RIO

PUC

Dissertação de Mestrado

Automated Breast Density Assessment In Mammograms Using Deep Learning

Juliana Heluy do Prado

Pontifícia Universidade Católica do Rio de Janeiro
Centro Técnico Científico
Departamento de Informática

Rio de Janeiro, 29 de Agosto de 2025

Pontifícia
Universidade
Católica do
Rio de Janeiro

Dissertação de Mestrado

Automated Breast Density Assessment in Mammograms Using Deep Learning

Juliana Heluy do Prado

Orientação: Professor Alberto Barbosa Raposo
Coorientação: Dr. Jan Jose Hurtado Jauregui

Dissertation presented to the Programa de Pós–graduação em Informática, do Departamento de Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática.

Rio de Janeiro, August 29th, 2025

Pontifícia
Universidade
Católica do
Rio de Janeiro

Automated Breast Density Assessment in Mammograms Using Deep Learning

Juliana Heluy do Prado

Dissertation presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the Examination Committee:

Professor Alberto Barbosa Raposo
Orientador
Departamento de Informática - PUC-Rio

Dr. Jan Jose Hurtado Jauregui
Co-Orientador
Instituto Tecgraf - PUC-Rio

Professor Paulo Ivson Netto Santos
Departamento de Informática - PUC-Rio

Dr. Augusto Ícaro Farias da Cunha
Immersed Inc

Rio de Janeiro, August 29th, 2025

Pontifícia
Universidade
Católica do
Rio de Janeiro

Juliana Heluy do Prado
Graduated in Information Systems from the Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio).

Ficha Catalográfica

Prado, Juliana Heluy do

Automated Breast Density Assessment in Mammograms Using Deep Learning / Juliana Heluy do Prado; advisor: Alberto Barbosa Raposo; co-advisor: Jan Jose Hurtado Jauregui.
– 2025.

54 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2025.

Inclui bibliografia

1. Informática – Teses. 2. Mamografia. 3. Densidade Mamária. 4. Aprendizado Profundo. 5. Classificação de Imagens. 6. Segmentação de Imagens. I. Raposo, Alberto Barbosa. II. Jauregui, Jan Jose Hurtado. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

To my father, my partner, my professors and my friends
for their support and encouragement.

Acknowledgments

I would like to express my special gratitude to those who made this journey possible.
First, to my advisor, Professor Alberto Raposo, whose support allowed me to complete this work.
To my co-advisor, Jan Hurtado, for his support. Without his assistance, this work would not have been possible.
To Luiz Fernando Trindade Santos, an extra thank you for accompanying me throughout my academic journey, from the challenges of my undergraduate thesis to the achievement of my master's degree.
To my friends and family, especially my father, Jason Prado, and Fernando Santos, who encouraged and supported me from beginning to end.
To Professor Marco Molinaro, for his invaluable help.
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.
To all of you, my heartfelt thanks.

Abstract

Prado, Juliana Heluy do; Raposo, Alberto Barbosa (Advisor); Jauregui, Jan Jose Hurtado (Co-Advisor). **Automated Breast Density Assessment in Mammograms Using Deep Learning**. Rio de Janeiro, 2025. 54p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Breast density is a critical factor in mammographic screening, as high-density tissue not only increases the risk of breast cancer but also reduces the sensitivity of lesion detection. This study proposes a comprehensive deep learning-based approach for automated breast density assessment in mammography images. The approach integrates a convolutional neural network (CNN) for classifying images into BI-RADS density categories and a semantic segmentation model for localizing dense fibroglandular tissue. To address limitations in existing datasets, we construct a balanced fusion dataset combining multiple sources, with refined annotations focused specifically on dense tissue regions. In addition to classification and segmentation, we introduce methodologies for generating structured, interpretable reports designed to assist clinical experts in decision-making. We evaluate multiple state-of-the-art models, achieving high performance across multiple metrics (AUC, F1-score, Dice coefficient, and IoU). The results demonstrate that the proposed system provides reliable and reproducible assessments, aligning closely with expert annotations and supporting the standardization of breast density evaluation in clinical practice.

Keywords

Mammography; Breast Density; Deep Learning; Image Classification; Image Segmentation.

Resumo

Prado, Juliana Heluy do; Raposo, Alberto Barbosa; Jauregui, Jan Jose Hurtado. **Avaliação Automatizada da Densidade Mamária em Mamografias Usando Aprendizado Profundo**. Rio de Janeiro, 2025. 54p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A densidade mamária é um fator crítico no rastreamento mamográfico, pois o tecido de alta densidade não apenas aumenta o risco de câncer de mama, mas também reduz a sensibilidade na detecção de lesões. Este estudo propõe uma estrutura abrangente baseada em deep learning para avaliação automatizada da densidade mamária em imagens de mamografia. A abordagem integra uma rede neural convolucional (CNN) para classificar as imagens nas categorias de densidade BI-RADS e um modelo de segmentação semântica para localizar o tecido fibroglandular denso. Para superar limitações de conjuntos de dados existentes, construímos um conjunto de dados balanceado e integrado, combinando múltiplas fontes, com anotações refinadas focadas especificamente nas regiões de tecido denso. Além da classificação e segmentação, introduzimos metodologias para gerar relatórios estruturados e interpretáveis, projetados para auxiliar especialistas clínicos na tomada de decisão. Avaliamos diversos modelos do estado da arte, alcançando alto desempenho em múltiplas métricas (AUC, F1-score, coeficiente de Dice e IoU). Os resultados demonstram que o sistema proposto fornece avaliações confiáveis e reproduzíveis, alinhando-se de perto às anotações de especialistas e apoiando a padronização da avaliação da densidade mamária na prática clínica.

Palavras-chave

Mamografia; Densidade Mamária; Aprendizado Profundo; Classificação de Imagens; Segmentação de Imagens.

Table of contents

1 Introduction **12**

2 Related work **16**
2.1 Breast Density Classification 16
2.2 Breast Density Segmentation 19
2.3 Related Works Summary Table 20

3 Methodology **22**
3.1 Datasets 23
3.2 Breast Density Classification Methodology 26
3.3 Dense Fibroglandular Tissue Segmentation Methodology 27

4 Breast Density Assessment Report Generation **29**
4.1 Workflow 30
4.2 Assessment Report Methodology 31

5 Results **39**
5.1 Evaluation of breast density classification 39
5.2 Evaluation of semantic segmentation of dense fibgroglandular tissue 42
5.3 Discussion of assessment reports 45
5.4 Analysis of the related works studies 46
5.5 Limitations 47

6 Conclusion **49**

7 Bibliography **51**

List of figures

Figure 1.1 Breast density categories (Source: (Ortegón et al., 2024)) 14

Figure 3.1 Fibroglandular tissue dense region annotations. First column: mammography image. Second column: original annotation included in (Sierra-Franco et al., 2024). Third column: our manual annotation focusing on dense regions. First row: a sample from density category A. Second row: a sample from density category B. Third row: a sample from density category C. Fourth row: a sample from density category D. The nipple, pectoral muscle, fibroglandular tissue, fatty tissue are colored in green, blue, magenta, and yellow, respectively. The background is colored in black. 26

Figure 4.1 Workflow for breast density assessment report generation. The process integrates the outputs of independent classification and segmentation models to compute quantitative metrics and enhanced visualizations that assist radiologists in mammogram interpretation. 30

Figure 4.2 Localization of dense tissue regions in the upper and lower quadrants of the breast. Left: predicted segmentation masks for the nipple ($\mathbf{M}_{\text{nip}}$), pectoral muscle ($\mathbf{M}_{\text{pec}}$), and dense tissue ($\mathbf{M}_{\text{den}}$). Middle: estimated PEC and PNL following the method of (Sierra-Franco et al., 2024). Right: segmentation of $\mathbf{M}_{\text{den}}$ into upper (purple) and lower (orange) quadrant regions. 32

Figure 4.4 Image patches extracted around the sampled points. Left: Original mammogram. Right: Pre-processed version used as input for the deep learning models. 34

Figure 4.5 Patches visualization. Each row represents a different patch and each column represents each of the proposed visualizations. The contour color maps the corresponding patch correspondence. 35

Figure 5.1 Confusion matrix for EfficientNetB3 on test set of Fusion dataset. 41

Figure 5.2 EfficientNetB3 classification examples. The red label indicates the ground truth density class (gt), the predicted density class (pred), and the prediction confidence (parenthesis). 41

Figure 5.3 Confusion matrix for EfficientNetB3 on INbreast dataset. 42

Figure 5.4 Confusion matrix for EfficientNetB3 on RSNA dataset. 42

Figure 5.5 U-Net segmentation examples. Each row represents a distinct case with varying breast densities (A, B, C, and D). First column: input mammography image. Second column: ground truth annotation. Third column: model's prediction. Foruth column: uncertainty map. The nipple, pectoral muscle, fibroglandular tissue, fatty tissue are colored in green, blue, magenta, and yellow, respectively. The background is colored in black. 44

List of tables

Table 2.1 Summary of Related Work on Automated Breast Density Assessment 21

Table 3.1 Breast density classification dataset 24
Table 3.2 External test-only datasets for cross-domain evaluation 24

Table 4.1 Breast density metrics example. 37

Table 5.1 Breast density classification results on test set of Fusion dataset. 40
Table 5.3 Semantic segmentation results on test set of Private dataset 43
Table 5.4 Analysis of Model Performance Metrics with Related Works: Classification 46
Table 5.5 Analysis of the Model Performance Metrics with Related Works: Segmentation 47

List of Abreviations

ANN – Artificial Neural Network

CNN – Convolutional Neural Network

IoU – Intersection over Union

MLO – Mediolateral oblique

CC – Craniocaudal

BI-RADS – Breast Imaging-Reporting and Data System

1
Introduction

Breast cancer represents one of the leading causes of mortality among women worldwide (World Health Organization, 2020). In Brazil, according to estimates from the National Cancer Institute (INCA), it is expected that approximately 73,610 (Instituto Nacional de Câncer, 2022a) new cases will be diagnosed annually by 2025. A determining factor for early detection of this neoplasm is related to breast density, which plays a crucial role both in tumor identification and in assessing the risk of disease development.

The breast is composed of two main types of tissue: fibroglandular tissue, responsible for milk production and transport, and adipose tissue, which corresponds to fat. Breasts with higher density, characterized by an elevated proportion of fibroglandular tissue relative to adipose tissue, present greater difficulty for tumor identification in mammography examinations. Studies, such as the one conducted by Nataliia Moshina (Moshina et al., 2018b), indicate that women with high breast density have a risk of developing breast cancer four to six times higher compared to those with predominantly fatty breasts. Thus, breast density assessment is essential for prevention and early diagnosis, complementing mammography.

Mammography is an imaging examination that uses low-dose radiation to generate detailed images of breast tissue. This method allows for the detection of alterations, such as nodules or micro calcifications, before they become palpable. According to data from the Ministry of Health, in Paraíba, non-palpable nodules are identified in 39% of breast cancer cases diagnosed by mammography (Tot, 2014; Boyd et al., 2010).

The main projections used in mammography are craniocaudal (CC) and mediolateral oblique (MLO) (Moshina et al., 2018a). In the CC projection, the breast is compressed vertically, providing a broad view of breast tissue.

