

فارس عبدالغني

breast cancer MAMO Classify V8.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:404706606

Submission Date

Aug 5, 2026, 3:35 AM GMT+3

Download Date

Aug 8, 2026, 3:05 AM GMT+3

File Name

breast cancer MAMO Classify V8.docx

File Size

4.9 MB

60 Pages

11,273 Words

70,223 Characters

75% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely revised by AI.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.



70 AI-generated 75%

Likely generated or likely generated and revised by AI.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely human generated as AI generated and likely AI generated as human generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.



Breast Cancer Mammogram Classification and Segmentation

Prepared By

Name: Moayad Khayat

ID No.: 261002503

Name: Fares Ashraf

ID No.: 221000769

Submitted in fulfilment of the requirements for the degree Master of
Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Mohamed Eassa

College of Informatics, Midocean University, Comoros

Acknowledgements

We would like to thank our supervisor, Dr. Mohamed Essa, for his continuous guidance and valuable feedback throughout this research. His expertise in both artificial intelligence and galaxy studies has greatly influenced the direction of this work. His comments challenged us to think more critically and approach the research from a scientific perspective rather than a purely technical one.

We are also thankful to the faculty members at the College of Informatics, Midocean University, for providing a supportive academic environment that inspired independent thinking and research development.

Finally, we would like to express our appreciation to our families for their patience, encouragement, and support during this journey.

Abstract

Breast cancer remains the most frequently diagnosed malignancy among women worldwide, and mammography screening is the primary method for early detection. While deep learning models have reached high accuracy on mammogram classification and tumor segmentation tasks individually, the clinical workflow still requires a radiologist to interpret imaging outputs, determine the cancer type and stage, and write a structured diagnostic report. These steps are time-consuming and subject to inter-observer variability.

This thesis proposes a two-stage pipeline that automates the full path from raw mammogram input to structured clinical report output. In the first stage, a deep learning model classifies mammographic findings (benign versus malignant, cancer subtype) and segments tumor regions within the image. In the second stage, the classification label, segmentation mask, and extracted imaging features are passed as structured input to MedGemma, a medical vision-language foundation model developed by Google. Through prompt engineering, including system prompting, few-shot examples, and chain-of-thought reasoning, MedGemma generates a clinical report that describes the cancer type, estimated stage based on available imaging features, lesion morphology, and recommended follow-up actions.

The system is evaluated on publicly available mammography datasets, including the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) and VinDr-Mammo. Classification and segmentation performance is measured using standard metrics (accuracy, F1-score, Dice coefficient, IoU). Report quality is assessed through both automated metrics (BLEU, ROUGE-L) and clinical accuracy evaluation against radiologist-written reference reports. The primary objectives of this research are to build the classification and segmentation pipeline, to develop and evaluate prompt engineering strategies for MedGemma-based report generation, and to assess the clinical usefulness of the generated reports. This work contributes a reproducible, end-to-end framework that connects imaging analysis to natural language reporting without requiring model fine-tuning, relying instead on carefully designed prompts to extract clinical reasoning from an off-the-shelf medical foundation model.

Contents

Acknowledgements	2
Abstract	3
List of Abbreviations	5
List of figures	7
Chapter 1: Introduction	1
1.1 Overview	1
1.2 Problem Definition	2
1.3 Thesis Objectives	3
1.4 Thesis Contribution	4
1.5 Thesis Organization	4
Chapter 2: Theoretical Background and Related Work	5
2.1 Introduction	5
2.2 Background	6
2.2.1 Breast Cancer and Mammographic Imaging	6
2.2.2 Deep Learning for Medical Image Classification	7
2.2.3 Medical Image Segmentation	9
2.2.4 Vision-Language Models And Medical AI	10
2.2.5 Prompt Engineering For Medical AI	11
2.2.6 Automated Radiology Report Generation	13
2.2.7 MedGemma	14
2.2.8 Mammography Dataset : CBIS-DDSM And VinDr-Mammo.....	16
2.3 Related Work	17
.....	17
2.4 Related Work Summary and Gap Analysis	24

