.

5.3.1.2 Training augmentation steps (excluded from test set)

To prevent overfitting and help the model generalize better, the training includes the addition of several random augmentation steps. These include Random Horizontal Flip, where images are randomly flipped horizontally with a probability of 50%; Random Rotation, which applies a random rotation between -30 and +30 degrees; and Random Zoom, which applies a random zoom between -20% and +20%.

5.3.2 Training procedure

The model was trained using the PyTorch Lightning framework to ensure modular and boilerplate-free code. The key components of the training procedure include the Adam (Adaptive Moment Estimation) optimizer to update the model weights and a binary cross-entropy with logits loss, which was minimized. This loss function is well-suited for binary classification tasks as it combines a sigmoid layer and the binary cross-entropy loss in one single class, leveraging the log-sum-exp trick for greater numerical stability. Additionally, a ReduceLROnPlateau scheduler was employed to dynamically decrease the learning rate if the validation loss stagnated, helping the model to escape local minima and converge more effectively. Finally, two key callbacks were used: EarlyStopping to halt training if no improvement in validation performance was observed for a set number of epochs, preventing overfitting, and ModelCheckpoint to save the model weights corresponding to the best validation performance achieved during training.

5.3.3 Evaluation metrics

Model performance was assessed using a standard set of classification metrics, tracked using torchmetrics. In a medical context, understanding the trade-offs between these metrics is vital. These metrics include Accuracy, the proportion of total correct predictions, which can be misleading on imbalanced datasets. We also evaluated Precision ($TP/(TP+FP)$), which measures the accuracy of positive predictions. High precision is crucial to minimize false alarms and avoid unnecessary follow-up procedures for healthy patients. Recall (Sensitivity) ($TP/(TP+FN)$) measures the model's ability to identify all actual positive cases, which is crucial in a screening context to minimize the risk of missing a cancer diagnosis. Furthermore, Specificity ($TN/(TN+FP)$) measures the model's ability to correctly identify true negative cases. High specificity is complementary to high precision and indicates that the model is effective at correctly clearing healthy

patients. The F1-Score $(2 * (Precision * Recall)/(Precision + Recall))$, the harmonic mean of precision and recall, provides a single score that balances the two.

To further probe the reliability of the test-set estimate reported in Chapter 6, two complementary, post-hoc robustness analyses were conducted after the original evaluation.

First, a patient-level bootstrap resampling procedure was applied to the fixed test set: predictions on all test-set images were retained together with their patient identifiers, and the set of test patients was resampled with replacement (2,000 replicates), recomputing accuracy, precision, recall, F1-score, and specificity on each replicate to obtain 95% confidence intervals for the reported metrics.

Second, a nested, patient-wise, stratified group k-fold cross-validation (StratifiedGroupKFold, k = 5) was performed over the union of the training and test sets, such that every patient in the dataset served as part of an independent, held-out test fold exactly once, while model training and early-stopping selection for each outer fold used only the remaining patients (with a further inner split reserved exclusively for early-stopping monitoring, never for evaluation). This nested procedure removes the dependency of the reported performance on the composition of any single, fixed test partition.

A k-fold cross-validation restricted to the training set alone was deliberately not adopted as the primary robustness check: with only 28 sick patients available for training, partitioning this already small set further would reduce the effective training sample per fold even more. The nested cross-validation described above, which folds the test set itself into the cross-validation procedure, was chosen instead because it directly targets the risk of a biased or unrepresentative test partition.

5.4 Model Export and Inference Optimization

A critical step in the methodology is the transition from a research artifact to a production-ready service. After the training procedure is completed and the best-performing weights are identified via validation metrics, a dedicated *Export Pipeline* is triggered. This process is designed to decouple the model from the experimental environment, ensuring portability and scalability.

This pipeline performs **Device Mapping**, where the model's tensors are moved from the CUDA-enabled training environment to a CPU-compatible state. This transformation ensures that the inference service can be deployed on standard cloud infrastructure,

allowing for horizontal scaling according to demand without the financial or hardware constraints of specialized GPU instances.

To guarantee that real-time predictions remain consistent with the experimental results, the model is encapsulated within a custom `ModelWithPipeline` wrapper. Rather than serializing only the neural network weights, this wrapper integrates the entire preprocessing logic—including resizing, padding, and normalization statistics—directly into the model's forward pass. By loading the model through this unified pipeline, the system eliminates the risk of "training-serving skew", where discrepancies between preprocessing implementations could lead to degraded diagnostic accuracy. This architecture ensures that the inference service remains lightweight and independent, preventing resource bottlenecks between the high-performance training stage and the high-availability serving stage. Furthermore, the wrapper makes the model an autonomous unit. If a future version requires a different input resolution or normalization method, the model can be promoted via the MLflow Registry and consumed by the FastAPI service without any modifications to the API's source code. This enables seamless, zero-downtime updates in a clinical environment. This unified approach ensures that the resulting diagnostic tool is not only accurate but also robust, maintainable, and ready for high-availability deployment.

5.5 Usability Evaluation

To assess the practical viability and user experience of the deployed diagnostic tool (the Gradio web interface), a formal usability evaluation was conducted. The System Usability Scale (SUS) was employed as the primary assessment metric. The SUS is a widely adopted, standardized ten-item questionnaire evaluated on a 5-point Likert scale, ranging from "Strongly Disagree" (1) to "Strongly Agree" (5). It provides a global view of subjective assessments of usability, balancing both positive and negative statements regarding system complexity, ease of use, and need for technical support.

A group of 6 participants interacted with the deployed system and subsequently completed the SUS questionnaire. The final score was calculated by normalizing the responses and multiplying the sum by 2.5, yielding a final metric on a 0 to 100 scale.

Chapter 6

Results and discussions

With the system architecture and experimental methodology established, this chapter presents the empirical results of the research. The primary goal is to provide a quantitative and qualitative assessment of the trained EfficientNet-B0 model's performance on the task of breast cancer prediction from thermographic images, using a multi-step fine-tuning strategy.

The analysis begins with a presentation of the key performance metrics on the held-out, patient-separated test set, directly comparing them against the established benchmark. This is followed by an examination of the model's training dynamics to discuss convergence and generalization. To validate our architectural choices, the results of ablation studies are presented.

6.1 Experimental setup

For clarity, the results presented here were obtained under the following conditions:

- Model: An EfficientNet-B0 architecture, pre-trained on ImageNet and adapted using the multi-step fine-tuning strategy with progressive unfreezing, as described in Chapter 5.
- Dataset: This chapter utilizes two distinct datasets as defined in subsections 5.1.4 and 5.1.5.
- Evaluation: Performance was measured using the metrics of Accuracy, Precision, Recall and F1-Score, as defined in Chapter 5.

- Comparability: All metrics were logged in a centralized dashboard, allowing for a side-by-side analysis of how image segmentation impacts model performance.

6.2 Quantitative performance analysis

After the multi-stage training was complete, the best-performing model checkpoint was evaluated on the unseen, patient-separated test set. The model achieved a promising level of performance, demonstrating its potential for generalization.

6.2.1 Performance on full frontal thermograms

The first experiment compared the EfficientNet-B0 model against the baseline using the original frontal thermograms. As shown in Table 6.1, the EfficientNet-B0 model achieved an accuracy of 89.15% and a precision of 91.89%, compared to the baseline's 83% accuracy and 66% precision. The model's recall was 85%, a decrease from the baseline 97%.

Table 6.1: Final performance metrics on the patient-separated test set (full frontal images) compared to the benchmark study by Çevik et al [61].

Metric	**Our Model (Full Frontal Images)**	**Benchmark (Çevik et al., 2024)**
Accuracy	89.15%	83.00%
Precision	91.89%	66.00%
Recall	85.00%	97.00%
F1-Score	0.88	0.79
Specificity	93.02%	N/A

6.2.2 Impact of segmentation

To validate the hypothesis that image segmentation could improve predictive performance, the model was trained and evaluated on the pre-segmented dataset. Table 6.2 shows that using segmented images further increased performance, with accuracy reaching 92.29% and precision reaching 100%.

6.2.3 Analysis of training dynamics

The multi-step fine-tuning strategy proved effective in enabling stable convergence. During the initial training phase where only the classifier head was active, the validation

Table 6.2: Comparison of performance metrics across the baseline model and the EfficientNet-B0 model with both full frontal and segmented image datasets.