The MLO projection, performed at a diagonal angle, covers from the outer portion of the breast to the axillary region, including deeper tissues and lymph nodes, which are fundamental for assessing possible disease dissemination. Due to its standardization and comprehensiveness, the MLO projection was selected as the focus of this study, although the proposed method can be equally applied to the CC projection. In addition to identifying tumors in early stages, mammography plays a crucial role in breast density assessment, which, as mentioned, constitutes a significant risk factor. Dense breasts can obscure lesions, such as nodules, micro calcifications, or alterations in tissue density, making interpretation by radiologists more difficult.

Breast density classification is performed according to the BI-RADS (Breast Imaging-Reporting and Data System) system (Smith et al., 2019), developed by the American College of Radiology to standardize the interpretation of breast imaging examinations. This system categorizes breast density into four levels:

- **Category A**: Predominantly fatty breasts;
- **Category B**: Breasts with areas of scattered fibroglandular tissue;
- **Category C**: Heterogeneously dense breasts;
- **Category D**: Extremely dense breasts.

This classification has direct implications for both breast cancer development risk and mammography sensitivity, as denser breasts reduce examination accuracy (Instituto Nacional de Câncer, 2022b). Figure 1.1 illustrates these categories. Additionally, the most recent version of BI-RADS recommends that reports include a detailed description of the location of dense regions in the breast. For this purpose, the breast is segmented into specific regions (superior, inferior, posterior, and center-anterior), facilitating communication and analysis of radiological findings.

Deep learning methods have demonstrated noticeable success in medical image analysis (Ma et al., 2024; Hurtado et al., 2024; Wang; Wang; Zhang,

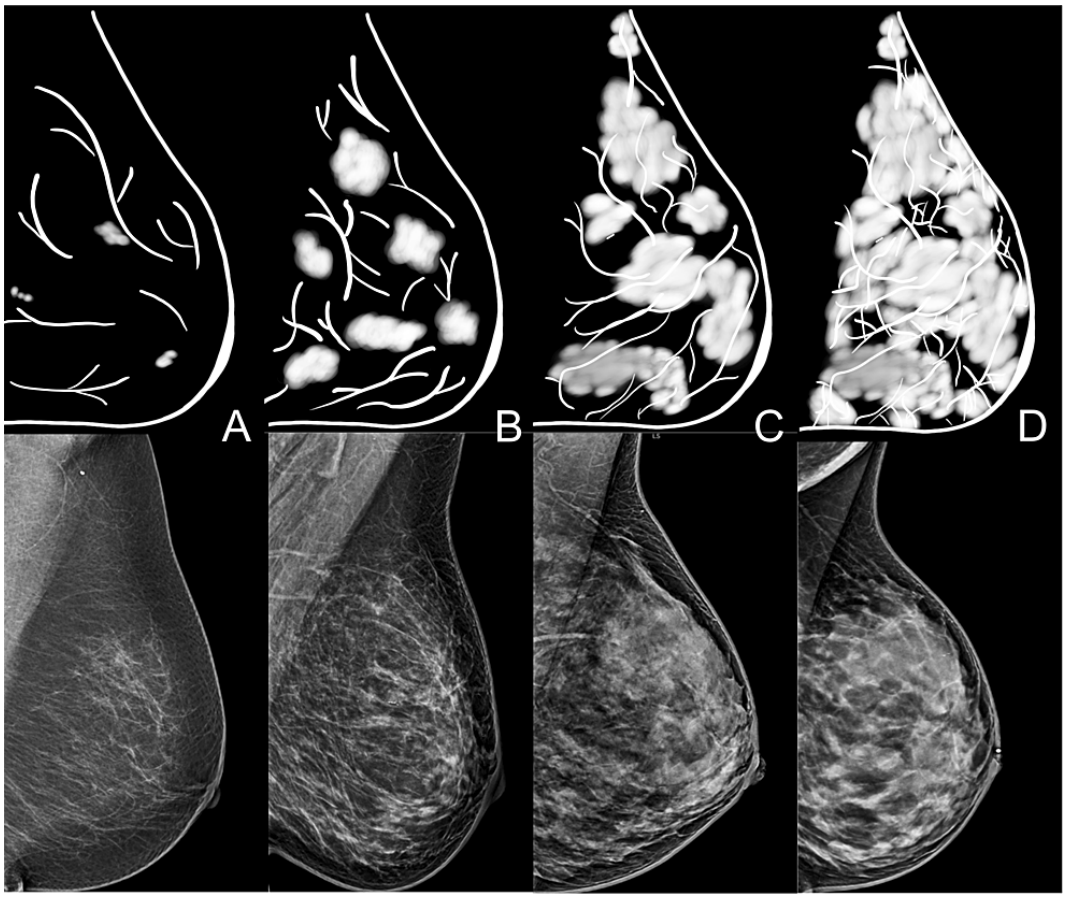

Figure 1.1: Breast density categories (Source: (Ortegón et al., 2024))

2025), including mammography images. These methods can be employed for lesion detection and localization (Michael et al., 2021), tumor classification (Salama; Aly, 2021), image enhancement (Singla et al., 2022), landmarks estimation (Sierra-Franco et al., 2024), cancer risk prediction (Yala et al., 2019), and breast density estimation (Saffari et al., 2020). These advancements automate complex tasks, reduce radiologist workload, and enable earlier and more precise breast cancer detection. Specifically, in breast density assessment, deep learning models have been employed to segment dense tissue regions or classify entire images into predefined density categories.

This thesis presents a deep learning approach for breast density analysis, integrating a classifier for BI-RADS category prediction and a segmentation model for dense tissue localization. The proposed method aims to enhance clinical decision-making by providing objective, reproducible, and interpretable results. Our main contributions can be summarized in the following points:

- Introduce a combined mammography breast density classification dataset with balanced class distributions.
- Develop a novel mammography segmentation dataset with manually annotated dense fibroglandular tissue regions.
- Present deep learning training methodologies for both breast density classification and dense fibroglandular tissue segmentation, achieving high performance on the evaluated datasets.
- Propose a set of algorithms for generating comprehensive breast density assessment reports that integrate classification and segmentation outputs, providing both numerical metrics and visual representations.

The remainder of this document is structured as follows. Chapter 2 reviews related work, Chapter 3 details our methodology, Chapter 5 presents experimental results, and Chapter 6 concludes this work.

2
Related work

This section reviews the state-of-the-art literature on automated breast density assessment in mammography, highlighting key advancements in deep learning techniques. We organize the discussion into two primary areas: first, approaches focused on image classification to categorize breast density according to established systems like BI-RADS, which enable global density evaluation; and second, methods for semantic segmentation to precisely delineate dense fibroglandular tissue, facilitating more granular analysis and integration with clinical workflows. This review underscores the evolution from traditional feature-based methods to modern neural network architectures, identifying gaps that our work addresses through dataset fusion and report generation.

2.1
Breast Density Classification

Breast density classification is a critical study due to its direct link with cancer risk, though this study does not focus on this link per se, it does take into account the Bi-RADS method used for density classification that is globally used as the standard. However, it is important to acknowledge that Bi-RADS was not a standalone invention, but a combination of years of previous studies. Earlier methods, such as those proposed by John Wolfe and Norman Boyd, laid the foundation for density assessment. From this previous study, it was also possible to advance in the automation of breast density classification which has had recent progress leveraged by computational techniques, particularly deep learning.

Prior to the global adoption of Bi-RADS classification system, others had been studying a way to standardize the classification and reporting of mammography findings. Wolfe, for one, introduced a qualitative classification system based on mammographic parenchymal patterns in 1976, categorizing

breasts into four types: N1 (predominantly fatty, lowest risk), P1 (ductal prominence up to 25% volume), P2 (ductal prominence over 25%), and DY (dense, dysplastic, highest risk) (Wolfe, 1976). Wolfe's visual assessment linked density to cancer risk by using xeromammography, which is a specialized type of mammography. He reported a 4–6 times higher risk for DY compared to N1 (Brisson; Diorio; Mâsse, 2003). However, its subjectivity led to significant inter-observer variability, limiting reproducibility (Mendelson et al., 2015).

Since Wolfe's study still didn't show promise for clinical real world adoption, another significant study was developed by Boyd et al., in the 1980s. This system created a density assessment with a quantitative approach, classifying mammograms into six categories based on the percentage of dense tissue (0%, <10%, 10–25%, 25–50%, 50–75%, >75%) (Boyd et al., 1982). Using film-screen mammography, Boyd's method employed visual estimation or planimetric tools, demonstrating that women with >75% dense tissue had a 4–6 times higher cancer risk (Boyd et al., 2007). Despite this classification method being rigorous, it was extremely labor-intensive and its persistent variability hindered clinical adoption. Both Wolfe and Boyd methods were largely replaced by the BI-RADS system in the 1990s, which offered standardized, clinically actionable four-category density reporting (A–D) (American College of Radiology, 2013) and considering that it has become a globally accepted standard for breast imaging reporting, this study uses the 5th Edition of the Bi-RADS system.

Ever since BI-RADS was established as the standard classification system, studies were able to focus on using computational methods to detect breast density. Recent studies focus particularly in deep learning, since there was a big investment to automate and standardize breast density classification aligning with BI-RADS categories.

Major studies suggest a clear evolution, such as the one introduced by Mohamed et al. (2018) which focuses on distinguishing between nuanced BI-

RADS categories (e.g., scattered vs. heterogeneously dense) (Mohamed et al., 2018). Using a dataset of 22,000 digital mammogram images from 1,427 women screened between 2005 and 2016 at a single institution, the model was trained both from scratch and with transfer learning using a pretrained VGG16 model. Their results were of an AUC score of 0.9421 for training from scratch and 0.9265 for transfer learning. However, after they refined the dataset by removing potentially wrong labeled images, their AUCs improved to 0.9882 and 0.9857, respectively.

Using a convolutional neural network (CNN), Lehman et al. (2019) developed a model to classify breast density using BI-RADS classification going from A to D and focused on clinical implementation (Lehman et al., 2019). They trained the model using over 40,000 full-field digital mammography (FFDM) exams that were simultaneously assessed by board-certified radiologists. Their model achieved a kappa score of 0.85, matching inter-radiologist agreement, and outperformed commercial software (e.g., Volpara, Quantra; kappa 0.32–0.61). When they implemented the model in real-time clinical practices, it was highly accepted by radiologists and addressed the variability in qualitative density assessments (6%–85% classified as heterogeneously/extremely dense).

Lastly, in another effort to classify BI-RADS density, Ciritsis et al. (2019) used a deep CNN that could separate into BI-RADS categories (A-D) (Ciritsis et al., 2019). The study utilized a dataset of 1,200 full-field digital mammography (FFDM) exams, including cranio-caudal (CC) and medio-lateral oblique (MLO) views from 5221 different patients (58.3 ± 11.5 years) which were downloaded from the archives of their institution and automatically sorted according to the BI-RADS density (A-D) provided by the corresponding radiological reports. They achieved an accuracy of 89%, and their kappa score reflected the high agreement they had with radiologist assessments. The model also reduced inter-observer variability by 15%, according to their

study, and compared to manual evaluations, demonstrated potential for density assessment standardization in clinical practice.