List of Abbreviations

- **AI: Artificial Intelligence**
- **AUC: Area Under the Curve**
- **BI-RADS: Breast Imaging Reporting and Data System**
- **BLEU: Bilingual Evaluation Understudy**
- **CAD: Computer-Aided Detection/Diagnosis**
- **CBIS-DDSM: Curated Breast Imaging Subset of DDSM**
- **CLAHE: Contrast Limited Adaptive Histogram Equalization**
- **CNN: Convolutional Neural Network**
- **CoT: Chain-of-Thought**
- **DCIS: Ductal Carcinoma In Situ**
- **DL: Deep Learning**
- **DSC: Dice Similarity Coefficient**
- **FFDM: Full-Field Digital Mammography**

- **Grad-CAM: Gradient-weighted Class Activation Mapping**
- **IDC: Invasive Ductal Carcinoma**
- **ILC: Invasive Lobular Carcinoma**
- **IoU: Intersection over Union**
- **LLM: Large Language Model**
- **ML: Machine Learning**
- **NLP: Natural Language Processing**
- **ROUGE: Recall-Oriented Understudy for Gisting Evaluation**
- **TNM: Tumor-Node-Metastasis**
- **ViT: Vision Transformer**
- **VLM: Vision-Language Model**

List of figures

- Figure 1.1: Overview of the proposed two-stage pipeline: deep learning classification and segmentation followed by MedGemma prompt-engineered report generation.
- Figure 2.1: Schematic representation of the U-Net segmentation architecture for mammographic lesion delineation.
- Figure 2.2: Comparison of prompt engineering strategies (zero-shot, few-shot, chain-of-thought) applied to MedGemma for clinical report generation.
- Figure 2.3: Convolution operation on a mammogram patch showing a 3x3 kernel applied to a grayscale input region.
- Figure 2.4: Visual example of Dice coefficient computation between a predicted segmentation mask and ground-truth annotation for a breast mass.
- Figure 2.5: Illustration of the self-attention mechanism in Vision Transformers, showing query-key-value computation for image patch tokens.
- Figure 2.6: Worked example of BLEU score computation for a generated radiology report against a reference report.

List of Tables

Table 2.1: Comparison of related studies in mammogram AI classification, segmentation, and automated report generation.

Chapter 1: Introduction

1.1 Overview

Breast cancer is the most commonly diagnosed cancer worldwide. According to the World Health Organization, 2.3 million new cases were recorded in 2022 alone, and early detection through screening mammography remains the most effective strategy for reducing mortality [1]. A screening mammogram can identify tumors before they become palpable, but interpreting mammographic images and producing a clinical report requires trained radiologists who are in short supply across most healthcare systems. The average radiologist reads between 40 and 100 mammograms per day, and studies have consistently shown that inter-reader agreement on BI-RADS classification varies between 60% and 80%, depending on lesion complexity [2, 3].