Metric	(Çevik et al., 2024)	EfficientNet-B0 (Full Frontal)	EfficientNet-B0 (Segmented)
Accuracy	83.00%	89.15%	**92.29%**
Precision	66.00%	91.89%	**100%**
Recall	**97.00%**	85.00%	85.83%
F1-Score	0.79	0.88	**0.92**
Specificity	N/A	93.02%	**100%**

loss decreased rapidly, indicating that the general-purpose features from the pre-trained backbone were already highly informative. As more layers were unfrozen, there were small, temporary increases in validation loss, which is expected as the model adjusts a larger number of parameters. The ReduceLROnPlateau scheduler was crucial during these stages, helping the model navigate these new optimization landscapes and settle into better minima. The EarlyStopping callback typically halted the training process after the final unfreezing stage, indicating that further end-to-end training provided diminishing returns and that the model had reached a stable point of convergence without significant overfitting. This controlled training process was essential for achieving the strong generalization performance seen on the test set.

6.2.4 Robustness of the test-set estimate: bootstrap confidence intervals and nested cross-validation

As described in Section 5.3.3, the performance reported in Table 6.1 was obtained from a single, fixed patient-wise split (28 sick and 129 healthy patients for training; 10 sick and 43 healthy patients for testing). Given the small number of patients in the test partition, two additional analyses were conducted to assess how sensitive this estimate is to the particular patients that happened to fall into the test set.

Table 6.3 reports patient-level bootstrap 95% confidence intervals (2,000 resamples), computed from an independent replication of the same training and testing protocol described in Chapter 5.

Table 6.3: Bootstrap 95% confidence intervals (patient-level resampling, $n = 2{,}000$) for the test-set metrics.

Metric	Point	Bootstrap mean	Bootstrap std	95% CI lower	95% CI upper
Accuracy	86.75%	86.86%	5.09%	75.90%	95.35%
Precision	91.43%	90.93%	6.52%	77.42%	100.00%
Recall	80.00%	79.90%	10.12%	58.33%	95.45%
F1-score	0.853	0.847	0.0726	0.6909	0.9474
Specificity	93.02%	93.11%	3.98%	84.08%	100.00%

Table 6.4 reports the results of the nested, patient-wise Stratified Group k-fold cross-validation (k = 5 outer folds), in which the training and test sets were merged and every patient served, exactly once, as an unseen test case.

Table 6.4: Nested cross-validation results ($k = 5$ outer folds, patient-wise, stratified by class).

Metric	Mean	Std. deviation
Accuracy	85.46%	5.89%
Precision	90.44%	4.24%
Recall	79.71%	8.30%
F1-score	0.85	0.05
Specificity	91.83%	3.69%

Table 6.5 places these results alongside the originally reported test-set performance (Table 6.1) for direct comparison.

Table 6.5: Comparison of original test-set performance, bootstrap confidence intervals, and nested cross-validation results.

Metric	Original test set (Table 6.1)	Bootstrap 95% CI	Nested CV mean ± std
Accuracy	89.15%	[75.90%, 95.35%]	85.46% ± 5.89%
Precision	91.89%	[77.42%, 100.00%]	90.44% ± 4.24%
Recall	85.00%	[58.33%, 95.45%]	79.71% ± 8.30%
F1-score	0.88	[0.6909, 0.9474]	0.85 ± 0.05
Specificity	93.02%	[84.08%, 100.00%]	91.83% ± 3.69%

The originally reported accuracy (89.15%) falls within the bootstrap confidence in-

terval and close to the nested cross-validation mean for every metric, indicating that it was not an outlier product of a particularly favourable test partition. The nested cross-validation estimate, which uses every patient in the dataset as a test case exactly once, is systematically somewhat lower than the fixed-split estimate. This is consistent with the expectation that a single held-out split, however carefully patient-separated, still carries higher variance than an evaluation that averages over the entire patient population. We therefore regard the nested cross-validation mean as the more statistically robust estimate of the model's expected generalization performance, while the fixed-split result in Table 6.1 remains a valid, patient-wise-separated evaluation that is consistent with it.

6.3 Comparison with recent works

To contextualize the performance of the proposed EfficientNet-B0 model, we compared our results against recent studies in the literature that utilized the same source dataset (DMR-IR). Table 6.6, adapted from the extensive review provided by Çevik et al. [61], summarizes these works, focusing on the relationship between the data splitting strategy and reported accuracy.

A critical pattern emerges from this comparison. Studies that employ a "Random" data splitting technique—which typically leads to data leakage by mixing images from the same patient across training and test sets—consistently report near-perfect accuracies ranging from 92% to 100%. In contrast, when a rigorous "Patient-by-patient" split is enforced to ensure valid generalization, performance metrics tend to drop to more realistic levels.

Our work aligns methodologically with Çevik et al. [61], prioritizing valid generalization. However, as shown in the table, our segmented EfficientNet-B0 model achieves an accuracy of 92.29% for the segmented set and 89.15% for the full frontal set, outperforming the 83.00% reported by Çevik et al. under similar rigorous constraints.

6.4 Ablation studies

In order to systematically validate our methodology's key components, we conducted a series of ablation studies. These studies involved selectively removing or altering various parts of our pipeline, which ranged from the model architecture to the preprocessing steps, allowing us to isolate and quantify their individual contributions to the final performance.

Table 6.6: Performance comparison with recent studies using the DMR-IR dataset.

Method	Study	Approach	Model	Data Split Technique	Accuracy (%)
Machine Learning	Pramanik et al.[40]	Detection & Classification	MLP	Randomly	90.48
	Karim et al.[26]	Classification	SVM	NA	91.25
	Ghayoumi Zadeh et al.[24]	Classification	Autoencoder	NA	94.87
Deep Learning	Baffa and Lattari[19]	Classification	CNN	Randomly	98.0 (static) 95.0 (dynamic)
	Fernández-Ovies et al.[22]	Classification	CNN	Randomly	100.00
	Tello-Mijares et al.[54]	Detection & Classification	CNN	NA	100.00
	Civilibal et al.[62]	Segmentation & Classification	Mask R-CNN and transfer learning	Randomly	100.00
	Zuluaga-Gomez et al.[60]	Classification	CNN	Randomly	92.00
	Civilibal et al.[13]	Detection & Classification	Mask R-CNN	Randomly	97.10
	Çevik et al.[61]	Detection & Classification	YOLOv5	On a patient-by-patient basis	83.00
	This Work (Baseline)	**Classification**	**EfficientNet-B0**	**On a patient-by-patient basis**	**89.15**
	This Work (Segmented)	**Classification**	**EfficientNet-B0**	**On a patient-by-patient basis**	**92.29**

6.4.1 Preprocessing ablation: the role of padding

In convolutional neural networks, features located at the very edge of an image can be partially lost or underrepresented during the initial convolution operations. Our preprocessing pipeline (detailed in Section 5.3.1) addresses this by adding padding after resizing the image while maintaining its aspect ratio. This process creates a border around the original image, effectively moving crucial edge features towards the center so their information can be fully captured by the network's filters.

To validate this choice, we compared it against an alternative pipeline where the padding step was simply excluded. In this approach, the image undergoes the same resizing and center-cropping, but without padding, which can lead to the loss of information at the image's borders. We trained a "No Padding" model using this alternative step to quantify the impact of our padding strategy.

6.4.1.1 Full frontal image dataset - padding ablation Analysis

The results on the full frontal image dataset, shown in Table 6.7, confirm that our padding strategy yields superior performance. The model trained without padding, which loses information at the image borders during the center-cropping step, shows a significant drop across all key metrics. Notably, Accuracy fell by over 10 percentage points, and Recall decreased from 85.00% to 72.50%. This suggests that the loss of peripheral spatial information degrades critical features that the model relies on for accurate classification.

Table 6.7: Performance comparison of padding pipeline on the full frontal dataset.

Metric	Full frontal model with Padding	Full frontal model without Padding
Accuracy	89.15%	78.31%
Precision	91.89%	80.56%
Recall	85.00%	72.50%
F1-Score	0.88	0.76
Specificity	93.02%	83.72%

6.4.1.2 Segmented image dataset - padding ablation analysis

The negative impact of image distortion is even more pronounced on the cleaner, segmented dataset (Table 6.8). While Precision remained a perfect 100%, the Recall plummeted from 85.83% to just 61.67%. This dramatic decrease means the no-padding model missed over a third of the positive cases. The noticeable improvement in all metrics when using padding validates our hypothesis that preserving edge features and image geometry is crucial for the model's ability to interpret the complete thermal map.

Table 6.8: Performance comparison of padding pipeline on the segmented dataset.