2.2 Breast Density Segmentation

In addition to breast density classification, several studies have proposed the segmentation of dense fibroglandular tissue regions (He et al., 2015). Matsubara et al. used a two-stage segmentation approach to classify breast tissue into fatty, semidense, and dense types based on histogram variance and discriminant analysis (Matsubara et al., 2001). Ferrari et al. proposed and later refined a breast density segmentation method using a weighted Gaussian mixture model, with tissue classes and model parameters estimated via the EM algorithm and guided by information-theoretic criteria (Ferrari et al., 2000; Ferrari et al., 2004). El-Zaart models mammography image histograms as mixtures of Gamma distributions, estimating parameters and selecting thresholds at histogram valleys to compute dense regions (El-Zaart, 2010). Highnam et al. proposed a relative physics-based method that integrates phase congruency, which is invariant to imaging conditions, with an iterative approach for identifying the fatty, uncompressed breast edge using realistic edge models (Highnam et al., 2010). To reduce dependency on the input image intensity distribution, Torres et al. introduced the Morphological Area Gradient (MAG) measurement, enabling adaptive segmentation of high-density regions (Torres et al., 2019).

Using a data-driven approach, Keller et al. applied c-means clustering and support vector machines to estimate dense regions in preprocessed mammograms, excluding the background and pectoral muscle (Keller et al., 2012; Keller et al., 2015). Dubrovina et al. proposed a deep learning framework employing a CNN to segment dense fibroglandular tissue along with the nipple, pectoral muscle, and fatty tissue (Dubrovina et al., 2018). Kallenberg et al. adopted an unsupervised learning strategy to pretrain a feature extractor

and used a CNN for patch-wise segmentation of dense regions (Kallenberg et al., 2016). Lee and Nishikawa employed a fully convolutional neural network with fine-tuning, initializing the weights using the ImageNet dataset (Lee; Nishikawa, 2018). Saffari et al. introduced a Generative Adversarial Network for dense region segmentation, utilizing a U-Net architecture for the generator (Saffari et al., 2020). Larroza et al. proposed a novel U-Net-based architecture trained on two ground-truth masks generated by different annotators using a threshold tool (Larroza et al., 2022). To improve U-Net performance in segmenting dense regions, Hu et al. introduced an attention mechanism that enables the model to focus on regions of interest, enhancing its ability to distinguish between different tissue types (Hu; Liu; Wang, 2022). Finally, Tiryaki and Kaplanoğlu proposed a U-Net architecture with a ResNet50 feature extractor, combined with a postprocessing step, for segmenting dense tissue, fatty tissue, and the pectoral muscle (Tiryaki; Kaplanoğlu, 2022).

2.3 Related Works Summary Table

To provide a comprehensive overview of the existing research on automated breast density assessment, we present a summary of key studies in Table 2.1. This table encapsulates critical aspects of each study, including the methodologies employed, datasets utilized, key performance metrics, and notable observations.

Table 2.1: Summary of Related Work on Automated Breast Density Assessment

Reference	**Method**	**Dataset**	**Key Metrics**	**Notes**
Mohamed et al. (2018)	CNN (VGG16, transfer learning)	22,000 mammograms (single institution)	AUC: 0.988 (refined dataset)	Focuses on nuanced BI-RADS categories; refined data improves performance.
Lehman et al. (2019)	CNN	40,000 FFDM exams	Kappa: 0.85 (matches inter-radiologist)	Outperforms commercial software; real-time clinical use.
Ciritsis et al. (2019)	Deep CNN	1,200 FFDM exams	Accuracy: 89%, Kappa: high (not specified)	Reduces inter-observer variability by 15%.
Saffari et al. (2020)	GAN with U-Net	Not specified	Dice: õ0.80 (dense regions)	Segmentation-focused; adversarial training for realism.
Larroza et al. (2022)	U-Net variant	Custom (threshold-based masks)	IoU: 0.70-0.85	Handles noisy labels from multiple annotators.
Hu et al. (2022)	U-Net with attention	Custom	Dice: 0.82	Improves focus on dense regions.
Tiryaki et al. (2022)	U-Net with ResNet50	Custom	Dice: 0.85 (dense), IoU: 0.75	Multi-label (dense, fatty, pectoral).

3
Methodology

We propose a multi-step methodology for assessing breast density in mammography images, designed to provide both quantitative and qualitative insights to aid in medical evaluation.

First, we employ a dedicated deep learning model to predict the BI-RADS density category from an input mammography image. This classification model is trained using a supervised learning strategy, leveraging a large annotated dataset of mammograms to classify breast tissue into one of the BI-RADS categories, ranging from predominantly fatty to extremely dense, without any integration with other models.

Second, in a distinct and separate step, we utilize a deep learning model specifically for segmenting dense fibroglandular tissue regions within the mammography images. This segmentation model is also trained under a supervised learning approach, with the primary goal of identifying and delineating dense areas crucial for assessing breast density, operating independently of the classification model. These segmented regions are significant because dense tissue can obscure potential anomalies, making accurate identification more challenging.

Although these two models operate independently, their outputs are later used at the analysis stage. Specifically, we use the BI-RADS category predicted by the classification model together with the segmented dense tissue regions to generate a comprehensive report, providing detailed statistics and visual representations. This summary report offers valuable insights for clinicians, delivering a high-level overview and specific visual cues to support mammogram interpretation. In the following, we present each of these steps in detail, along with the datasets used for training these independent deep learning models.

3.1 Datasets

The limitations of the breast density BI-RADS classification and segmentation datasets presented significant challenges in the development and evaluation of deep learning models, including a limited number of samples, class imbalances, variability in image acquisition protocols, and differences in label consistency between annotators. These factors can lead to biased models, reduced generalization in diverse populations, and decreased reliability in real-world clinical applications.

3.1.1 Classification dataset

This study uses two large full-field digital mammography datasets, VinDr-Mammo and the private dataset introduced in (Sierra-Franco et al., 2024), both of which exhibit class imbalances. For simplicity, we refer to the dataset from (Sierra-Franco et al., 2024) as Private throughout this study. Our experiments are conducted exclusively on the MLO view, as it offers a more comprehensive and standardized representation of breast tissues. However, the proposed methodologies are extendable to the CC view or multi-view processing.

The main issue with the VinDr-Mammo dataset is the imbalanced distribution of density classes, particularly the small number of samples in density category A. This underrepresentation can lead to biased models that struggle to classify low-density cases accurately. In the counterpart, the Private dataset suffers from a limited number of samples in density category D, which is critical because high breast density is associated with an increased risk of cancer and reduced mammographic sensitivity. To address these limitations, we construct a fusion dataset by combining samples from VinDr-Mammo and Private while preserving their respective machine learning train, validation, and test splits. This approach ensures a more balanced distribution of density

Table 3.1: Breast density classification dataset

Dataset	Split	A	B	C	D
VinDr-Mammo	Train	32	612	4892	864
	Validation	8	152	1223	216
	Test	10	190	1530	270
Private	Train	903	663	347	14
	Validation	209	231	229	130
	Test	84	96	95	90
Fusion	Train	935	1275	5239	878
	Validation	217	383	1452	346
	Test	94	286	1625	360

Table 3.2: External test-only datasets for cross-domain evaluation

Dataset	A	B	C	D
INbreast	18	87	92	9
RSNA	553	883	510	61

classes, enhancing the dataset's diversity and representativeness for robust model training and evaluation. Table 3.1 provides a detailed overview of the dataset creation process, including the distribution of samples across density classes and splits.

Additionally, we incorporate the INbreast (Moreira et al., 2012) and RSNA (Carr et al., 2022) datasets exclusively for testing and cross-domain generalization analysis, enabling the evaluation of model performance on data from different imaging domains. To ensure consistency with the 5th edition of BI-RADS, the annotations of the selected images were reviewed and, if necessary, modified by two image processing specialists, who were trained by two experienced radiologists with over 10 years of expertise in mammographic image assessment.

For the INbreast dataset, we selected all available MLO images, resulting in 18, 87, 92, and 9 samples for density categories A, B, C, and D, respectively. For the RSNA dataset, we filtered images to include only MLO views acquired from machines with identifiers 93, 190, 197, and 216, while excluding cases with breast implants or specialized acquisitions. This filtering process yielded 553, 883, 510, and 61 samples for density categories A, B, C, and D, respectively.

Table 3.2 summarizes the test-only datasets used for cross-domain evaluation.

3.1.2 Segmentation dataset

We formulate the localization of dense fibroglandular tissue regions as a semantic segmentation problem, generating per-pixel probability maps to accurately identify these regions. To construct a dataset for training deep learning models, we utilize the MLO view dataset introduced in (Sierra-Franco et al., 2024), which includes segmentation annotations for the nipple, pectoral muscle, fibroglandular tissue, and fatty tissue. However, the fibroglandular tissue annotations in this dataset encompass not only dense regions but also scattered and fatty areas where clusters of ducts or glands may be present. While this broader annotation approach provides a more comprehensive representation of breast composition, it does not directly align with the goal of breast density assessment, which focuses specifically on the extent of dense tissue. To address this problem, we generate refined annotations that specifically delineate dense regions, excluding scattered fibroglandular structures in predominantly fatty areas.

To generate refined annotations focused on dense regions, we developed an interactive annotation tool designed to facilitate fibroglandular tissue segmentation. The process begins with the fibroglandular tissue annotations from (Sierra-Franco et al., 2024), which are then adapted to isolate dense regions more accurately. The tool features a brush and eraser interface, enabling annotators to manually refine the segmentation by selecting or removing specific regions. Additionally, we integrated a contour evolution algorithm based on geodesic active contours, which dynamically adjusts the annotation to bright regions by leveraging image gradients and intensity variations. The annotation process was conducted on full MLO images from (Sierra-Franco et al., 2024) by the same annotators of the density classification dataset that were trained by clinical experts. Figure 3.1 presents examples

of the fibroglandular tissue annotations introduced in this study, alongside the annotations from (Sierra-Franco et al., 2024), highlighting the differences in dense regions segmentation. Note that the fatty tissue annotation is also adjusted, while the nipple and pectoral muscle annotations remain unchanged. The number of samples for each density category and split are described in the Private dataset of Table 3.1.

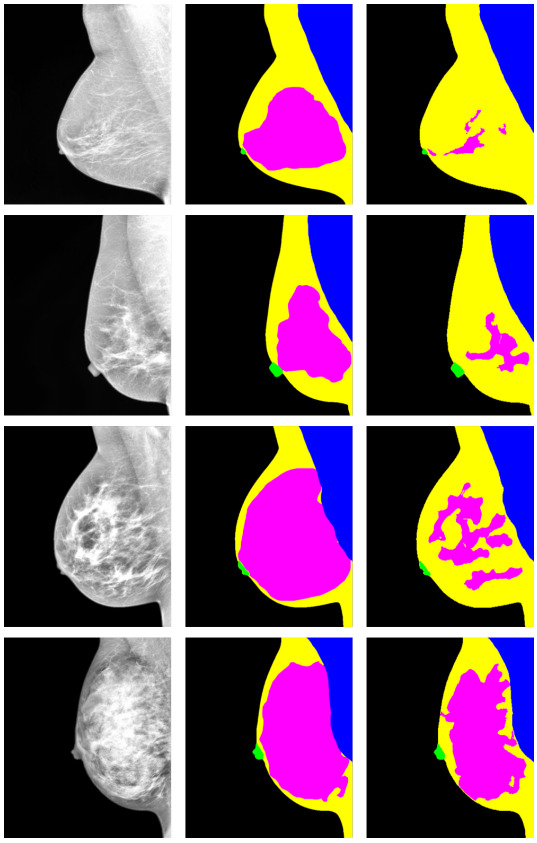

Figure 3.1: Fibroglandular tissue dense region annotations. First column: mammography image. Second column: original annotation included in (Sierra-Franco et al., 2024). Third column: our manual annotation focusing on dense regions. First row: a sample from density category A. Second row: a sample from density category B. Third row: a sample from density category C. Fourth row: a sample from density category D. The nipple, pectoral muscle, fibroglandular tissue, fatty tissue are colored in green, blue, magenta, and yellow, respectively. The background is colored in black.