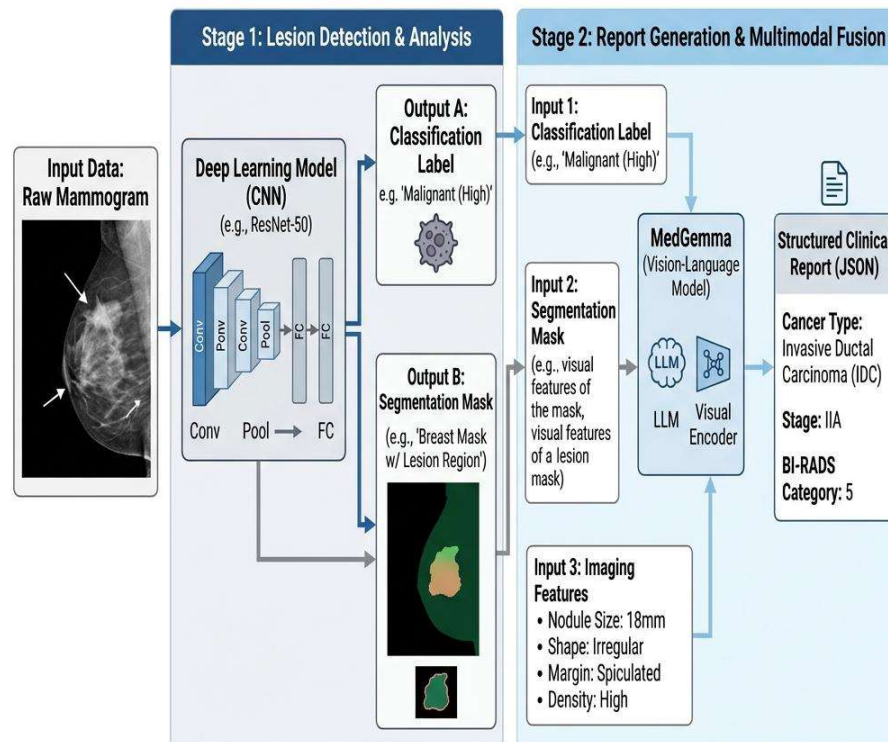
Deep learning has transformed medical image analysis over the past decade. Convolutional neural networks and their successors can now classify mammographic findings as benign or malignant with accuracy rates exceeding 90% on standard benchmarks [4]. Separately, segmentation models based on the U-Net architecture and its variants can delineate tumor boundaries with Dice coefficients above 0.90 in controlled settings [5]. These results are promising, but they address only part of the clinical workflow. A classification label or a segmentation mask is not a clinical report. The radiologist still needs to describe the lesion morphology, assign a BI-RADS category, estimate the cancer stage based on available features, specify the cancer type if identifiable, and recommend next steps. That interpretive and communicative step is where most of the clinical time goes.

Recent advances in vision-language models (VLMs) have opened a new direction. Models like MedGemma, developed by Google and trained on large-scale medical imaging and text data, can process both visual and textual inputs and generate coherent clinical narratives [6]. Rather than fine-tuning these models, which requires substantial computational resources and annotated data, prompt engineering offers a lighter alternative. By designing the right combination of system instructions, structured input formatting, and reasoning prompts, a VLM can be guided to produce clinically useful outputs without any weight modification [7].

This thesis proposes a two-stage pipeline. The first stage uses a deep learning model to classify and segment mammographic images. The second stage feeds the classification results, segmentation outputs, and extracted features into MedGemma through engineered prompts to generate a structured clinical report. The report

includes the cancer type, estimated stage based on available imaging features, lesion characteristics, and recommended actions. The goal is to automate the path from raw mammogram to written report, reducing radiologist workload and improving reporting consistency. [1, 2, 6, 7]

TWO-STAGE MEDICAL AI PIPELINE FOR MAMMOGRAM ANALYSIS & REPORT GENERATION



- Figure 1.1: Overview of the proposed two-stage pipeline: deep learning classification and segmentation followed by MedGemma prompt-engineered report generation.

1.2 Problem Definition

Current AI research in breast cancer imaging addresses individual subtasks in isolation. Classification models assign a benign/malignant label. Segmentation models outline the lesion boundary. Report generation models, where they exist, have been developed primarily for chest X-rays and lack support for mammography-specific terminology and BI-RADS formatting [8]. No published system connects these three tasks in a single pipeline that takes a mammogram as input and produces a structured clinical report as output.

Three specific gaps define the problem addressed in this thesis:

First, the disconnect between image analysis and clinical communication. A segmentation mask showing a 2.3 cm spiculated mass in the upper outer quadrant of the left breast is clinically meaningless without a narrative that identifies the finding, classifies it, and recommends a course of action. Current systems stop at the image analysis stage.

Second, the absence of prompt engineering strategies tailored for mammography report generation. While prompt engineering has been explored for chest X-ray interpretation and general medical question answering, no study has systematically designed and evaluated prompts that guide a VLM to generate mammography-specific reports covering cancer type, stage, and lesion features [7, 9].

Third, the reliance on model fine-tuning for domain adaptation. Fine-tuning requires curated instruction datasets, GPU resources, and the risk of catastrophic forgetting. Prompt engineering avoids these costs and can be iteratively refined without retraining. Whether prompt engineering alone can produce clinically acceptable mammography reports from a foundation model like MedGemma remains an open question [6, 10].