Metric	Segmented model with Padding	Segmented model without Padding
Accuracy	92.29%	80.83%
Precision	100%	100%
Recall	85.83%	61.67%
F1-Score	0.92	0.76
Specificity	100%	100%

The results indicate that our padding strategy yields superior performance. The lower metrics of the "no-padding" model suggest that distorting the image's aspect ratio degrades critical spatial features. More importantly, it validates our hypothesis that preserving edge features through padding is crucial for the model's ability to interpret the complete thermal map. The noticeable improvement in recall and overall accuracy confirms that including padding to prevent the loss of information at the image borders is a vital preprocessing step.

6.4.2 Data augmentation ablation

To combat potential overfitting and improve model generalization, a set of data augmentation techniques (horizontal flips, rotations and zooms as described in Chapter 5.3.1.2)

were applied to the training set. To quantify their benefit, an ablation study was conducted by training a version of the model on the original, unaltered training data. The comparison was performed for both the full frontal image dataset and the pre-segmented dataset.

6.4.2.1 Full frontal image dataset - augmentation ablation analysis

The results for the model trained on full frontal thermograms are presented in Table 6.9. The comparison clearly demonstrates the profound impact of data augmentation. While precision remains high in both models, the model trained with an augmented dataset achieved a dramatic improvement in Recall, jumping from 50.00% to 85.00%. This indicates that without augmentation, the model fails to identify half of the positive cases. The substantial increases in overall Accuracy and F1-Score further confirm the value of augmentation as a critical component of the training procedure.

Table 6.9: Performance comparison between the model trained with and without data augmentation on the full frontal image dataset.

Metric	**Full frontal model with Augmentation**	**Full frontal model without Augmentation**
Accuracy	89.15%	73.49%
Precision	91.89%	90.91%
Recall	85.00%	50.00%
F1-Score	0.88	0.64
Specificity	95.35%	93.02%

6.4.2.2 Segmented image dataset - augmentation ablation analysis

The same positive trend was observed when evaluating the models on the pre-segmented dataset, as shown in Table 6.10. While the baseline performance without augmentation is higher on this cleaner dataset, the addition of augmentation still provides a significant boost. Precision remained perfect at 100% in both cases, but Recall saw a substantial increase from 66.67% to 85.83%. This reinforces the conclusion that by exposing the model to a wider variety of synthetic examples, data augmentation helps it learn more robust and generalizable features, making it significantly better at identifying true positive cases.

Table 6.10: Performance comparison between the model trained with and without data augmentation on the segmented image dataset.

Metric	Segment model with Augmentation	Segmented model without Augmentation
Accuracy	92.29%	83.33%
Precision	100%	100%
Recall	85.83%	66.67%
F1-Score	0.92	0.80
Specificity	100%	100%

6.5 Model explainability (XAI) and visual analysis

While quantitative metrics like accuracy and precision are essential for evaluating model performance, they do not explain why a model makes a specific prediction. In a high-stakes medical context, this lack of transparency can be a major barrier to clinical adoption. To address this, we employed Explainable AI (XAI) techniques to gain insight into the model decision-making process.

Specifically, we used Gradient-weighted Class Activation Mapping (Grad-CAM), a visualization technique that produces a heatmap highlighting the regions of an input image that were most influential in the model's classification decision. The heatmap is generated by analyzing the gradient flow into the final convolutional layer of the CNN feature extractor, effectively showing where the model "looks" when making a prediction.

This analysis serves two key purposes:

- Building Trust and Verifying Logic: By visualizing the model focus, we can verify if it is concentrating on physiologically relevant areas—such as regions of thermal asymmetry or vascular anomalies, rather than irrelevant artifacts. This helps build confidence that the model has learned meaningful patterns and in the case of a non-healthy prediction it helps the doctor to know where to investigate to confirm the diagnosis.
- Debugging and Error Analysis: For incorrect predictions, Grad-CAM can help diagnose the failure. It can reveal if the model was distracted by noise, focused on an ambiguous region, or failed to identify a subtle thermal signature.

The image below shows an example of a Grad-CAM heatmap applied to a positive prediction from the full frontal dataset, illustrating the areas the model identified as most

indicative of a malignant condition.

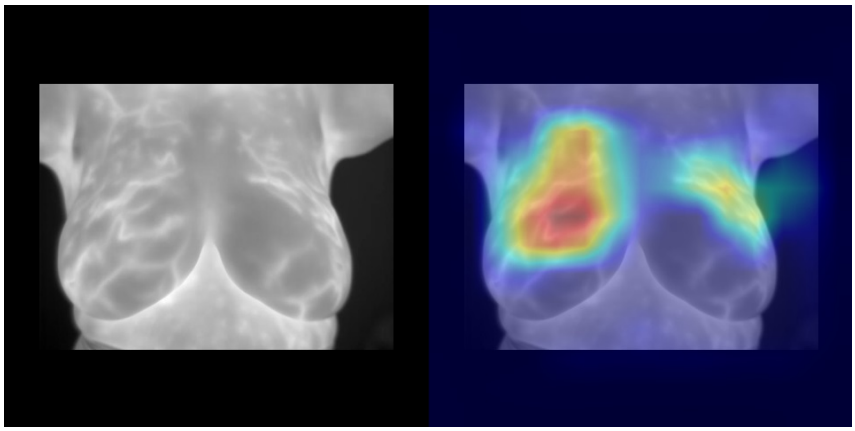

Figure 6.1: Grad-CAM visualization on a full frontal thermogram. The heatmap highlights the specific regions of thermal asymmetry that most influenced the model's positive prediction.

6.5.1 Qualitative relationship between predictions and image characteristics

Beyond the single example shown in Figure 6.1, a qualitative inspection of the Grad-CAM outputs and raw images across several patients of the sick class provides additional, non-clinical insight into how the reported metrics relate to properties of the input data. This analysis does not attempt to establish a radiological diagnosis, which would require expert clinical judgement, but instead examines image-level characteristics that are observable without medical training: activation location relative to breast tissue, image contrast and framing, and the consistency of predictions across images of the same patient.

Correctly classified cases with high prediction confidence (typically above 97%) consistently show Grad-CAM activation concentrated within the breast contours, coinciding with regions of visible thermal asymmetry or a pronounced vascular pattern already apparent in the raw grayscale image. Figure 6.2 illustrates this for one such patient: the heatmap's highest-activation region overlaps with the breast that displays the more irregular thermal pattern in the original image, rather than with background, arms, or the dark calibration marker occasionally visible near the sternum. This indicates that, at least for these high-confidence predictions, the model is attending to anatomically plausible regions rather than incidental artifacts.

Not every correct prediction, however, is associated with a concentrated, high-confidence activation. Figure 6.3 shows a patient correctly classified with a much lower confidence

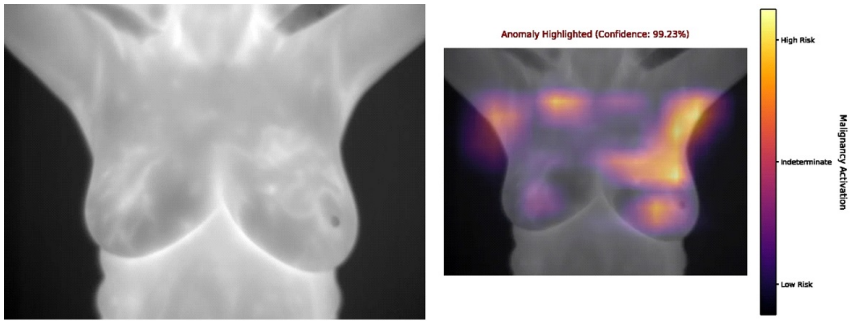

Figure 6.2: Raw thermogram (left) and corresponding Grad-CAM heatmap (right) for a correctly classified sick patient (prediction confidence 99.23%). The highest-activation region coincides with the breast presenting the more irregular thermal pattern in the raw image.

(55.41%), close to the decision boundary. In this case, the raw image itself presents visibly lower thermal contrast, the breast contours are less sharply defined and no clear vascular pattern is discernible, and the corresponding Grad-CAM activation is more diffuse, scattered mainly over the axillary and lateral chest-wall regions rather than concentrated within the breast tissue. This suggests a qualitative relationship between the sharpness of the thermal signal available in the raw image and both the model's confidence and the anatomical plausibility of its attention pattern.