3.2 Breast Density Classification Methodology

Using the labels from the fusion dataset, we model breast density assessment as a multi-class image classification problem, where the goal is to

predict the probability that a given mammography image belongs to one of the four breast density categories. To achieve this, we employ CNN architectures under a fully supervised learning paradigm. For a given CNN, we propose the following model adaptations and training configurations. At the end of the CNN, we include a fully connected layer with 256 neurons, followed by a final classification layer with four neurons and a softmax activation function, producing a probability distribution over the four density categories. The model input consists of single-channel images with dimensions 384×384 and intensity values normalized to the range $[0, 1]$. For training, we use a categorical cross-entropy loss function, a batch size of 16, and an initial learning rate of 10^{-4}, optimized using an adaptive strategy. Training is conducted for a maximum of 500 epochs, with early stopping applied based on validation loss, using a patience threshold of 50 epochs to prevent overfitting.

The input mammography images undergo preprocessing following the methodology outlined in (Sierra-Franco et al., 2024). This includes normalization using the 2nd and 98th percentiles as the minimum and maximum intensity values, contrast enhancement via Contrast Limited Adaptive Histogram Equalization (CLAHE) with a kernel size of 1/8 the image height and width, and intensity rescaling to the range $[0, 1]$. Left breast images are horizontally flipped to standardize laterality and reduce domain variability. During training, data augmentation is applied by adjusting image brightness within a range of 80% to 120%.

3.3 Dense Fibroglandular Tissue Segmentation Methodology

Using the pixel-level labels created for the dense fibroglandular tissue regions dataset, we employ fully supervised learning approaches for semantic segmentation. For this task, we propose the use of encoder-decoder deep learning architectures, which are widely adopted in the medical domain due to their ability to learn from limited annotated data, produce high-resolution

segmentations, and adapt across different imaging modalities. These architectures leverage a contracting path to capture contextual information and an expansive path to recover spatial details, enabling precise localization of dense tissue regions. Further, the encoder component, referred to as the backbone, can be replaced with various CNN architectures to create a robust feature extractor.

Building on the experiments presented in (Sierra-Franco et al., 2024) and our preliminary findings, we standardize the training configuration as follows. For a given encoder-decoder segmentation architecture, we adopt EfficientNetB3 as the backbone (Tan; Le, 2019). The model input consists of a single-channel image of size 384×384, with intensity values normalized between 0 and 1. The output is a per-pixel probability map of dimensions $384 \times 384 \times C$, where C denotes the number of classes, including an implicit background class for unannotated pixels. Since the segmentation task is framed as a multi-class per-pixel classification problem, the final layer employs a softmax activation function. For the training phase, we employ a hybrid loss function combining Categorical Focal Loss and Jaccard Loss functions, with a batch size of 4, a learning rate of 10^{-3}, and a maximum of 200 epochs, integrating early stopping with patience of 30.

For preprocessing, we apply the same procedure used in the breast density classification task. During training, beyond brightness augmentation, we incorporate the data augmentation strategy proposed in (Hurtado et al., 2025) to enhance generalization across images from different vendor equipment. This strategy combines style transfer with annotation-guided intensity manipulation to generate synthetic images, effectively increasing the diversity of the training set and improving model robustness to variations in acquisition conditions.

4
Breast Density Assessment Report Generation

The primary objective of the breast density assessment report is to provide a more comprehensive and interpretable analysis of mammograms to support radiologists in clinical decision-making. While the BI-RADS classification offers a categorical indication of breast density, it alone does not capture the full spatial and morphological characteristics of dense tissue that may affect image interpretation.

The proposed report expands this analysis by integrating the outputs of the independent classification and segmentation models, presenting quantitative measurements and enhanced visual representations in a unified format. This allows clinicians to correlate global density estimates with localized patterns of dense tissue distribution. By combining objective numerical metrics with visual cues, the report aims to reduce subjectivity in breast density assessment, improve consistency across evaluations, and highlight areas that may obscure potential lesions.

Ultimately, this report is designed not to replace radiologist expertise, but to enhance it, serving as a decision support tool that improves diagnostic confidence, facilitates comparison between examinations, and promotes a more standardized and transparent analysis of mammograms.

4.1 Workflow

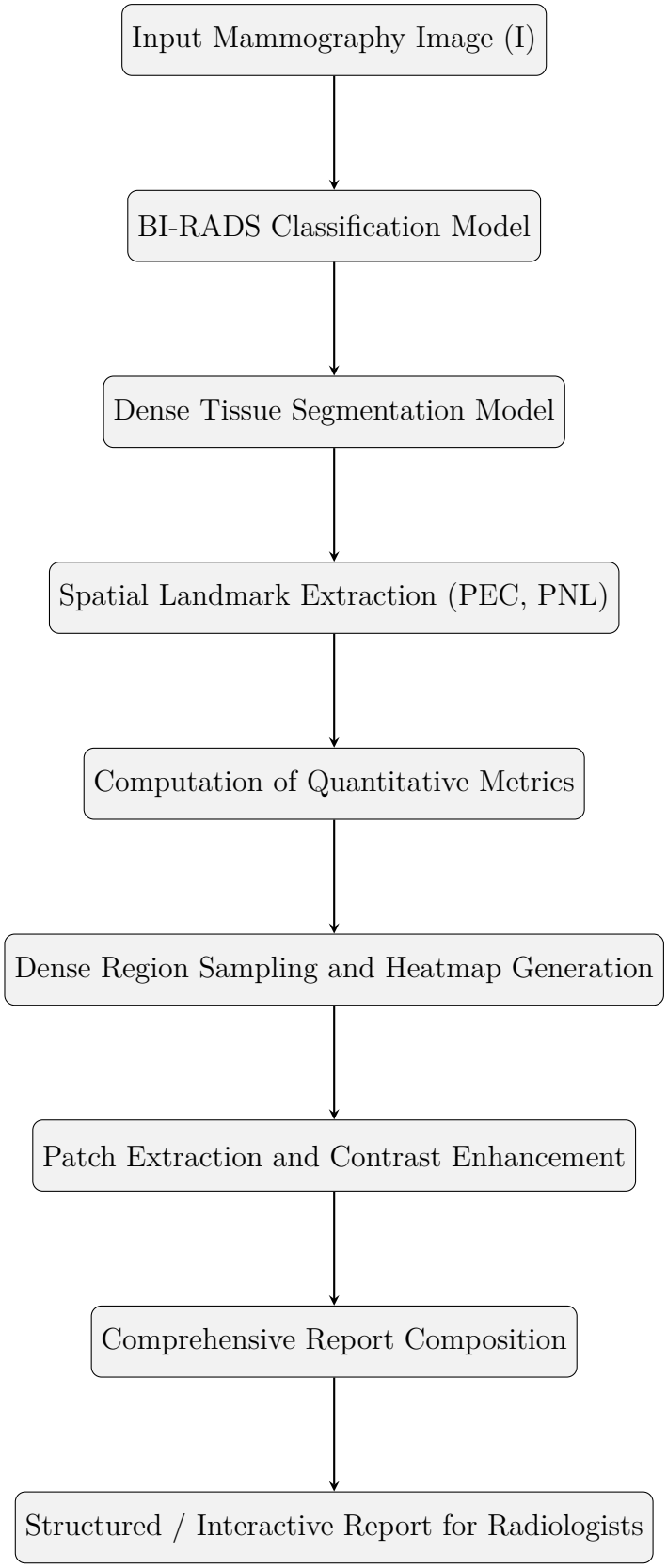

Figure 4.1: Workflow for breast density assessment report generation. The process integrates the outputs of independent classification and segmentation models to compute quantitative metrics and enhanced visualizations that assist radiologists in mammogram interpretation.

4.2 Assessment Report Methodology

By using the BI-RADS density category prediction with the segmented dense tissue regions, we generate a comprehensive report that combines statistical summaries and visual representations to support clinical interpretation.

Let us denote the original mammography image as $\mathbf{I} : \Omega \subset \mathbb{Z}^2 \to \mathcal{G}$, where $\Omega = [0, H-1] \times [0, W-1]$ is the discrete spatial domain of the image of height H and width W, and $\mathcal{G} \subset \mathbb{R}$ represents the grayscale intensity range, i.e., $\mathcal{G} = [0, 4095]$. Additionally, we assume a regular spatial sampling with uniform spacing s along both the vertical and horizontal dimensions.

Since the original image is preprocessed to match the expected input size of the segmentation model, the resulting probability maps are initially produced at a fixed resolution of 384×384. To obtain segmentation masks at the original image resolution, $H \times W$, the predicted probability maps are first resized using linear interpolation and then binarized through thresholding. We denote the resulting binary masks at the original resolution for the nipple, pectoral muscle, dense fibroglandular tissue, and fatty tissue as $\mathbf{M}_{\text{nip}}$, $\mathbf{M}_{\text{pec}}$, $\mathbf{M}_{\text{den}}$, and $\mathbf{M}_{\text{fat}}$, respectively.

To obtain a spatial distribution of the breast, the pectoral line (PEC) and posterior nipple line (PNL) are reference landmarks useful to divide the image into upper and lower breast quadrants. To estimate these lines, we adopt the method proposed in (Sierra-Franco et al., 2024), which uses the inner contour of the Convex Hull image of $\mathbf{M}_{\text{pec}}$ to approximate PEC. Then, PNL is computed by approximating a line segment that crosses PEC perpendicularly and passes over the centroid of $\mathbf{M}_{\text{nip}}$. Then, we can segment $\mathbf{M}_{\text{den}}$ into two regions, one region over the PNL and the other region under the PNL, representing the upper and lower breast quadrant regions, respectively.

To analyze the spatial distribution of breast tissue, the pectoral line (PEC) and the posterior nipple line (PNL) serve as anatomical reference landmarks for dividing the image into upper and lower breast quadrants. To

estimate these lines, we adopt the method proposed by (Sierra-Franco et al., 2024). Specifically, PEC is approximated using the inner contour of the convex hull of the pectoral muscle mask $\mathbf{M}_{\text{pec}}$. The PNL is then defined as the line perpendicular to the PEC that passes through the centroid of the nipple mask $\mathbf{M}_{\text{nip}}$. Using this reference, the dense tissue mask $\mathbf{M}_{\text{den}}$ is segmented into two regions: the portion above the PNL corresponds to the upper breast quadrant, while the portion below corresponds to the lower breast quadrant. Figure 4.2 illustrates this process using an example in which the upper quadrant contains a higher concentration of dense tissue regions.