1.3 Thesis Objectives

The primary objectives of this research are:

Develop and evaluate a fine-tuned multimodal AI agent that supports Tumor Board discussions by integrating various data sources. The specific objectives are as follows:

1. **Classify Mammograms:** Develop a deep learning classification model that categorizes mammographic findings as benign or malignant and identifies the cancer subtype (e.g., invasive ductal carcinoma, invasive lobular carcinoma, ductal carcinoma in situ) where imaging features permit.

2. **Segment Tumors:** Build a segmentation model based on the U-Net architecture (or a suitable variant) that delineates tumor boundaries in mammographic images with high spatial accuracy.
3. **Engineer Prompts:** Design and evaluate multiple prompt engineering strategies (zero-shot, few-shot, chain-of-thought, and structured system prompting) to guide MedGemma in generating clinical reports from classification and segmentation outputs.
4. **Generate Reports:** Produce structured clinical reports that include the cancer type, estimated stage based on available imaging features, lesion morphology description, BI-RADS category, and recommended follow-up actions.
5. **Evaluate End-to-End:** Assess the full pipeline from raw mammogram to generated report using both computational metrics (accuracy, F1, Dice, IoU, BLEU, ROUGE-L) and clinical accuracy evaluation against radiologist-written reference reports.

1.4 Thesis Contribution

This research contributes across three domains:

- **Technical Contributions:** A two-stage pipeline architecture that connects deep learning image analysis to vision-language model report generation without requiring fine-tuning. A systematic comparison of prompt engineering strategies for mammography-specific clinical reporting with MedGemma.
- **Clinical Contributions:** An automated reporting framework that produces structured mammography reports covering cancer type, stage estimation, and lesion characteristics. The framework reduces the manual effort required from radiologists and provides a consistent report format.
- **Research Contributions:** A reproducible evaluation methodology for assessing the clinical accuracy of AI-generated mammography reports. A publicly documented set of prompt templates designed for MedGemma-based medical report generation, usable by other researchers as a starting point.

1.5 Thesis Organization

The rest of this thesis is organized as follows:

- **Chapter 2** presents the theoretical background covering breast cancer imaging, deep learning classification and segmentation, vision-language models, prompt engineering, and automated report generation. It also reviews related work and identifies the research gaps that motivate the proposed approach.
- **Chapter 3** describes the proposed methodology in detail, including the dataset preparation pipeline, the classification and segmentation model architectures, the prompt engineering strategy design, and the system integration.
- **Chapter 4** presents the experimental results, evaluating classification accuracy, segmentation quality, and report generation performance against radiologist baselines.
- **Chapter 5** concludes the thesis with a summary of findings, a discussion of limitations, and directions for future work.

Chapter 2: Theoretical Background and Related Work

2.1 Introduction

This chapter presents the theoretical framework and reviews existing literature relevant to the proposed two-stage reporting pipeline. The framework connects deep learning mammogram analysis (classification and segmentation) with automated clinical report generation via prompt engineering on a medical vision-language model.

The discussion covers three key areas. First, it outlines the clinical background of breast cancer screening, mammographic views, and standardized BI-RADS reporting lexicons. Second, it explains the underlying computational methods: convolutional neural networks for classification, U-Net encoder-decoder models for

spatial segmentation, vision-language architectures for multimodal reasoning, and prompt engineering as an inference-time steering mechanism. Third, it provides a comparative review of fourteen prior studies addressing subtasks in mammogram classification, segmentation, and automated reporting.

Section 2.2 covers the theoretical background. Section 2.3 reviews prior empirical studies across four research categories. Section 2.4 synthesizes these findings into a formal gap analysis that motivates the proposed pipeline architecture.

Foundational studies in multimodal biomedical AI confirm that combining visual features with text representations yields higher diagnostic utility than relying on either modality in isolation [11].