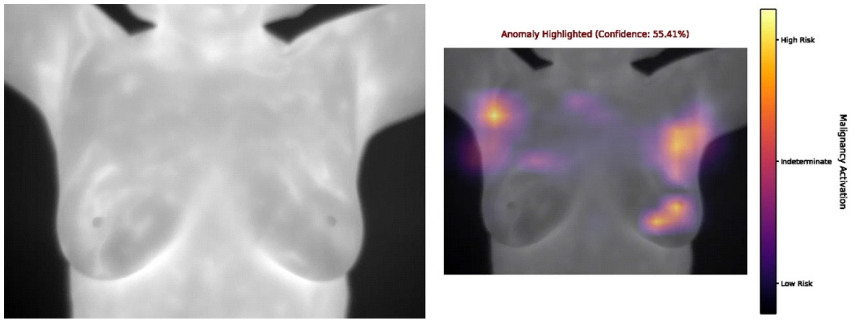

Figure 6.3: Raw thermogram (left) and Grad-CAM heatmap (right) for a correctly classified sick patient near the decision boundary (prediction confidence 55.41%). Activation is diffuse and concentrated over the axillary region rather than the breast tissue itself.

The dataset used in this experiment includes, for several sick patients, multiple images captured within the same acquisition session as part of the dynamic thermography protocol (Section 2.1.2). This allows a further qualitative observation: for patients with at least one misclassified image, the misclassified frame is, to visual inspection, nearly indistinguishable from other frames of the same patient that were classified correctly, same pose, similar framing, and no obvious difference in image quality or artifacts. Figure 6.4 illustrates this for one patient: the two raw images differ mainly in the presence of the calibration marker and minor pose variation, yet one was classified correctly and the other

was not. This indicates that, at least in these cases, misclassification is better explained by subtle, frame-to-frame variation in the thermal signal captured within a single session than by a single, visually obvious defect in the misclassified image.

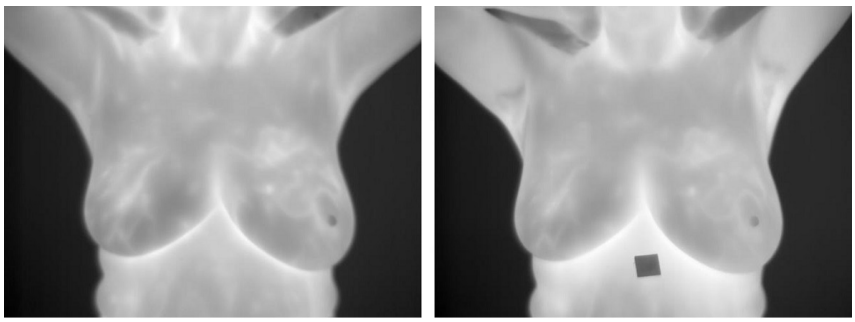

Figure 6.4: Two raw thermograms of the same sick patient, captured within the same acquisition session. Left: correctly classified. Right: misclassified as healthy. The two images are visually near-identical to non-expert inspection.

A possible alternative explanation for this result deserves particular attention: since the misclassified frame is the one of the two images that happens to display the dark calibration marker, the error could, in principle, reflect a systematic bias toward this incidental artifact rather than genuine, subtle variation in the underlying thermal signal. To directly investigate this hypothesis, Grad-CAM was additionally computed for the misclassified frame itself and compared against the heatmap of the correctly classified frame (Figure 6.5). In both frames, the region occupied by the calibration marker consistently falls within the lowest-activation band of the color scale, indicating that the marker did not attract the model's attention in either case. Moreover, the overall spatial pattern of activation is qualitatively very similar between the two frames: both show the highest activation concentrated in the same three regions: the left axilla, a vertical streak along the upper-right chest/arm boundary, and the lower-right breast. The misclassified frame differs from the correctly classified one mainly in the intensity, and consequently the overall confidence, of this pattern (47.72%, just below the 0.5 classification threshold) rather than in its spatial location (compared to 99.23% for the correctly classified frame). This evidence is inconsistent with the model relying on the calibration marker as a shortcut feature, and instead supports the interpretation that the observed misclassification stems from a threshold-level sensitivity to subtle frame-to-frame variation in the same anatomical regions, rather than from a systematic bias toward an incidental image artifact.

Finally, the raw images also reveal considerable heterogeneity in acquisition quality across patients, independent of classification outcome. Figure 6.6 contrasts a patient with

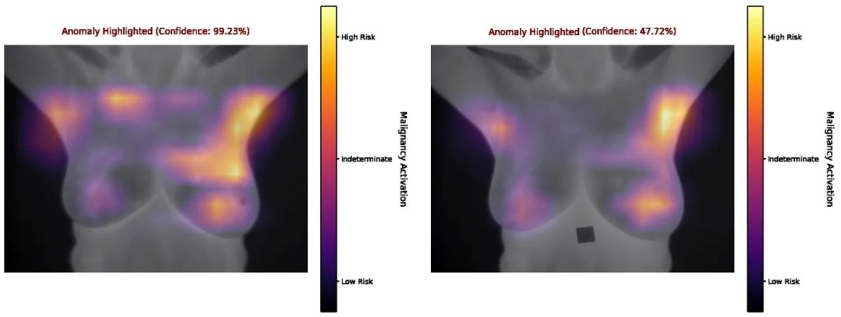

Figure 6.5: Grad-CAM heatmaps for the two frames of the same sick patient shown in Figure 6.4. Left: correctly classified frame (confidence 99.23%). Right: misclassified frame (confidence 47.72%). The calibration marker, visible in the right-hand raw image in Figure 6.4, falls within the lowest-activation band in both frames, and the overall spatial pattern of activation is similar between them.

a sharply defined, high-contrast thermal pattern against a patient whose image presents visibly lower contrast and less defined breast contours. This heterogeneity is consistent with the study's already acknowledged limitation regarding the absence of disease-staging information (Section 7.6): thermal and morphological signatures may be more explicit, and consequently easier for the model to detect, in cases with more pronounced physiological changes, while more subtle presentations may be comparatively harder to capture reliably regardless of model architecture.

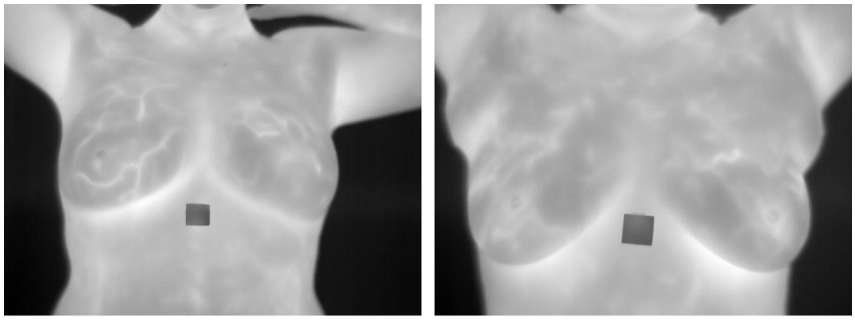

Figure 6.6: Raw thermograms of two different sick patients, illustrating the heterogeneity in image contrast and definition present in the dataset. Left: a patient with a clearly visible, asymmetric vascular pattern. Right: a patient with a comparatively low-contrast, less defined thermal signature.

Taken together, these qualitative observations do not replace the quantitative evaluation presented in Sections 6.2 and 6.2.4, but they provide converging, non-clinical evidence that (i) the model's attention, as captured by Grad-CAM, tends to concentrate on anatomically plausible regions in high-confidence correct predictions; (ii) prediction confidence tends to be lower, and model attention more diffuse, in cases with lower raw image contrast; and (iii) at least part of the observed misclassification rate is attributable

to frame-level variability within a single patient's acquisition session and to heterogeneity in raw image quality across patients, rather than to a single, identifiable failure mode of the model.

6.6 Usability Analysis

Beyond the predictive performance of the EfficientNet-B0 model, the practical success of the proposed MLOps pipeline relies heavily on user acceptance. To evaluate this, the results from the System Usability Scale (SUS) questionnaire were analyzed. The individual responses from the six participants are detailed in Table 6.11.