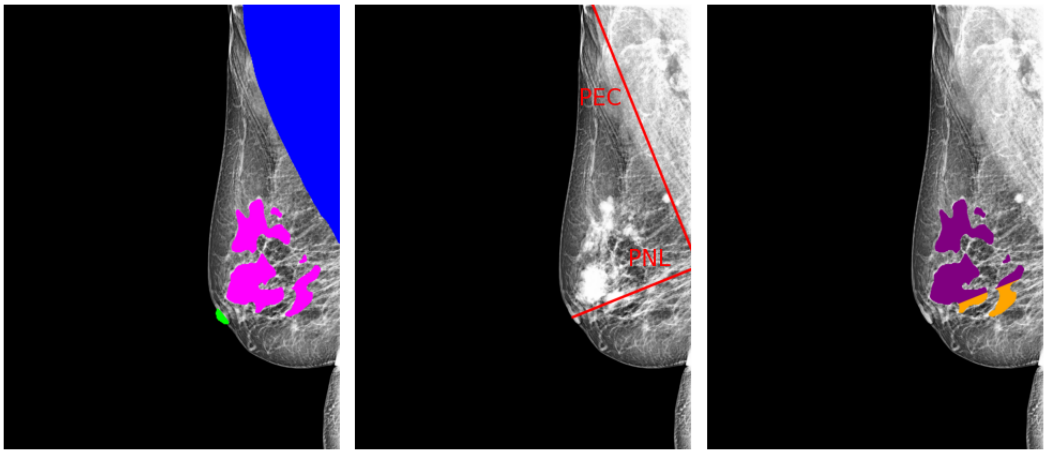

Figure 4.2: Localization of dense tissue regions in the upper and lower quadrants of the breast. Left: predicted segmentation masks for the nipple ($\mathbf{M}_{\text{nip}}$), pectoral muscle ($\mathbf{M}_{\text{pec}}$), and dense tissue ($\mathbf{M}_{\text{den}}$). Middle: estimated PEC and PNL following the method of (Sierra-Franco et al., 2024). Right: segmentation of $\mathbf{M}_{\text{den}}$ into upper (purple) and lower (orange) quadrant regions.

Although $\mathbf{M}_{\text{den}}$ is intended to represent regions of dense tissue, these areas are not always sufficiently thick to obscure underlying abnormalities. To address this limitation, we propose generating both a heatmap $\mathbf{H}_{\text{den}}$ and a binary mask $\mathbf{M}_{\text{roi}}$ that highlight regions with a higher potential to occlude such abnormalities. This process begins with the application of an exact Euclidean distance transform to the interior of $\mathbf{M}_{\text{den}}$, where each pixel is assigned a value corresponding to its distance from the nearest boundary of the mask. Pixel spacing s is taken into account to express distances in real-world units. Pixels on the boundary or outside of $\mathbf{M}_{\text{den}}$ are assigned a distance of zero. Let

us denote this distance map as $\mathbf{D}_{\text{den}}$. Next, the values of $\mathbf{D}_{\text{den}}$ are clipped using a threshold τ_{heatmap} and normalized to the range $[0, 1]$, producing a heatmap $\mathbf{H}_{\text{den}}$ that emphasizes regions with a higher likelihood of occluding abnormalities, as illustrated in Figure **??**. Additionally, a binary mask $\mathbf{M}_{\text{roi}}$ representing high-density regions can be obtained by thresholding $\mathbf{D}_{\text{den}}$ using a fixed value τ_{roi}, which we set to 0.5cm, based on the typical size of clinically relevant abnormalities. This binary mask, also shown in Figure **??**, defines a hard region of interest that warrants special attention.

The goal is to sample representative spots from $\mathbf{M}_{\text{roi}}$ for detailed analysis, as these regions may correspond to dense tissue areas that could potentially obscure underlying abnormalities, thereby increasing the difficulty of interpretation for radiologists. To perform this sampling, we first apply k-means clustering to the non-zero coordinates of $\mathbf{M}_{\text{roi}}$, using $k = 30$ to obtain a set of initial candidate centroids. We then apply Poisson disk sampling to these centroids with a maximum radius of 0.75cm to ensure spatial uniformity among the selected points. The rationale behind this approach is to first identify meaningful regions through clustering, and then refine the selection by enforcing a minimum distance between sampled locations to avoid redundancy and improve coverage. An example of these sampled points is shown in Figure **??**.

Using these sampled locations, we extract image patches that can be analyzed in isolation, applying different intensity enhancement schemes that leverage local information to emphasize subtle structures or variations in tissue density. This enables the interpreter to obtain a multi-perspective view of each region, combining raw and enhanced visualizations to better assess the presence of potential abnormalities. Figure 4.4 presents examples of extracted patches with a size of 3cm $\times$ 3cm, shown both on the original and the pre-processed images. Note that in the original image the presence of an abnormality can overshadow other dense regions, making them appear less prominent.

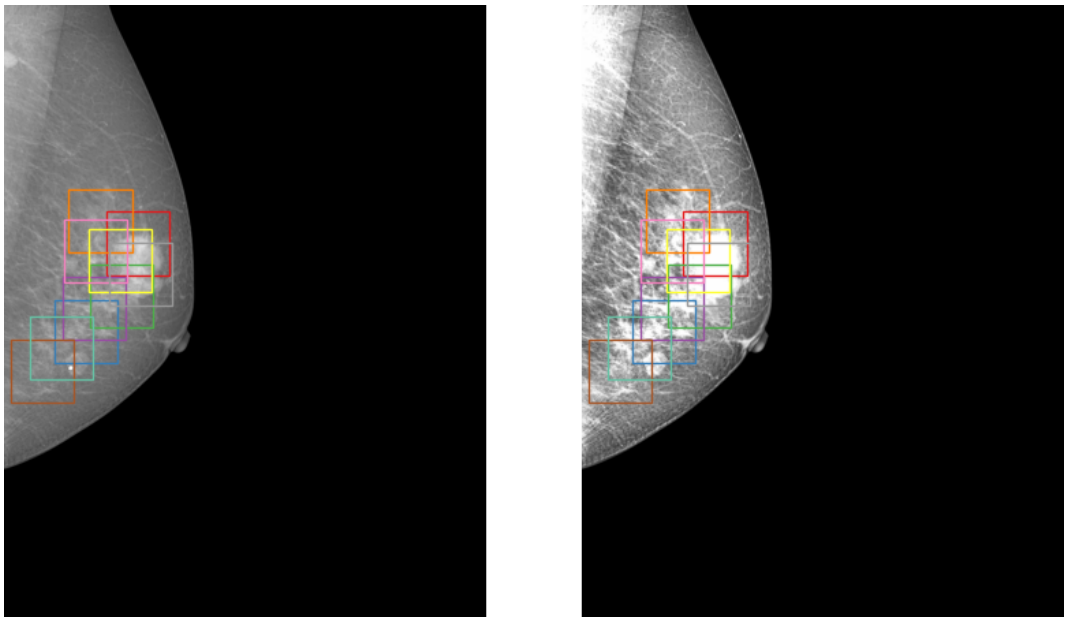

Figure 4.4: Image patches extracted around the sampled points. Left: Original mammogram. Right: Pre-processed version used as input for the deep learning models.

Once the patches are extracted, various intensity enhancement techniques can be applied to highlight regions of interest. We propose five distinct visualization strategies. The first displays the patch with its original intensity values. The second applies local min-max normalization to enhance contrast within the patch. The third uses histogram equalization, computed on the patch itself, to improve global contrast. The fourth applies adaptive histogram equalization, which enhances local contrast by considering intensity variations in small regions. Finally, the fifth approach estimates a bright region of interest within the patch using Otsu's thresholding method and then normalizes the entire patch using the minimum and maximum intensity values within that region.

Figure 4.5 illustrates these five visualizations applied to the patches shown in Figure 4.4, with colored contours indicating the correspondence between the two figures. These enhanced visualizations help to reveal subtle abnormalities that may be less apparent in the original view, particularly in the patches highlighted in red, yellow, and cyan. Moreover, the simultaneous presentation of multiple views enhances the perception of lesion texture,

supporting more detailed and informed visual analysis.

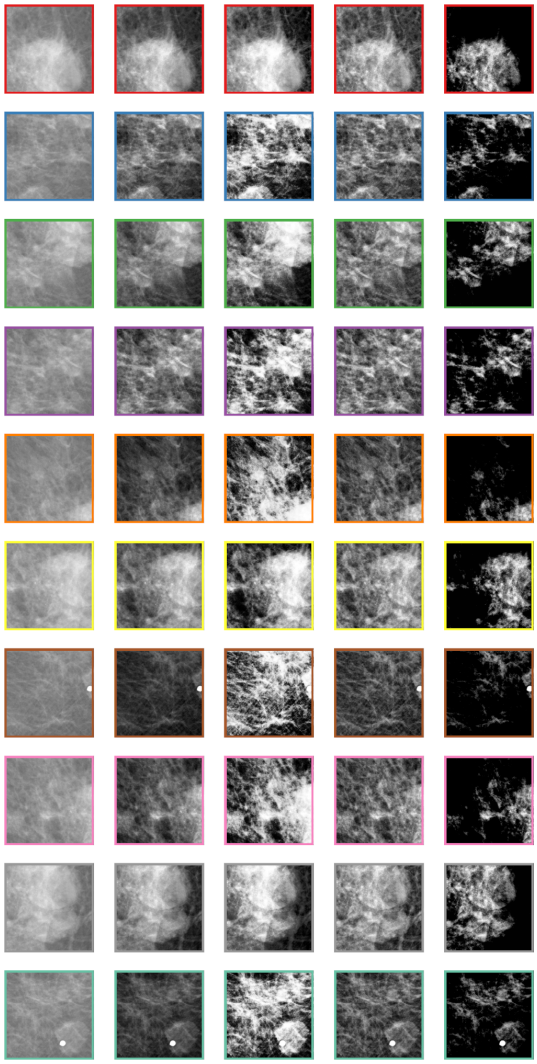

Figure 4.5: Patches visualization. Each row represents a different patch and each column represents each of the proposed visualizations. The contour color maps the corresponding patch correspondence.

Based on previous clinical studies, we propose the following set of metrics to assess breast density:

- **BI-RADS classification:** a categorical density score derived directly from our classification model, following the BI-RADS standard.
- **Absolute dense area:** the total area of dense fibroglandular tissue, computed in cm^2 as $\text{area}(\mathbf{M}_{\text{den}}) \times s^2$, where $\mathbf{M}_{\text{den}}$ is the binary mask of dense regions and s is the pixel spacing in centimeters.

- **Density proportion:** the percentage of dense tissue relative to the entire breast area, given by:

$$\frac{\text{area}(\mathbf{M}_{\text{den}})}{\text{area}(\mathbf{M}_{\text{den}}) + \text{area}(\mathbf{M}_{\text{fat}})} \times 100$$

  where $\mathbf{M}_{\text{fat}}$ is the binary mask of fatty tissue.

- **Quadrant-based Dense Area:** the dense tissue area within the upper and lower quadrants, computed as $\text{area}(\mathbf{M}_{\text{quad}}) \times s^2$, where $\mathbf{M}_{\text{quad}}$ corresponds to the quadrant-specific dense tissue masks.

- **Maximum dense region thickness:** an estimate of the maximum local thickness of dense tissue, which may indicate regions likely to obscure abnormalities. It is computed as $\max(\mathbf{D}_{\text{den}}) \times s$, where $\mathbf{D}_{\text{den}}$ is the Euclidean distance transform of $\mathbf{M}_{\text{den}}$.

- **Number of dense region spots:** the count of representative dense regions, obtained by clustering and uniformly sampling the interior of $\mathbf{M}_{\text{roi}}$. This serves as an indicator of the spatial dense tissue distribution.

- **Distance from nipple to closest dense spot:** Computed as the Euclidean distance between the centroid of $\mathbf{M}_{\text{nip}}$ and the closest dense spot identified within $\mathbf{M}_{\text{den}}$.