2.2 Background

2.2.1 Breast Cancer and Mammographic Imaging

Breast cancer develops through the uncontrolled proliferation of abnormal cells within mammary tissue, most frequently originating in the lactiferous ducts (invasive ductal carcinoma, IDC) or lobules (invasive lobular carcinoma, ILC). IDC represents roughly 70-80% of clinical diagnoses, whereas ILC accounts for 10-15%. Non-invasive precursors, notably ductal carcinoma in situ (DCIS), describe neoplastic proliferation confined to the ductal epithelium without basement membrane invasion. Distinguishing between these histological subtypes directly influences surgical planning, systemic therapy selection, and prognostic evaluation [1].

Screening mammography remains the standard imaging modality for early detection. Routine screening examinations capture two standard projections per breast: the craniocaudal (CC) view and the mediolateral oblique (MLO) view. Full-field digital mammography (FFDM) has superseded analog film-screen systems in modern clinical practice, offering enhanced dynamic range, superior soft-tissue contrast, and direct compatibility with digital image processing pipelines [2].

The American College of Radiology maintains the Breast Imaging Reporting and Data System (BI-RADS) to standardize mammographic vocabulary and report formatting. The BI-RADS lexicon defines specific terms for mass shape (round, oval, irregular), margin characteristics (circumscribed, obscured, microlobulated, indistinct, spiculated), and calcification morphology (amorphous, coarse heterogeneous, fine pleomorphic, fine linear). These descriptors determine the final BI-RADS assessment category, ranging from 0 (incomplete, requiring additional evaluation) to 6 (known biopsy-proven malignancy). The 2025 BI-RADS update refined breast density descriptors and expanded lexicon definitions across ultrasound and MRI modalities [3, 12].

To demonstrate how morphological descriptors drive clinical management, consider a representative mammographic finding: a 1.8 cm irregular mass with spiculated margins located in the upper outer quadrant of the right breast. Under the BI-RADS lexicon, an "irregular" shape combined with "spiculated" margins indicates high risk. A radiologist evaluating these features assigns a BI-RADS 5 category (highly suggestive of malignancy), which carries a positive predictive value exceeding 95% and mandates core needle biopsy [3]. Conversely, a round, circumscribed mass of identical dimensions typically receives a BI-RADS 3 classification (probably benign), recommending short-interval 6-month surveillance with a malignancy risk below 2%. The classification model in our pipeline must capture these subtle morphological distinctions, as downstream prompt generation depends directly on accurate feature extraction.

Cancer staging uses the TNM (Tumor-Node-Metastasis) classification system. While definitive staging requires pathological node assessment and systemic imaging for metastases, mammography provides an accurate estimate of primary tumor dimensions (the T component). T-stage categories based on maximum invasive diameter include: T1 (≤ 20 mm), T2 (> 20 mm to ≤ 50 mm), and T3 (> 50 mm). T4 indicates direct involvement of the chest wall or skin regardless of lesion size. For example, a 5.5 cm spiculated mass associated with skin thickening aligns with at least T3 staging. These imaging-derived measurements supply the quantitative inputs for the stage estimation component in our report generation framework [13].

2.2.2 Deep Learning for Medical Image Classification

Image classification maps an input tensor to categorical probability scores. In mammography, classification tasks involve separating benign from malignant lesions, predicting histological subtypes, or assigning BI-RADS risk categories..

Convolutional Neural Networks (CNNs) have served as the core architecture for medical vision tasks since AlexNet demonstrated deep feature extraction [14]. A convolutional layer applies learnable spatial filters across an input image tensor I . For a kernel K of size $m \times n$, the output feature map value at spatial coordinate (i, j) is defined as:

$$(I * K)(i, j) = \sum_{a=0}^{m-1} \sum_{b=0}^{n-1} I(i + a, j + b) \cdot K(a, b) \quad (2.1)$$

where $*$ represents the two-dimensional convolution operator. Nonlinear activation functions follow each convolution operation. The Rectified Linear Unit (ReLU), $f(x) = \max(0, x)$, remains standard because it provides linear gradients for positive inputs, preventing gradient saturation during backpropagation [49].