Table 6.11: System Usability Scale (SUS) questionnaire and individual user responses

Question	**U1**	**U2**	**U3**	**U4**	**U5**	**U6**
1. I think that I would like to use this system frequently.	3	5	5	4	4	5
2. I found the system unnecessarily complex.	4	1	1	-	1	1
3. I thought the system was easy to use.	3	5	5	5	4	5
4. I think that I would need the support of a technical person to be able to use this system.	3	1	1	4	1	1
5. I found the various functions in this system were well integrated.	3	5	5	4	5	5
6. I thought there was too much inconsistency in this system.	3	1	1	2	2	1
7. I would imagine that most people would learn to use this system very quickly.	3	5	5	5	5	5
8. I found the system very cumbersome to use.	3	1	1	1	1	1
9. I felt very confident using the system.	3	4	5	4	4	5
10. I needed to learn a lot of things before I could get going with this system.	3	1	1	1	1	1

The evaluated platform achieved a final average SUS score of 85.83. This indicates an excellent level of usability. The overall grade reported high confidence in using the system and disagreed with statements suggesting the tool was unnecessarily complex or required technical support. This confirms that the decoupled microservices architecture, combined with the Gradio interface, successfully bridges the gap between complex deep learning backend operations and an intuitive clinical user experience.

Chapter 7

Discussion

This chapter provides a comprehensive interpretation of the results presented in Chapter 6, placing them within the broader context of the research questions and the current state of the field. The aim is to move beyond the quantitative findings to explore their deeper meaning and implications.

The discussion will begin by interpreting the key performance findings, analyzing the trade-offs revealed by the evaluation metrics and what they signify for the model's potential real-world application. It will then revisit the central theme of methodological rigor, emphasizing the impact of our approach on establishing a realistic performance benchmark. Following this, the broader implications of the MLOps pipeline as a solution to the research-to-deployment gap will be explored. The chapter will conclude with a transparent assessment of the study's limitations, which in turn informs the directions for future work.

7.1 Interpretation of key findings

The results from the previous chapter paint a picture of a model with a distinct performance profile, offering both significant strengths and clear limitations.

7.1.1 The precision-recall trade-off and clinical utility - first model versus baseline

A key finding is the model's performance on the full frontal images, achieving 89.15% accuracy and a balanced 91.89% precision with 85% recall. This presents a compelling trade-off when compared to the benchmark study, which favored a very high recall (97%)

at the cost of much lower precision (66%).

- High Precision is Critical: The first model precision of nearly 92% means that when it flags a thermogram as potentially malignant (a positive prediction), it is correct more than 9 out of 10 times. This is vital in a clinical context as it minimizes the rate of false positives. Fewer false positives translate to less unnecessary patient anxiety, a reduction in costly and invasive follow-up procedures (like biopsies), and greater trust in the system as a reliable diagnostic aid.
- Strong Recall for Screening: A recall of 85% indicates the model correctly identifies 85 out of 100 malignant cases. While this means some cases are missed (false negatives), it represents a strong ability to detect the presence of disease. In contrast, the benchmark's higher recall comes with the significant drawback that one-third of its positive alerts are incorrect

Ultimately, our model's superior F1-Score (0.88 vs. 0.79) demonstrates a more balanced and clinically useful profile. It strikes an effective balance between reliably identifying sick patients and avoiding alarming healthy ones, positioning it as a potent tool to assist clinicians.

7.1.2 The impact of segmentation - second model

The secondary experiment conducted on the pre-segmented dataset from DMR-IR [1] yielded significantly higher performance metrics across the board, reaching 92.29% accuracy and a perfect 100% precision. This finding strongly suggests that the presence of background information (e.g., chest wall, arms, sternum) in the non-segmented images presents a major challenge for the model, acting as a source of noise.

By working with images where the breast region is already isolated, the model can dedicate its full capacity to analyzing relevant physiological features. This underscores the critical importance of an effective segmentation step in the overall pipeline. The performance leap observed in this experiment validates the hypothesis that accurate localization of the region of interest is as crucial as the classification architecture itself. Future work should prioritize the integration of an automated segmentation module, as it is the most direct path to boosting the model's overall performance and clinical utility.

7.2 The critical impact of methodological rigor

A core contribution of this thesis is its unwavering commitment to methodological rigor, specifically through the enforcement of strict patient-wise data separation.

This work serves as a crucial data point in the literature, helping to establish a more sober and realistic performance benchmark for deep learning models on this task. It challenges the validity of overly optimistic reports and argues that patient-wise separation should be a non-negotiable standard for any future research in this domain. By providing both the methodology and a more plausible set of results, this thesis contributes to building a more reliable and trustworthy foundation for the field.

A further methodological refinement, addressed a residual limitation of the original evaluation protocol: with only 10 sick patients in the fixed test partition, a single point estimate carries a non-trivial risk of being favourably or unfavourably biased by the specific patients drawn into that partition. Section 6.2.4 reports two complementary analyses conducted to quantify and mitigate this risk: patient-level bootstrap resampling of the fixed test set, and a nested, patient-wise cross-validation in which every patient in the dataset serves as an unseen test case exactly once. Both analyses corroborate the original result, with the nested cross-validation mean falling close to the originally reported accuracy, and thus provide additional, more statistically grounded evidence that the model's reported performance reflects genuine generalization rather than a favourable test split.

7.3 Implications of the MLOps pipeline

Beyond the predictive model itself, the engineering of the end-to-end MLOps pipeline represents a significant contribution. This work sought to "bridge the gap between research prototypes and practical deployment," and the resulting system serves as a concrete solution.

7.3.1 A blueprint for reproducible research

The pipeline provides a practical, open-source blueprint for other researchers in medical imaging. By integrating tools for orchestration (Airflow), experiment tracking (MLflow), and containerization (Docker), it addresses the common challenges of reproducibility and scalability that often plague academic research. This framework allows for faster, more reliable iteration and easier collaboration.

7.3.2 Democratizing deployment

The choice of modern, open-source tools was intentional, aiming to "democratize ML deployment, even in academic settings with limited resources". This thesis demonstrates that building a robust, end-to-end system does not require proprietary software or massive enterprise-level infrastructure. This lowers the barrier to entry for creating high-quality, deployable machine learning solutions, empowering researchers to move their work closer to real-world impact.

7.4 Implications of the microservices-based MLOps pipeline

Beyond the predictive model, the engineering of the end-to-end MLOps pipeline represents a significant contribution to the field of medical systems. This work successfully addressed the "research-to-deployment gap" by transitioning from a monolithic script to a decoupled, service-oriented architecture.

7.4.1 Resource efficiency and horizontal scalability

A major implication of our architecture is the optimization of high-cost hardware. By separating the training engine from the inference API (CPU-bound), the system demonstrates that a robust diagnostic tool can be served on lightweight infrastructure. This horizontal scalability ensures that as user adoption increases, additional inference nodes can be deployed without the financial or technical burden of specialized GPU instances, a critical factor for the democratization of AI in resource-constrained medical environments.

7.4.2 Governance as a clinical prerequisite

The synergy between Apache Airflow and MLflow establishes a level of technical auditability that is often absent in academic prototypes. Every diagnostic model in our registry carries a complete lineage, linking current performance to original raw data and specific code versions. In a clinical context, this traceability is not just a convenience but a prerequisite for regulatory compliance and professional trust.

7.5 Engineering implications: beyond predictive accuracy

The results achieved by the EfficientNet-B0 model are inseparable from the infrastructure that supports it. A significant part of this work success lies in the transition from a monolithic experimental setup to a service-oriented architecture.

7.5.1 Operational robustness and the unified pipeline

The implementation of the ModelWithPipeline wrapper represents a strategic engineering choice to decouple high-level preprocessing logic from the underlying infrastructure. By encapsulating image normalization, padding, and resizing within the model artifact itself, we effectively mitigated the risk of training-serving skew, a common failure point where discrepancies in data preparation lead to degraded clinical accuracy.

Practically, this architectural pattern facilitates seamless model updates. Since the preprocessing requirements are "baked" into the model, upgrades to newer architectures or different input resolutions can be deployed via the MLflow Registry without modifying the inference API's source code. This ensures high system availability and horizontal scalability, allowing the diagnostic tool to be served across multiple CPU-based nodes while maintaining strict data consistency and operational integrity.

This approach addresses a critical "research-to-deployment" gap. In many medical imaging studies, the model's performance is often tied to specific, manually configured environments. Our methodology, however, provides a plug-and-play capability: the model becomes an autonomous unit of knowledge that carries its own instructions for data handling, ensuring that the performance benchmarks reported in Chapter 6 are preserved regardless of the deployment environment.

7.6 Limitations of the study

A critical and honest assessment of the study's limitations is essential for contextualizing the findings and guiding future work.