- **Distance from pectoral to closest dense spot:** Estimated as the minimum distance between the pectoral line (PEC) and any dense point in $\mathbf{M}_{\text{den}}$.

- **Dense region solidity:** Calculated over the full dense tissue mask $\mathbf{M}_{\text{den}}$ as the ratio between the total area of dense tissue and the area of its convex hull.

- **Dense region compactness:** A measure of shape regularity for $\mathbf{M}_{\text{den}}$, defined as $\frac{4\pi \cdot \text{Area}}{\text{Perimeter}^2}$.

- **Dense region aspect ratio:** Computed as the ratio between the width and height of the bounding box enclosing $\mathbf{M}_{\text{den}}$.

Solidity and compactness are geometric regularity metrics that help assess whether the dense tissue region is compact or spatially heterogeneous. When combined with the density proportion, these metrics can provide insight into overall breast density patterns. In particular, a breast with a high density proportion and solidity and compactness values close to 1 may indicate an extremely dense breast, where dense tissue is not only abundant but also uniformly distributed. Table 4.1 presents an example of the proposed breast density metrics applied to the mammogram illustrated in Figure 4.2.

Table 4.1: Breast density metrics example.

Description	**Value**
BI-RADS classification	C
Absolute dense area	18.70cm^2
Density proportion	14.10%
Upper quadrant-based dense area	15.64cm^2
Lower quadrant-based dense area	3.05cm^2
Maximum dense region thickness	1.98cm
Number of dense region spots	8
Distance from nipple to closest dense spot	1.92cm
Distance from pectoral to closest dense spot	3.48cm
Dense region solidity	0.53
Dense region compactness	0.073
Dense region aspect ratio	0.80

All the proposed metrics and visualizations can be incorporated into either static or interactive report formats, depending on the clinical workflow requirements. In a static configuration, key measurements and visual components such as the BI-RADS classification, dense tissue distribution maps, and patch-based views can be compiled into a structured report suitable for in-

tegration into radiology documentation. Alternatively, an interactive version of the report allows clinicians to explore individual elements more flexibly, including the ability to navigate between visualizations, adjust contrast enhancement, or examine spatial metrics in detail. This adaptability supports a wide range of use cases, from standard screening environments to more exploratory, research-oriented applications.

Although the proposed metrics were designed to provide clinically interpretable information, they have not yet been validated by radiology specialists. Their medical relevance lies in their ability to quantify aspects of breast density that are directly associated with diagnostic challenges, such as the extent, distribution, and morphology of dense fibroglandular tissue. These measurements complement the BI-RADS categorical classification by introducing objective, reproducible descriptors that can help reduce inter-observer variability and improve standardization in breast density reporting.

Regarding reproducibility, all metrics are computationally derived from the outputs of the segmentation and classification models, following fixed and well-defined mathematical formulations. However, some depend on image preprocessing parameters (e.g., brightness normalization, threshold values for binarization) and morphological filtering steps, which can slightly affect quantitative results across datasets with different acquisition settings or clinical configurations. Despite these dependencies, the full pipeline operates automatically, without requiring manual intervention, annotations, or subjective region definitions.

Future work will involve clinical validation through collaboration with radiologists to confirm the practical interpretability and diagnostic value of each proposed metric.

5
Results

This section presents the results obtained from the proposed method for automated breast density assessment. The performance of the classification and segmentation models is evaluated using standard quantitative metrics to assess their accuracy and generalization across different datasets. In addition, the generated breast density assessment reports are analyzed to illustrate how the integration of categorical predictions, spatial metrics, and visual representations can enhance the interpretation of mammography images.

5.1
Evaluation of breast density classification

For the breast density classification problem, we evaluate the performance of four deep learning models: MobileNetV2 (Sandler et al., 2018), InceptionResNetV2 (Szegedy et al., 2017), EfficientNetB0 (Tan; Le, 2019), and EfficientNetB3 (Tan; Le, 2019). These models are selected due to their varying architectures, computational efficiency, and proven effectiveness in medical image analysis. To assess classification performance numerically, we consider four metrics: precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). Precision and recall provide insights into the models' ability to correctly classify positive instances, while F1-score balances these metrics to account for class imbalances. AUC evaluates the overall discriminative power of each model across different decision thresholds.

Table 5.1 presents the classification performance for each breast density category across the evaluated models using the mentioned metrics. The results indicate that all models achieve consistently high metric values, achieving a performance close to an ideal classifier on the test set of Fusion dataset. This suggests that the models effectively differentiate between density categories with minimal misclassification. Additionally, we observe that smaller models

Table 5.1: Breast density classification results on test set of Fusion dataset.

Metric	**Method**	**A**	**B**	**C**	**D**	**Mean**
Precision	MobileNetV2	0.9889	0.9527	0.9908	0.9943	0.9817
	InceptionResNetv2	1.0000	0.9931	0.9988	0.9972	0.9973
	EfficientNetB0	0.9895	0.9965	0.9969	0.9972	0.9950
	EfficientNetB3	1.0000	0.9965	0.9982	1.0000	0.9987
Recall	MobileNetV2	0.9468	0.9860	0.9932	0.9667	0.9732
	InceptionResNetv2	1.0000	1.0000	0.9982	0.9944	0.9981
	EfficientNetB0	1.0000	0.9965	0.9988	0.9861	0.9953
	EfficientNetB3	1.0000	0.9965	0.9994	0.9944	0.9976
F1-Score	MobileNetV2	0.9674	0.9691	0.9920	0.9803	0.9772
	InceptionResNetv2	1.0000	0.9965	0.9985	0.9958	0.9977
	EfficientNetB0	0.9947	0.9965	0.9978	0.9916	0.9952
	EfficientNetB3	1.0000	0.9965	0.9988	0.9972	0.9981
AUC	MobileNetV2	0.9732	0.9896	0.9865	0.9828	0.9830
	InceptionResNetv2	1.0000	0.9995	0.9977	0.9970	0.9986
	EfficientNetB0	0.9998	0.9980	0.9960	0.9928	0.9967
	EfficientNetB3	1.0000	0.9980	0.9977	0.9972	0.9982

tend to degrade performance, as evidenced by the results of MobileNetV2, the most lightweight architecture among the selected models. In our preliminary experiments, we evaluated even more compact models, such as MobileNetV3-Small (Howard et al., 2019), which exhibited significantly poorer performance. This suggests that an excessively reduced model capacity may hinder the ability to capture the complex patterns necessary for accurate breast density classification.

Figure 5.1 presents the confusion matrix obtained using the EfficientNetB3 model, highlighting the model's performance across the four breast density categories. Notice that there are four misclassifications within density categories B, C, and D, suggesting areas where the model struggles to distinguish between adjacent density levels. Additionally, Figure 5.2 provides visual examples of both correctly and incorrectly classified cases. These visualizations offer insights into the model's decision-making process, illustrating challenging cases where misclassifications occur, as well as examples where the model successfully differentiates between density categories.

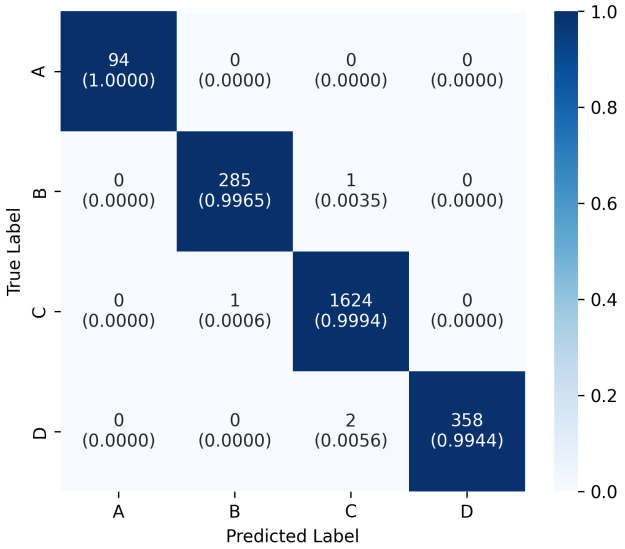

Figure 5.1: Confusion matrix for EfficientNetB3 on test set of Fusion dataset.

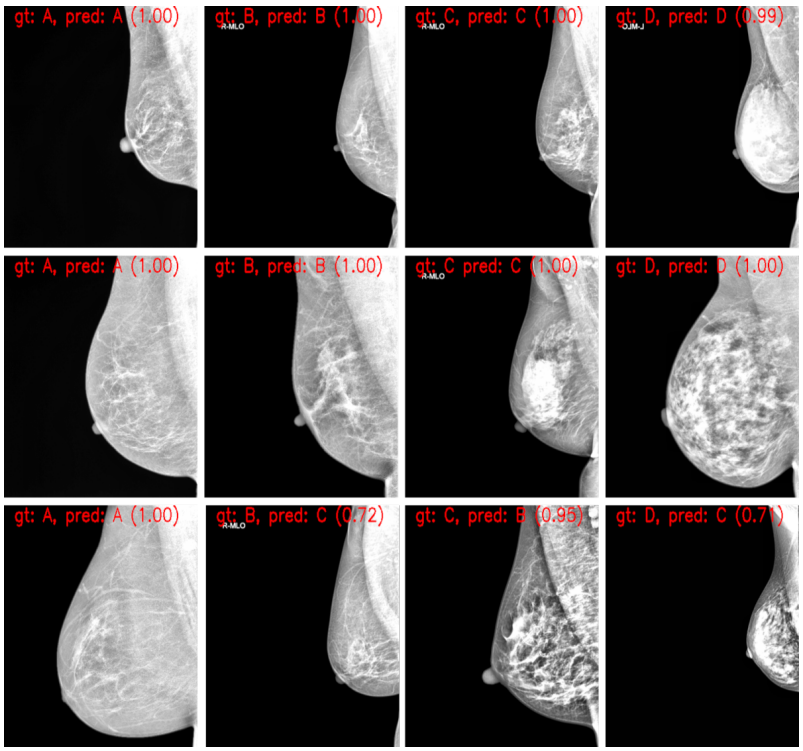

Figure 5.2: EfficientNetB3 classification examples. The red label indicates the ground truth density class (gt), the predicted density class (pred), and the prediction confidence (parenthesis).

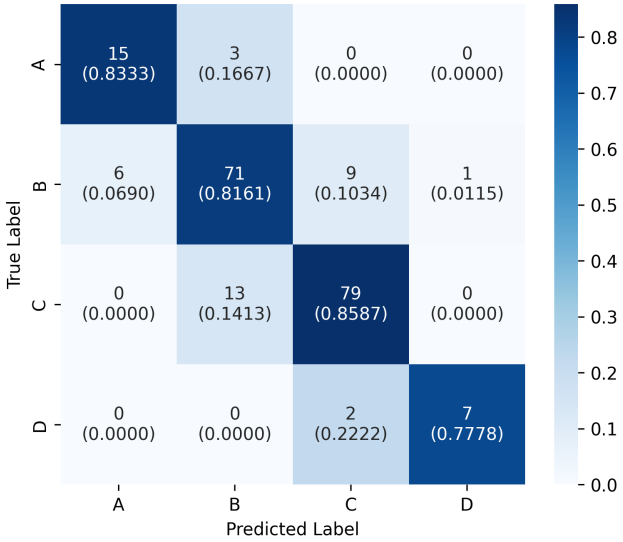

Figure 5.3: Confusion matrix for EfficientNetB3 on INbreast dataset.