To make the convolution concrete, consider a 3x3 kernel applied to a grayscale mammogram patch (see Figure 2.3).

This is a horizontal edge detector. When this kernel slides over a region of the mammogram where pixel intensities transition sharply from dark (adipose tissue) to bright (dense lesion), the output value is large and positive. Where the region is uniform, the output is close to zero. A CNN learns hundreds of such kernels at each layer. Early layers detect edges and textures; intermediate layers combine these into shapes (round margins, spiculated borders); and deep layers assemble shape combinations into diagnostic categories (benign mass, malignant mass, calcification cluster) [14, 15].

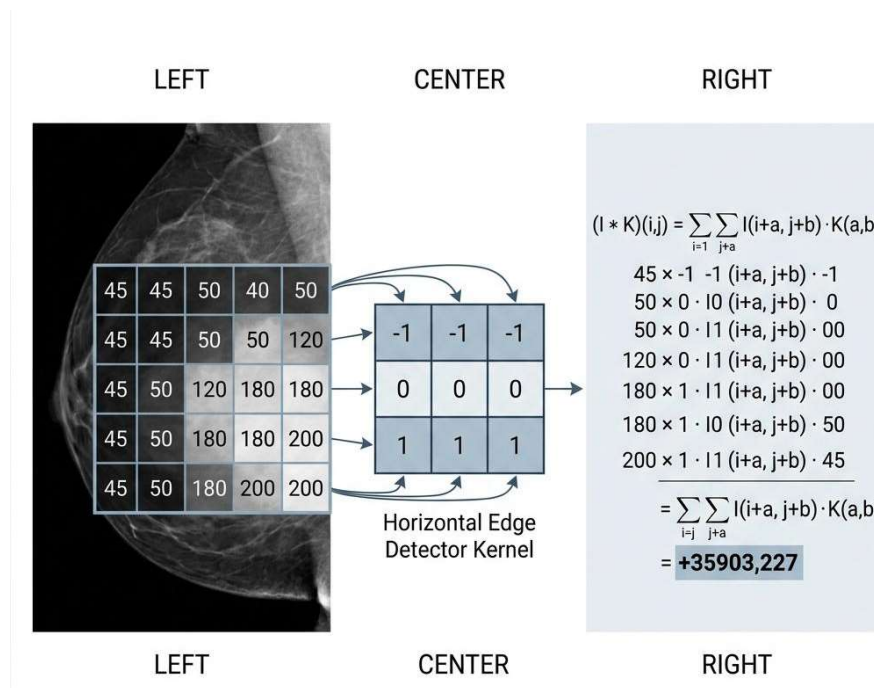


Figure 2.3: Convolution operation on a mammogram patch showing a 3x3 kernel applied to a grayscale input region.

Subsequent architectures introduced specific innovations. ResNet (2015) used skip connections (also called residual connections) to train networks of over 100 layers without degradation [15]. The skip connection adds the input of a layer block directly to its output:

$$y = F(x, \{W_i\}) + x$$

(2.2)

where x is the input, $F(x, \{W_i\})$ is the residual mapping learned by the stacked layers with weights W_i , and y is the output. This formulation allows the network to learn identity mappings when additional depth offers no benefit, which prevents the training signal from degrading as the network gets deeper. DenseNet (2017) connected each layer to every other layer in a feed-forward fashion. EfficientNet (2019) used compound scaling to balance network depth, width, and resolution for better accuracy-efficiency tradeoffs [16].

For mammogram classification, the network's final layer produces a probability distribution over the classes via the softmax function. For a classification problem with C classes, the softmax output for class c given the logit vector z is:

$$P(z) = \frac{e^{z_c}}{\sum_{k=1}^C e^{z_k}}$$

(2.3)

In the binary case (benign versus malignant), $C = 2$, and the model outputs two probabilities that sum to 1. For multi-class classification covering subtypes (IDC, ILC, DCIS, benign), $C = 4$.