- Dataset Limitations: The primary limitation is the size and diversity of the dataset. While carefully curated, the data was sourced from only two hospitals within a single geographic region (Rio de Janeiro, Brazil). This may limit the model's ability

to generalize to different populations, ethnicities, and body types. Furthermore, the performance may vary with images captured using different models of infrared cameras or under different clinical protocols.

- Lack of Clinical Validation: The system has not been evaluated in a real clinical workflow. While a preliminary SUS evaluation demonstrated excellent system usability, the direct impact on a clinician's diagnostic accuracy and overall workflow efficiency in a live clinical setting remains to be fully quantified. The findings of this thesis represent a technical validation, which is a necessary precursor to, but not a substitute for, formal clinical trials.

- Lack of Disease Staging Information: The dataset does not differentiate between the stages of cancer progression. In patients with more advanced disease, the thermal and morphological traits being predicted (such as significant hotspots or physical deformations of the breast area) can be more explicit and thus easier for a model to detect. The model high performance could be influenced by these more obvious cases. Consequently, its true effectiveness in detecting the subtle thermal patterns of early-stage cancer, which is the primary goal of a screening tool, may be lower than expected.

- Operational Complexity: While the microservices approach offers scalability, it introduces a higher baseline for operational complexity. Maintaining a distributed system with Airflow, MLflow, and Docker requires specialized DevOps knowledge that might not be available in all clinical research settings.

Chapter 8

Conclusion and future work

This final chapter concludes the thesis by summarizing the research undertaken, its primary findings, and its contributions to the field of medical image analysis. It serves to consolidate the work, restate its significance, and outline a clear and promising roadmap for future investigation. By reflecting on what has been accomplished and what lies ahead, this chapter frames the thesis as a foundational step toward the development of reliable, accessible, and impactful AI-driven diagnostic tools.

8.1 Summary of work

This thesis addressed the pressing need for safer and more accessible breast cancer screening methods by exploring the potential of machine learning applied to medical thermography. The research moved beyond the development of a standalone predictive model to design, implement, and rigorously evaluate a complete, end-to-end MLOps pipeline. By bridging the gap between academic research and functional deployment, this work establishes a new methodological benchmark for the field, prioritizing both diagnostic accuracy and engineering scalability. Another important point of this work was a commitment to methodological rigor, specifically by implementing a strict patient-wise data separation strategy to prevent the data leakage that compromises the validity of many existing studies.

An EfficientNet-B0 model, adapted using a multi-step fine-tuning strategy, was trained and evaluated within this robust framework. It yielded a realistic performance profile characterized by high precision and strong recall. The resulting system, complete with a user-friendly interface, demonstrates a viable path from a research concept to a deployable application, providing a solid foundation for future development and clinical integration.

8.2 Restatement of key contributions

The principal contributions of this research can be summarized as follows:

- Decoupled MLOps Architecture: This thesis successfully designed and implemented a complete, open-source MLOps pipeline for medical image analysis using modern tools like Airflow, MLflow, and Docker. This pipeline provides a blueprint for creating scalable, reproducible, and deployable ML systems in an academic setting. This modular design ensures that each component can scale independently, providing a blueprint for reproducible and resilient ML systems in medical informatics. Furthermore, the implementation of the code also makes every component easily customizable making it easy to re-use it for different applications, adding different base models, training configurations or new endpoints to the API or UI.

- An EfficientNet-B0 Classification Model: An EfficientNet-B0 architecture was adapted using a multi-step fine-tuning strategy and rigorously evaluated, demonstrating its effectiveness for thermogram classification. The final model achieved promising performance on the patient-separated test set, with an accuracy of 89.15% and a precision of 91.89%, which improved further to 92.29% accuracy and 100% precision when evaluated on the pre-segmented dataset.

- Elimination of Training-Serving Skew: Through the development of a custom ModelWithPipeline wrapper, this work ensured that all preprocessing logic is encapsulated within the model artifact. This "plug-and-play" capability guarantees that performance benchmarks are preserved during transition to production, preventing the common discrepancies between experimental and real-world data preparation.

- Dynamic Model Governance: The integration with MLflow as a centralized model registry allows the system to transition between experimental versions and production-ready models without manual intervention. The practical implementation supports a "live-swap" mechanism, where the API dynamically fetches the latest "Production-tagged" model, ensuring that clinical users always benefit from the most accurate validated version.

- Methodological Rigor: A central contribution was the critical analysis and mitigation of the data leakage problem. By enforcing a strict patient-wise separation, this work helps to establish a more realistic performance benchmark for the field and champions a higher standard of validation.

- An Accessible Interface: The development of an integrated user interface using Gradio demonstrated a clear path from a trained model to an interactive, deployable application, thereby bridging a common gap in ML research.

In conclusion, this thesis provides a methodologically sound and technologically advanced framework for the deployment of AI in medical imaging. It demonstrates that the synergy between deep learning, rigorous data handling, and automated orchestration can transition research prototypes into reliable, practical, and life-saving diagnostic systems.

8.3 Future works

It is hoped that this dissertation provides a useful baseline for subsequent studies. Building upon the findings presented here, the research could be extended in several promising directions. The following areas represent suggested avenues for future research:

8.3.1 Multi-centric data acquisition

To improve model generalization and reduce potential bias, future work should focus on acquiring more diverse thermography data from multiple institutions, encompassing varied patient demographics and imaging equipment. A concrete and immediate next step is the integration of publicly available multi-centric datasets into the current pipeline. For instance, the dataset acquired at the San Juan de Dios Hospital & Consultorio Rosado in Cali, Colombia (publicly available via the Mendeley Data Repository), provides thermograms of 119 women captured under controlled protocols using a FLIR A300 camera[46]. It includes essential clinical metadata such as age, BMI, and pathology reports[46]. Combining this external dataset with the current DMR-IR database would significantly enhance the robustness and reliability of the trained models for diverse populations.

8.3.2 Advanced data augmentation

Research into data augmentation techniques, such as using Generative Adversarial Networks (GANs) to synthesize realistic thermograms, could help expand the dataset. Furthermore, future augmentation strategies should build upon the group's prior innovations, such as using the 20 images of the dynamic protocol acquired after cooling the patient. These images are captured for 5 minutes at 15-second intervals following a slight cooling of the breast surface via patient ventilation, and have already been successfully utilized in

works such as Resmini et al.[44] and Silva et al.[17]. Leveraging these temporal dynamics could yield highly realistic augmented samples. However, this must be done with extreme care to avoid reintroducing data leakage across patient splits.

8.3.3 3D Visualization of Thermographic Examinations

An improved understanding and interpretation of what the thermographic examination reveals can be achieved through the 3D representation of 2D thermographic breast acquisitions, as proposed by Moura et al. [35]. This three-dimensional representation can be highly valuable in breast reconstructions following total (radical mastectomies) or partial removal surgeries. Furthermore, it can be extremely useful in the planning of aesthetic plastic surgeries, since this examination, regarding the external shape of the breast, is equivalent to the photographs already employed daily by plastic surgeons in their clinical practice[14, 39, 36].

8.3.4 Segmentation and localization

Regarding segmentation and localization, it is essential to distinguish between two different aspects. The first is the segmentation of the breasts from the full frame acquired by the camera, which typically serves as a preliminary step in the pipeline. The UFF research group has extensive experience in this area, having produced three master's dissertations dedicated exclusively to this task (two for frontal views and one for lateral views). These works included ground truth verifications by mastologists, which are available in the dataset, alongside several published papers. The experiment with segmented data, which yielded substantially higher results, confirms that implementing a robust, automated segmentation module should be the highest priority for future work. The UFF research group has already demonstrated significant success in this area, notably through the Region of Interest (ROI) extraction in thermographic images using Genetic Algorithms proposed by Mendes et al. [33]. Future efforts should focus on upgrading these classical methods by integrating deep learning segmentation models (e.g., a U-Net architecture) as a preliminary step in the pipeline.

Another aspect is indicating where the problem is located within the patient. This step is not performed at the beginning, but rather in conjunction with the diagnostic phase. It encompasses several clinical details: identifying in which breast the problem is located (right or left), determining which part of the breast is affected (typically localized by

quadrant or retroareolar region), and pinpointing exactly where the problem is, including its size and depth (which is crucial for planning subsequent treatments). While the current system successfully uses Grad-CAM to generate heatmaps highlighting the regions that influenced the diagnosis, a key limitation of the current model is its inability to precisely localize and outline these suspicious regions within the full thermogram to extract those exact clinical metrics. Addressing this in future works, evolving from attention heatmaps to exact anatomical localization, would increase the clinical utility of the tool for clinicians on monitoring breast cancer treatments.