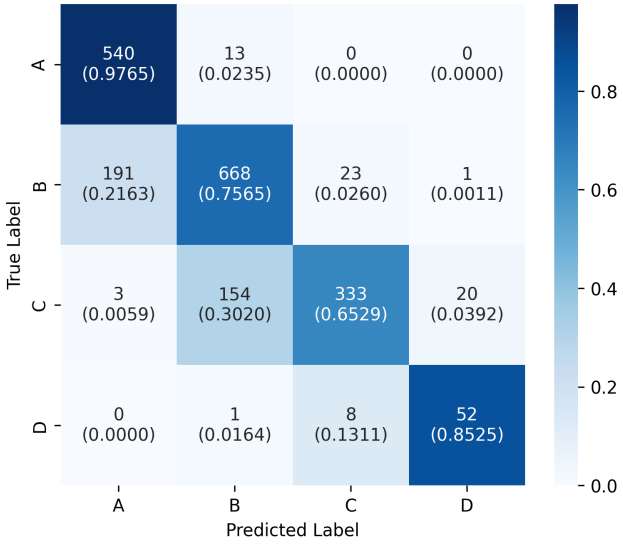

Figure 5.4: Confusion matrix for EfficientNetB3 on RSNA dataset.

5.2 Evaluation of semantic segmentation of dense fibgroglandular tissue

To evaluate semantic segmentation models for dense fibroglandular tissue region segmentation, we adopt the same segmentation architectures tested in (Sierra-Franco et al., 2024), i.e., U-Net (Ronneberger; Fischer; Brox, 2015), LinkNet (Chaurasia; Culurciello, 2017), Pyramid Scene Parsing Network (PSPNet) (Zhao et al., 2017), and Feature Pyramid Network (FPN) (Lin et al., 2017). These architectures are widely used in medical image analysis due to

Table 5.3: Semantic segmentation results on test set of Private dataset

Metric	**Method**	**Nipple**	**Pectoral**	**Dense Fib. Tissue**	**Fat. Tissue**	**Mean**
Precision	U-Net	0.8623	0.9862	0.8612	0.9503	0.9150
	LinkNet	0.8612	0.9857	0.8237	0.9563	0.9067
	PSPNet	0.8289	0.9820	0.8243	0.9485	0.8959
	FPN	0.8319	0.9793	0.8417	0.9501	0.9008
Recall	U-Net	0.8715	0.9801	0.7814	0.9671	0.9000
	LinkNet	0.8672	0.9809	0.8206	0.9563	0.9063
	PSPNet	0.8874	0.9806	0.7942	0.9572	0.9048
	FPN	0.8951	0.9844	0.7964	0.9598	0.9089
Dice	U-Net	0.8499	0.9824	0.8012	0.9575	0.8978
	LinkNet	0.8490	0.9824	0.8045	0.9552	0.8978
	PSPNet	0.8416	0.9805	0.7894	0.9515	0.8908
	FPN	0.8445	0.9804	0.7999	0.9535	0.8946
IoU	U-Net	0.7549	0.9670	0.7055	0.9217	0.8373
	LinkNet	0.7551	0.9671	0.7038	0.9176	0.8359
	PSPNet	0.7423	0.9633	0.6811	0.9112	0.8245
	FPN	0.7492	0.9647	0.6994	0.9150	0.8321

their ability to capture both local and global contextual information. For numerical evaluation, we employ standard segmentation metrics, including precision, recall, Dice coefficient, and intersection over union (IoU), to assess model performance in delineating dense fibroglandular regions. Dice coefficient and IoU measure spatial overlap between predicted and ground truth segmentations.

Table 5.3 presents the numerical results, highlighting the performance of each architecture in segmenting dense fibroglandular tissue regions. The results offer insights into the effectiveness of different segmentation approaches and their potential applicability to breast density assessment. Additionally, this table presents the results obtained for the other structures, i.e., the nipple, the pectoral muscle, and the fatty tissue. Notice that U-Net and LinkNet present very similar Dice and IoU results for the dense fibroglandular tissue, slightly superior to the other models but without significant differences. U-Net achieves the highest precision, demonstrating a noticeable advantage over the other models. This suggests that U-Net is more selective in identifying dense fibroglandular tissue regions.

Figure 5.5 presents four prediction examples on mammography images with varying breast densities. The predicted dense fibroglandular tissue regions closely align with the annotations, demonstrating consistency with BI-RADS breast density classification. The fourth column of the figure displays uncertainty maps, visualized using a hot colormap with values ranging from [0,1]. These maps, generated through test-time augmentation, highlight regions where the model exhibits higher uncertainty. Notably, the uncertainty levels are minimal, indicating high confidence and robustness in the model's predictions.

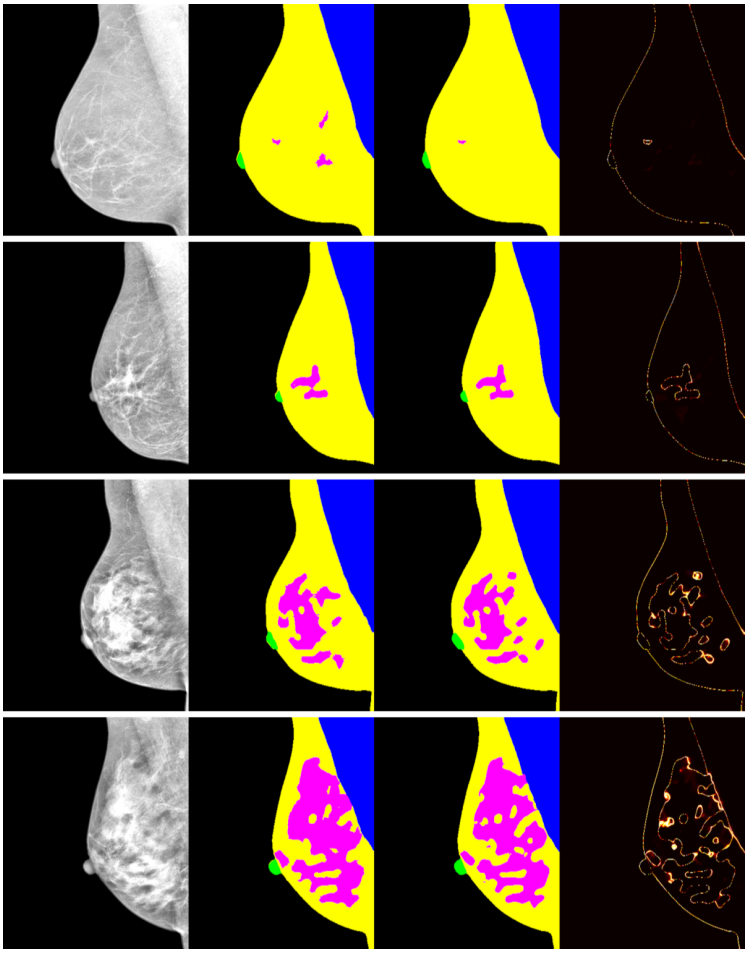

Figure 5.5: U-Net segmentation examples. Each row represents a distinct case with varying breast densities (A, B, C, and D). First column: input mammography image. Second column: ground truth annotation. Third column: model's prediction. Foruth column: uncertainty map. The nipple, pectoral muscle, fibroglandular tissue, fatty tissue are colored in green, blue, magenta, and yellow, respectively. The background is colored in black.

5.3 Discussion of assessment reports

The integration of multiple quantitative metrics and enhanced visualizations in our breast density assessment approach has important implications for clinical practice. While the BI-RADS classification remains the primary standard for categorizing breast density, it is inherently subjective and may be influenced by image quality, radiologist experience, and visual perception. Our system addresses this limitation by complementing the BI-RADS prediction with a comprehensive set of measurable features that provide deeper insight into the spatial and morphological characteristics of dense tissue.

In particular, the proposed report includes absolute and proportional measures of dense tissue, quadrant-based localization, and metrics related to shape regularity (e.g., compactness, solidity, and aspect ratio). These descriptors allow clinicians to better understand the distribution and structure of dense regions, going beyond categorical labels. For example, a case misclassified as BI-RADS C due to ambiguous global patterns may be reassessed if the density proportion, compactness, and solidity suggest characteristics closer to BI-RADS D. Conversely, cases classified as BI-RADS D but showing sparse and irregular dense regions may reflect a less obstructive and lower-risk profile.

Moreover, the spatial metrics such as the distances from the nipple and pectoral regions to the nearest dense tissue spot provide an additional layer of anatomical context. These are especially valuable for identifying critical areas where dense tissue may obscure lesions, and for understanding whether the risk is more localized or diffuse. The segmentation into upper and lower quadrants further supports targeted review and allows clinicians to verify whether dense regions follow typical patterns or present asymmetries worth investigating.

Enhanced patch-based visualizations provide radiologists with a more detailed and refined view of regions that may be clinically significant. In situations where the global density classification is ambiguous or potentially inaccurate, these localized visual cues help improve contrast and texture

perception, enabling more confident and informed interpretations.

Finally, even when the BI-RADS prediction does not align with clinical expectations, the supporting metrics and visualizations included in the report provide an objective basis for reevaluating the case. This multi-level approach can serve both as a second opinion tool and as a foundation for more standardized and explainable reporting. By combining classification, segmentation, quantitative metrics, and enhanced visualization, the proposed methodology supports more reliable and informed clinical decisions in breast cancer screening and diagnosis.

5.4 Analysis of the related works studies

As references, Tables 5.4 and 5.5 show the results of related works to the problems of density classification and segmentation, respectively, highlighting the performance of various models on these tasks. These tables provide a backdrop to contextualize the high metric values and robustness achieved by our method, even though direct comparisons are limited due to evaluations being performed on different datasets.

Table 5.4: Analysis of Model Performance Metrics with Related Works: Classification

Study/Model	**AUC**	**Accuracy/Kappa**
Mohamed et al. (2018)	0.988	-
Lehman et al. (2019)	0.95 (Kappa 0.85)	-
Ciritsis et al. (2019)	-	89% (Kappa high)

Table 5.5: Analysis of the Model Performance Metrics with Related Works: Segmentation

Study/Model	**Dice**	**IoU**
Saffari et al. (2020)	0.80	-
Larroza et al. (2022)	-	0.70-0.85
Hu et al. (2022)	0.82	-
Tiryaki et al. (2022)	0.85	0.75

5.5 Limitations

First, the performance and reliability of the metrics and visualizations depend on the accuracy of the underlying segmentation and classification models. Although we observed good qualitative results, model errors can propagate and affect the computed metrics and visual interpretations.

Second, the current approach focuses on single-view mammograms (MLO). While this simplifies analysis, clinical interpretation often relies on multiple views (e.g., CC and MLO), and extending the methodology to incorporate multi-view consistency remains an important area for future work.

Third, the clinical relevance and interpretability of certain quantitative metrics, such as compactness or spatial proximity to anatomical landmarks, have not yet been validated through radiologist studies. Their utility as decision-support tools, though promising, requires further investigation in real-world settings.

Fourth, the dataset used for development and evaluation may not fully capture the diversity of imaging devices, acquisition protocols, and population subgroups seen in clinical practice. Generalizability across institutions and imaging settings should be assessed through external validation.