The network is trained by minimizing the cross-entropy loss between the predicted probability distribution and the true label. For a single training example with true class c , the loss is:

$$L_c = -\log \log P(z)$$

(2.4)

Over a dataset of N labeled mammograms, the total training loss is the average of the individual losses. The optimizer (typically Adam or SGD with momentum) updates the network weights by computing gradients of this loss with respect to all parameters via backpropagation [49].

For mammogram classification specifically, transfer learning has proven effective. Models pre-trained on ImageNet are fine-tuned on mammography datasets, taking advantage of low-level feature representations (edges, textures) learned from natural images that transfer reasonably well to medical imaging tasks. Transfer learning works because the early convolutional layers learn generic visual features (edge orientations, color gradients, simple textures) that are common across image domains. When the pre-trained model is fine-tuned on mammograms, only the later layers and the classification head need to learn domain-specific features, which reduces the amount of labeled data required by roughly an order of magnitude compared to training from scratch [17]. Data augmentation techniques, including rotation, flipping, and Contrast Limited Adaptive Histogram Equalization (CLAHE), are routinely applied to address the limited size of annotated

mammography datasets. CLAHE operates by dividing the image into small tiles and performing histogram equalization within each tile independently, with a clip limit to prevent noise amplification. This produces locally enhanced contrast that makes subtle lesion boundaries more visible to both radiologists and classification models [17].

More recently, Vision Transformers (ViTs) have been applied to medical imaging. Unlike CNNs, which build hierarchical local feature maps through convolution, ViTs divide the image into patches, embed them as tokens, and process them through self-attention layers that can capture long-range spatial dependencies. The self-attention mechanism computes, for each patch token, a weighted sum of all other tokens where the weights are determined by learned compatibility functions. Formally, given a sequence of patch embeddings X , the self-attention output is:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

(2.5)

where $Q = X \cdot W_Q$, $K = X \cdot W_K$, and $V = X \cdot W_V$ are the query, key, and value projections (each a learned linear transformation), d_k is the dimensionality of the key vectors, and the division by $\sqrt{d_k}$ prevents the dot products from growing too large as d_k increases [18].