8.3.5 Exploring new architectures

The field of deep learning is constantly evolving , and future research should explore different model architectures to further improve performance. The UFF group has extensively proven the value of hybrid methodologies and temporal analysis for this task. Works such as Resmini et al.[43], which combined Genetic Algorithms and SVMs, and the continuous research by Araújo et al.[5, 6, 7] and Silva et al.[17, 50] using dynamic thermographic time series.

Expanding on this temporal analysis, a highly promising direction is the application of Video Vision Transformers (ViViTs) for sequential thermal data. Recent research by Falconi et al.[21] demonstrated that ViViTs, specifically employing the TimeSformer architecture, can effectively process the sequential nature of the UFF's Dynamic Infrared Thermography protocol. By treating dynamic thermography as a video sequence, their model achieved an Area Under the Curve score of 0.94, successfully outperforming previous baselines set by 3D Convolutional Neural Networks. Future models should continue to adapt these modern spatio-temporal attention architectures to capitalize on the dynamic data acquisition protocols already validated by the research group, exploring longer sequence lengths and hybrid combinations of ViViTs with the existing MLOps pipeline.

8.3.6 System and pipeline refinement

8.3.6.1 Fully automated retraining

The Airflow DAGs can be refined to create a fully automated closed-loop system. This engineering enhancement directly aligns with the group's broader goal of developing integrated clinical aids. This would involve implementing data drift monitoring and performance triggers that automatically launch retraining and deployment cycles when new

data becomes available or model performance degrades.

8.3.7 Extensive clinical usability testing

Building upon the excellent preliminary SUS score (85.83), broader usability testing should be conducted directly integrated into the daily workflows of specialized radiologists and oncologists to optimize the interface for specific clinical demands.

8.3.8 Clinical validation

The most critical next step is clinical validation. This would begin with a large-scale retrospective study on historical patient data to further validate the model's performance. Success in that stage would pave the way for a prospective pilot study to assess the system's real-world impact as a diagnostic aid in a clinical setting.

References

[1] DMR-IR (Database for Mastology Research with Infrared Image). Accessed: 2025-09-09.

[2] Abdulla, M.; Habaebi, M.; Islam, M. Real-time thermography for breast cancer detection with deep learning. *Discover Artificial Intelligence 4* (08 2024), 57.

[3] Al Husaini, M. A. S.; Habaebi, M. H.; Gunawan, T. S.; Islam, M. R.; Elsheikh, E. A. A.; Suliman, F. M. Thermal-based early breast cancer detection using inception v3, inception v4 and modified inception mv4. *Neural Comput. Appl. 34*, 1 (Jan. 2022), 333–348.

[4] Alzahrani, R.; Sikkandar, M.; Begum S, S.; Babetat, A.; Alhashim, M.; Alduraywish, A.; N B, P.; Ng, E. Early breast cancer detection via infrared thermography using a cnn enhanced with particle swarm optimization. *Scientific Reports 15* (07 2025), 25290.

[5] Araujo, A. S.; Issa, M. H. S.; Medeiros, P. R. T.; Sanchez, A.; Muchaluat-Saade, D. C.; Conci, A. Monitoring Breast Cancer Neoadjuvant Treatment using Thermographic Time Series . In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)* (Los Alamitos, CA, USA, June 2021), IEEE Computer Society, pp. 247–252.

[6] Araujo, A. S.; Issa, M. H. S.; Sánchez, ; Muchaluat-Saade, D. C.; Conci, A. Using thermography for breast cancer neoadjuvant treatment. In *2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP)* (2022), vol. CFP2255E-ART, pp. 1–4.

[7] Araujo, A. S.; Milena H.S., I.; Sánchez, ; Muchaluat-Saade, D. C.; Conci, A. Utilizing infrared thermography for assessing breast cancer treatment progression. In *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)* (2023), pp. 1–5.

[8] Araújo, A.; Issa, M.; Ángel Sánchez; Muchaluat-Saade, D.; Conci, A. Termografia como ferramenta de avaliação durante o tratamento neoadjuvante para câncer de mama. In *Anais do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde* (Porto Alegre, RS, Brasil, 2023), SBC, pp. 280–291.

[9] Attallah, O. Harnessing infrared thermography and multi-convolutional neural networks for early breast cancer detection. *Scientific Reports 15* (07 2025).

[10] Benitez Fuentes, J. D.; Morgan, E.; Aguilar, A.; Costa, A.; Shah, R.; Giusti, F.; Vignat, J.; Znaor, A.; Musetti, C.; Yip, C.-H.; Van Eycken, E.; Jedy-Agba, E.; Piñeros, M.; Soerjomataram, I. Global stage distribution of

breast cancer at diagnosis: A systematic review and meta-analysis. *JAMA Oncology 10* (11 2023).

[11] Borchartt, T.; Conci, A.; Lima, R.; Resmini, R.; Sanchez, A. Breast thermography from an image processing viewpoint: A survey. *Signal Processing 93* (10 2013), 2785–2803.

[12] Brookshire, G.; Kasper, J.; Blauch, N.; Wu, Y.; Glatt, R.; Merrill, D.; Gerrol, S.; Yoder, K.; Quirk, C.; Lucero, C. Data leakage in deep learning studies of translational eeg, 01 2024.

[13] Civilibal, S.; Cevik, K. K.; Bozkurt, A. A deep learning approach for automatic detection, segmentation and classification of breast lesions from thermal images. *Expert Systems with Applications 212* (2023), 118774.

[14] Costa, G. M.; Moura, E. L. S.; Borchartt, T. B.; Conci, A. Modeling the 3d breast surface using thermography. In *MICCAI Workshop on Artificial Intelligence over Infrared Images for Medical Applications (AIIIMA)* (2023), Springer.

[15] da Silva, L.; Saade, D.; Sequeiros Olivera, G.; Silva, A.; Paiva, A.; Bravo, R.; Conci, A. A new database for breast research with infrared image. *Journal of Medical Imaging and Health Informatics 4* (03 2014), 92–100.

[16] da Silva, L.; Seixas, F.; Fontes, C.; Muchaluat-Saade, D.; Conci, A. A computational method for breast abnormality detection using thermographs. pp. 469–474.

[17] da Silva, L. F.; dos Santos, A. A. S.; de Souza Bravo, R.; Silva, A. C.; Muchaluat-Saade, D. C.; Conci, A. Hybrid analysis for indicating patients with breast cancer using temperature time series. *Computer methods and programs in biomedicine 130* (2016), 142–53.

[18] de Freitas Oliveira Baffa, M. Segmentação automatizada de imagens térmicas de mama via redes neurais convolucionais para auxílio ao diagnóstico médico. *Revista de Informática Aplicada 19*, 1 (04 2017), 44–55.

[19] de Freitas Oliveira Baffa, M.; Grassano Lattari, L. Convolutional neural networks for static and dynamic breast infrared imaging classification. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (2018), pp. 174–181.

[20] Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nature Medicine 25* (01 2019).

[21] Falconi, L. G.; Pérez, M.; Benalcázar, M. E.; Conci, A. Video transformers for dynamic breast infrared imaging classification. In *Artificial Intelligence over Infrared Images for Medical Applications* (Cham, 2026), S. T. Kakileti, G. Manjunath, R. G. Schwartz, and E. Y. Ng, Eds., Springer Nature Switzerland, pp. 1–19.

[22] Fernández-Ovies, F.; Alférez, S.; DeAndrés-Galiana, E. J.; Cernea, A.; Fernández-Muñiz, Z.; Fernández-Martínez, J. L. *Detection of Breast Cancer Using Infrared Thermography and Deep Neural Networks*. 04 2019, pp. 514–523.

[23] García-Mojón, R.; Martín-Rodríguez, F.; Fernandez Barciela, M. Helping breast cancer diagnosis on mammographies using convolutional neural networks, 08 2024.

[24] Ghayoumi Zadeh, H.; Fayazi, A.; Binazir, B.; Yargholi, M. Breast cancer diagnosis based on feature extraction using dynamic models of thermal imaging and deep autoencoder neural networks. *Journal of Testing and Evaluation 49*, 3 (05 2021), 1516–1532.

[25] Goñi-Arana, A.; Pérez-Martín, J.; Díez, F. J. Breast thermography: a systematic review and meta-analysis. *Systematic Reviews 13*, 1 (2024), 295.

[26] Karim, C. N.; Ouslim, M.; Ryad, T. A new approach for breast abnormality detection based on thermography. *Medical Technologies Journal 2* (09 2018), 245–254.