Finally, while the reporting format is designed to be interpretable, its integration into clinical workflows and compatibility with existing reporting standards (e.g., structured reporting tools) has not been tested. Usability

studies involving radiologists would be valuable to refine the design and ensure practical adoption.

6
Conclusion

We proposed an integrated approach for mammographic breast density assessment that combines automated BI-RADS classification, dense tissue segmentation, and a set of interpretable quantitative and spatial metrics. Beyond providing a categorical prediction, our system delivers detailed anatomical and geometric information about dense tissue distribution, supporting more informed and explainable clinical assessments.

A key contribution of this work lies in the design of a comprehensive reporting mechanism that includes enhanced visualizations, shape regularity measures, and spatial descriptors such as distances to anatomical landmarks. These additions go beyond traditional approaches by offering objective, measurable insights into breast composition, which can be valuable in both routine screening and diagnostic follow-up.

We proposed a more balanced classification dataset through the fusion of existing sources, enabling more robust evaluation and fairer comparisons against other models by mitigating the effects of class imbalance. Additionally, we provided manually annotated dense tissue segmentation masks, offering a higher-quality ground truth compared to conventional threshold-based or semi-automated methods, which are often sensitive to noise, artifacts, and acquisition variability. Together, these contributions support more reliable benchmarking.

Although various alternative methodologies exist, including joint models that combine classification and segmentation within a single architecture, the aim of this work is not to benchmark against such approaches. Instead, the focus is on demonstrating the robustness and generalization of the proposed approach across different model architectures, while emphasizing interpretability and the use of complementary metrics. The underlying motivation is to

support clinically meaningful reporting and analysis, rather than to design an end-to-end unified system. Future work may explore more integrated strategies and include direct comparisons with existing methods.

In the next stages, efforts will be directed toward large-scale validation, assessment of clinical impact, and extension of the approach to multi-view mammograms and other imaging modalities. In the long term, this methodology has the potential to contribute to more standardized, explainable, and personalized risk assessment in breast cancer screening.

7
Bibliography

American College of Radiology. **Breast Imaging Reporting and Data System (BI-RADS)**. 5th. ed. Reston, VA: ACR, 2013.

BOYD, N. F. et al. Quantitative classification of mammographic density and breast cancer risk: Results from the canadian national breast screening study. **Journal of the National Cancer Institute**, v. 74, n. 2, p. 387–393, 1982.

BOYD, N. F. et al. Mammographic density and the risk and detection of breast cancer. **New England Journal of Medicine**, v. 356, n. 3, p. 227–236, 2007.

BOYD, N. F. et al. Mammographic density and breast cancer risk: current understanding and future prospects. **Breast Cancer Research**, v. 12, n. 6, p. 223, 2010.

BRISSON, J.; DIORIO, C.; MÂSSE, B. Wolfe's parenchymal pattern and percentage of the breast with mammographic density: Redundant or complementary classifications? **Cancer Epidemiology, Biomarkers & Prevention**, v. 12, n. 8, p. 728–732, 2003. PMID: 12917203.

CARR, C. et al. **RSNA Screening Mammography Breast Cancer Detection**. 2022. https://kaggle.com/competitions/rsna-breast-cancer-detection. Kaggle.

CHAURASIA, A.; CULURCIELLO, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In: IEEE. **2017 IEEE visual communications and image processing (VCIP)**. [S.l.], 2017. p. 1–4.

CIRITSIS, A. et al. Determination of mammographic breast density using a deep convolutional neural network. **British Journal of Radiology**, v. 92, n. 1093, p. 20180691, 2019.

DUBROVINA, A. et al. Computational mammography using deep neural networks. **Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization**, Taylor & Francis, v. 6, n. 3, p. 243–247, 2018.

EL-ZAART, A. Expectation–maximization technique for fibro-glandular discs detection in mammography images. **Computers in Biology and Medicine**, Elsevier, v. 40, n. 4, p. 392–401, 2010.

FERRARI, R. et al. Segmentation of the fibro-glandular disc in mammogrms using gaussian mixture modelling. **Medical and Biological Engineering and Computing**, Springer, v. 42, p. 378–387, 2004.

FERRARI, R. et al. Segmentation of mammograms: identification of the skin-air boundary, pectoral muscle, and fibro-glandular disc. In: MADISON, WI: MED. PHYS. **5th International Workshop on Digital Mammography**. [S.l.], 2000. p. 573–579.

HE, W. et al. A review on automatic mammographic density and parenchymal segmentation. **International journal of breast cancer**, Hindawi, v. 2015, 2015.

HIGHNAM, R. et al. Robust breast composition measurement-volpara tm. In: SPRINGER. **Digital Mammography: 10th International Workshop, IWDM 2010, Girona, Catalonia, Spain, June 16-18, 2010. Proceedings 10**. [S.l.], 2010. p. 342–349.

HOWARD, A. et al. Searching for mobilenetv3. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [S.l.: s.n.], 2019. p. 1314–1324.

HU, J.; LIU, Z.; WANG, Q. Breast density segmentation in mammograms based on dual attention mechanism. In: **Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences**. [S.l.: s.n.], 2022. p. 430–435.

HURTADO, J. et al. Improving the generalization of deep learning models in the segmentation of mammography images. **arXiv preprint arXiv:2503.22052**, 2025.

HURTADO, J. et al. Segmentation of four-chamber view images in fetal ultrasound exams using a novel deep learning model ensemble method. **Computers in Biology and Medicine**, Elsevier, v. 183, p. 109188, 2024.

Instituto Nacional de Câncer. **Estimativa 2023: incidência de câncer no Brasil**. Rio de Janeiro: INCA, 2022. Accessed: 25 Jul. 2025. Available from Internet: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa.

Instituto Nacional de Câncer. **A situação do câncer de mama no Brasil: síntese de dados dos sistemas de informação**. Rio de Janeiro: INCA, 2022. Accessed: 25 Jul. 2025. Available from Internet: https://www.inca.gov.br/publicacoes/livros/situacao-do-cancer-de-mama-no-brasil-sintese-de-dados-dos-sistemas-de-informacao.

KALLENBERG, M. et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. **IEEE transactions on medical imaging**, IEEE, v. 35, n. 5, p. 1322–1331, 2016.

KELLER, B. M. et al. Preliminary evaluation of the publicly available laboratory for breast radiodensity assessment (libra) software tool: comparison of fully automated area and volumetric density measures in a case–control study with digital mammography. **Breast cancer research**, Springer, v. 17, p. 1–17, 2015.

KELLER, B. M. et al. Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation. **Medical physics**, Wiley Online Library, v. 39, n. 8, p. 4903–4917, 2012.

LARROZA, A. et al. Breast dense tissue segmentation with noisy labels: A hybrid threshold-based and mask-based approach. **Diagnostics**, MDPI, v. 12, n. 8, p. 1822, 2022.

LEE, J.; NISHIKAWA, R. M. Automated mammographic breast density estimation using a fully convolutional network. **Medical physics**, Wiley Online Library, v. 45, n. 3, p. 1178–1190, 2018.

LEHMAN, C. D. et al. Mammographic breast density assessment using deep learning: Clinical implementation. **Radiology**, v. 290, n. 1, p. 52–58, 2019.

LIN, T.-Y. et al. Feature pyramid networks for object detection. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 2117–2125.

MA, J. et al. Segment anything in medical images. **Nature Communications**, Nature Publishing Group UK London, v. 15, n. 1, p. 654, 2024.

MATSUBARA, T. et al. An automated classification scheme for mammograms based on amount and distribution of fibroglandular breast tissue density. In: ELSEVIER. **International congress series**. [S.l.], 2001. v. 1230, p. 545–552.

MENDELSON, E. B. et al. Breast density: Clinical implications and assessment methods. **Radiographics**, v. 35, n. 2, p. 297–315, 2015.

MICHAEL, E. et al. Breast cancer segmentation methods: current status and future potentials. **BioMed Research International**, Hindawi Limited, v. 2021, p. 1–29, 2021.

MOHAMED, A. A. et al. A deep learning method for classifying mammographic breast density categories. **Medical Physics**, v. 45, n. 1, p. 314–321, 2018.

MOREIRA, I. C. et al. Inbreast: toward a full-field digital mammographic database. **Academic radiology**, Elsevier, v. 19, n. 2, p. 236–248, 2012.

MOSHINA, N. et al. Mammographic density and breast cancer risk in a norwegian screening cohort. **Acta Radiologica**, v. 59, n. 3, p. 290–297, 2018.

MOSHINA, N. et al. Automated volumetric analysis of mammographic density in a screening setting: worse outcomes for women with dense breasts. **Acta Radiologica**, v. 59, n. 3, p. 290–297, 2018.

ORTEGÓN, J. D. C. et al. Exploring dense breast density in mammography: A comparative analysis of breast cancer risk. **Cureus**, Cureus, v. 16, n. 11, 2024.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18**. [S.l.], 2015. p. 234–241.

SAFFARI, N. et al. Fully automated breast density segmentation and classification using deep learning. **Diagnostics**, MDPI, v. 10, n. 11, p. 988, 2020.

SALAMA, W. M.; ALY, M. H. Deep learning in mammography images segmentation and classification: Automated cnn approach. **Alexandria Engineering Journal**, Elsevier, v. 60, n. 5, p. 4701–4709, 2021.

SANDLER, M. et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 4510–4520.

SIERRA-FRANCO, C. A. et al. Towards automated semantic segmentation in mammography images for enhanced clinical applications. **Journal of Imaging Informatics in Medicine**, Springer, p. 1–21, 2024.

SINGLA, C. et al. Deep learning enhancement on mammogram images for breast cancer detection. **Materials Today: Proceedings**, Elsevier, v. 49, p. 3098–3104, 2022.

SMITH, R. A. et al. Cancer screening in the united states, 2019: a review of current american cancer society guidelines and current issues in cancer screening. **CA: A Cancer Journal for Clinicians**, v. 69, n. 3, p. 184–210, 2019.

SZEGEDY, C. et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In: **Proceedings of the AAAI conference on artificial intelligence**. [S.l.: s.n.], 2017. v. 31, n. 1.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. **International conference on machine learning**. [S.l.], 2019. p. 6105–6114.

TIRYAKI, V.; KAPLANOĞLU, V. Deep learning-based multi-label tissue segmentation and density assessment from mammograms. **IRBM**, Elsevier, v. 43, n. 6, p. 538–548, 2022.

TORRES, G. F. et al. Morphological area gradient: System-independent dense tissue segmentation in mammography images. In: IEEE. **2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**. [S.l.], 2019. p. 4855–4858.

TOT, T. The role of mammary gland structure in breast cancer risk. **Current Breast Cancer Reports**, v. 6, n. 4, p. 256–263, 2014.

WANG, J.; WANG, S.; ZHANG, Y. Deep learning on medical image analysis. **CAAI Transactions on Intelligence Technology**, Wiley Online Library, v. 10, n. 1, p. 1–35, 2025.

WOLFE, J. N. Breast patterns as an index of risk for developing breast cancer. **American Journal of Roentgenology**, v. 126, n. 6, p. 1130–1137, 1976.

World Health Organization. **Breast cancer**. Geneva: WHO, 2020. Accessed: 25 Jul. 2025. Available from Internet: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

YALA, A. et al. A deep learning mammography-based model for improved breast cancer risk prediction. **Radiology**, Radiological Society of North America, v. 292, n. 1, p. 60–66, 2019.

ZHAO, H. et al. Pyramid scene parsing network. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 2881–2890.