[27] Kolb, T. M.; Lichy, J.; Newhouse, J. H. Comparison of the performance of screening mammography, physical examination, and breast us and evaluation of factors that influence them: An analysis of 27,825 patient evaluations. *Radiology 225*, 1 (2002), 165–175. PMID: 12355001.

[28] Kreuzberger, D.; Kühl, N.; Hirschl, S. Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access PP* (01 2023), 1–1.

[29] Litjens, G.; Kooi, T.; Ehteshami Bejnordi, B.; Setio, A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.; Ginneken, B.; Sánchez, C. A survey on deep learning in medical image analysis. *Medical Image Analysis 42* (02 2017).

[30] Litjens, G.; Kooi, T.; Ehteshami Bejnordi, B.; Setio, A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.; Ginneken, B.; Sánchez, C. A survey on deep learning in medical image analysis. *Medical Image Analysis 42* (02 2017).

[31] Mahoro, E.; Akhloufi, M. Breast cancer classification on thermograms using deep cnn and transformers. *Quantitative InfraRed Thermography Journal 21* (09 2022), 1–20.

[32] Mashekova, A.; Zhao, M. Y.; Zarikas, V.; Mukhmetov, O.; Aidossov, N.; Ng, E. Y. K.; Wei, D.; Shapatova, M. Review of artificial intelligence techniques for breast cancer detection with different modalities: Mammography, ultrasound, and thermography images. *Bioengineering 12*, 10 (2025).

[33] Mendes, L.; Rodrigues, E.; Izidoro, S.; Conci, A.; Liatsis, P. Roi extraction in thermographic breast images using genetic algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)* (07 2020), pp. 111–115.

[34] Mirasbekov, Y.; Aidossov, N.; Mashekova, A.; Zarikas, V.; Zhao, Y.; Ng, E. Y. K.; Midlenko, A. Fully interpretable deep learning model using ir thermal images for possible breast cancer cases. *Biomimetics 9*, 10 (2024).

[35] Moura, E.; Costa, G.; Borchartt, T.; Conci, A. A tool for 3d representation of the 2d thermographic breast acquisitions. In *19th International Conference on Computer Vision Theory and Applications* (Rome, 2024), Proceedings of the 19th

International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, p. 138.

[36] Moura, E. L. S.; Conci, A. Reconstrução tridimensional a partir de termografias: Avaliação de modelos 3d e simetria das mamas. In *Anais do Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)* (2025), SBC.

[37] Munguía-Siu, A.; Vergara, I.; Espinoza-Rodríguez, J. H. The use of hybrid cnn-rnn deep learning models to discriminate tumor tissue in dynamic breast thermography. *Journal of Imaging 10*, 12 (2024).

[38] Nelson, H.; Tyne, K.; Naik, A.; Bougatsos, C.; Chan, B.; Humphrey, L. Screening for breast cancer: An update for the u.s. preventive services task force. *Annals of internal medicine 151* (11 2009), 727–37, W237.

[39] Pastura, F. C. H.; Oliveira, W. I.; Moura, E. L. S. Presenting a dataset for 3d breast surface representation using thermography. In *Artificial Intelligence over Infrared Images for Medical Applications: 4th International Conference, AIIIMA 2025, Virtual Event, November 15, 2025, Proceedings* (2025), Springer.

[40] Pramanik, S.; Bhattacharjee, D.; Nasipuri, M. Wavelet based thermogram analysis for breast cancer detection. In *2015 International Symposium on Advanced Computing and Communication (ISACC)* (2015), pp. 205–212.

[41] Princigalli, N. R. Análise de simetria em termografias: uma aplicação em mamas. Dissertação de mestrado, Universidade Federal Fluminense (UFF), Niterói, 2023.

[42] Rahman, M. A.; Khan, M. S. H.; Watanobe, Y.; Prioty, J. T.; Annita, T. T.; Rahman, S.; Hossain, M. S.; Aitijjo, S. A.; Taskin, R. I.; Dhrubo, V.; Hanip, A.; Bhuiyan, T. Advancements in breast cancer detection: A review of global trends, risk factors, imaging modalities, machine learning, and deep learning approaches. *BioMedInformatics 5*, 3 (2025).

[43] Resmini, R.; Silva, L.; Araujo, A. S.; Medeiros, P.; Muchaluat-Saade, D.; Conci, A. Combining genetic algorithms and svm for breast cancer diagnosis using infrared thermography. *Sensors 21*, 14 (2021).

[44] Resmini, R.; Silva, L. F. d.; Medeiros, P. R. T.; Araujo, A. S.; Muchaluat-Saade, D. C.; Conci, A. A hybrid methodology for breast screening and cancer diagnosis using thermography. *Computers in Biology and Medicine 135* (2021), 104553.

[45] Ring, E.; Ammer, K. Infrared thermal imaging in medicine. *Physiological measurement 33* (03 2012), R33–46.

[46] Rodriguez, S.; Loaiza-Correa, H.; Restrepo, A.; Reyes, L.; Olave, L.; Diaz, S.; Pacheco, R. Dataset of breast thermography images for the detection of benign and malignant masses. *Data in Brief 54* (05 2024), 110503.

[47] Ryan, L.; Agaian, S. Breast cancer detection using infrared thermography: A survey of texture analysis and machine learning approaches. *Bioengineering 12*, 6 (2025).

[48] Santos, L.; Lima, R.; Paiva, A.; Conci, A.; Espindola, N. A computing platform to analyze breast abnormalities using infrared images. *Medical Biological Engineering Computing 61* (12 2022).

[49] Sayed, B. A.; Eldin, A. S.; Elzanfaly, D. S.; Ghoneim, A. S. A comprehensive review of breast cancer early detection using thermography and convolutional neural networks. In *2023 International Conference on Computer and Applications (ICCA)* (2023), pp. 1–6.

[50] Silva, T. A. E. d.; Silva, L. F. d.; Muchaluat-Saade, D. C.; Conci, A. A computational method to assist the diagnosis of breast disease using dynamic thermography. *Sensors 20*, 14 (2020).

[51] Singh, D.; Singh, A. K. Role of image thermography in early breast cancer detection- past, present and future. *Computer Methods and Programs in Biomedicine 183* (2020), 105074.

[52] Starcke, J.; Spadafora, J.; Spadafora, J.; Spadafora, P.; Toma, M. The effect of data leakage and feature selection on machine learning performance for early parkinson's disease detection. *Bioengineering 12*, 8 (2025).

[53] Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks, 05 2019.

[54] Tello-Mijares, S.; Woo, F.; Flores, F. Breast cancer identification via thermography image segmentation with a gradient vector flow and a convolutional neural network. *Journal of Healthcare Engineering 2019*, 1 (2019), 9807619.

[55] U, L. D.; R, M. T. Thermal breast cancer detection using deep learning and grad-cam visualization. *Salud, Ciencia y Tecnologia 5*, None (None 2025), 1518–1518.

[56] Veerlapalli, P.; Dutta, S. A hybrid gan-based deep learning framework for thermogram-based breast cancer detection. *Scientific Reports 15* (06 2025).

[57] Wissler, E. H. Pennes' 1948 paper revisited. *Journal of Applied Physiology 85*, 1 (1998), 35–41. PMID: 9655751.

[58] World Health Organization. Breast cancer, 3 2024. Accessed: 2025-03-30.

[59] Zaharia, M. A.; Chen, A.; Davidson, A.; Ghodsi, A.; Hong, S. A.; Konwinski, A.; Murching, S.; Nykodym, T.; Ogilvie, P.; Parkhe, M.; Xie, F.; Zumar, C. Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull. 41* (2018), 39–45.

[60] Zuluaga, J. P.; Al Masry, Z.; Benaggoune, K.; Meraghni, S.; Zerhouni, N. A cnn-based methodology for breast cancer diagnosis using thermal images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging Visualization 9* (10 2020), 1–15.

[61] Çevik, K.; Civilibal, S.; Bozkurt, A.; Dandil, E. Enhanced lesion classification based on yolo architectures using thermal breast images on a patient by patient basis. *Traitement du Signal 41* (12 2024), 2989–2999.

[62] Çivilibal, S.; Çevik, K. K.; Bozkurt, A. Derin Öğrenme yardımıyla aktif termogramlar Üzerinden meme lezyonlarının sınıflandırması. *Süleyman Demirel University Faculty of Arts and Science Journal of Science 18*, 2 (2023), 140–156.