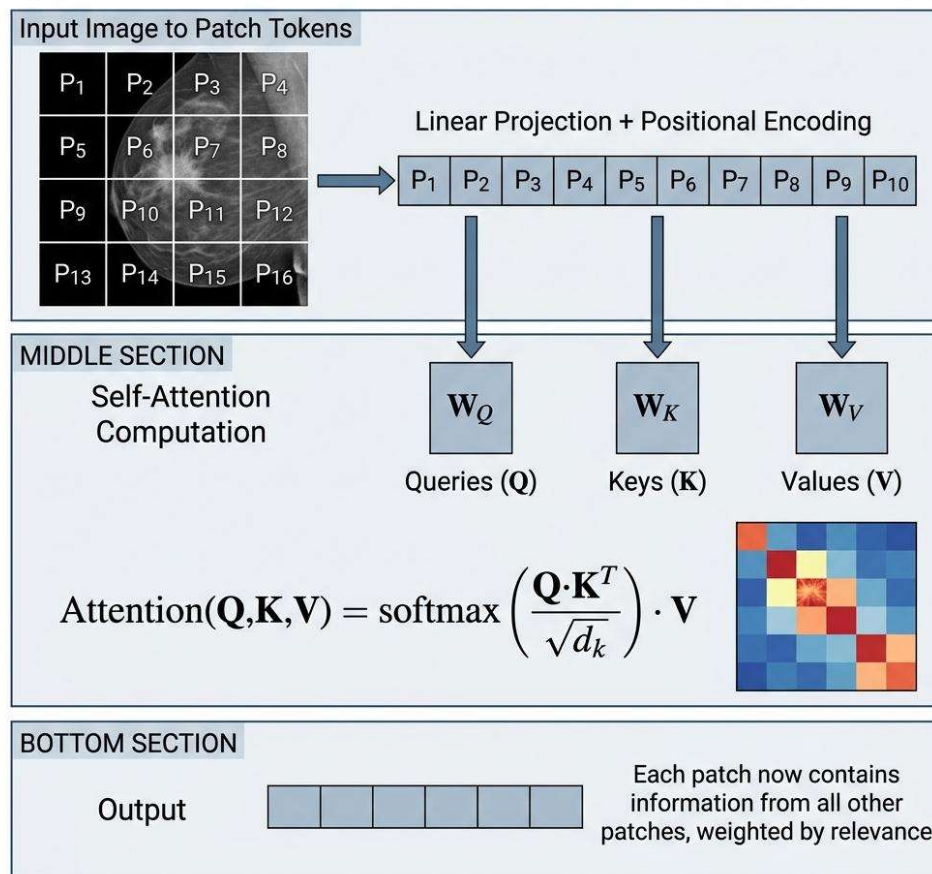


Figure 2.5: Illustration of the self-attention mechanism in Vision Transformers, showing query-key-value computation for image patch tokens.

In a mammogram, this means that a patch containing a suspicious spiculated margin can attend to distant patches showing associated architectural distortion or skin thickening, even if those patches are far apart in the image. CNNs can only capture such relationships after many stacked convolutional layers, each with a limited receptive field. ViTs capture them in a single attention layer.

Hybrid architectures that combine CNN feature extraction with Transformer-based attention have shown competitive or superior performance on mammography

benchmarks [18]. For example, a hybrid model might use a ResNet-50 backbone to extract a feature map of size 16x16x2048 from a 512x512 mammogram, then treat each spatial position in that feature map as a token and pass the 256 tokens through a Transformer encoder. This gives the model both the inductive bias of convolutions (translation invariance, local connectivity) and the global context modeling of self-attention.

Explainability remains a concern for clinical deployment. Gradient-weighted Class Activation Mapping (Grad-CAM) computes a heatmap by weighting the activations of the last convolutional layer by the gradient of the target class score with respect to those activations. Formally, the importance weight for the k -th feature map channel is:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y_c}{\partial A_k(i, j)}$$

(2.6)

where y_c is the score for class c , $A_k(i, j)$ is the activation of the k -th channel at spatial location (i, j) , and Z is the number of spatial locations. The final heatmap is the ReLU of the weighted combination: $L_{\text{Grad-CAM}} = \text{ReLU}(\sum_k \alpha_k \cdot A_k)$. The ReLU discards negative contributions, keeping only features that positively influence the class prediction [19]. These visualizations help radiologists assess whether the model is attending to the correct anatomical features. If the heatmap highlights the lesion region, the classification is likely based on relevant morphological features; if it highlights background tissue or artifacts, the prediction should be treated with suspicion.

2.2.3 Medical Image Segmentation

Segmentation assigns a label to each pixel in an image, delineating the boundaries of structures of interest. In mammography, the target structures are masses (solid lesions) and calcification clusters, though broader breast region segmentation (separating breast tissue from background) is also common as a preprocessing step.

Unlike classification, which produces a single label for the entire image, segmentation produces an output map of the same spatial dimensions as the input, where each pixel is labeled as either "lesion" (1) or "background" (0) in the binary case. This is sometimes called dense prediction because the model must make a prediction at every spatial location.

The U-Net architecture, introduced by Ronneberger et al. in 2015, remains the most widely used segmentation model in medical imaging [20]. U-Net uses an encoder-decoder structure with skip connections that transfer fine-grained spatial information from early encoder layers to corresponding decoder layers. This design allows the network to produce precise boundary delineation even with limited training data, which is typical in medical imaging applications.

The encoder follows a standard CNN pattern: each block contains two 3x3 convolutions (each followed by ReLU), then a 2x2 max pooling operation that halves the spatial dimensions. If the input mammogram is 512x512 pixels, the encoder produces feature maps at resolutions 256x256, 128x128, 64x64, and 32x32, with progressively more channels (64, 128, 256, 512) at each stage. The decoder reverses this process: each block applies a 2x2 up-convolution (transposed

convolution) that doubles the spatial dimensions, concatenates the result with the corresponding encoder feature map via the skip connection, and applies two 3x3 convolutions. The skip connections are what make U-Net work well for medical imaging: the encoder captures "what" features are present (a mass, a calcification) while the decoder uses the skip connections to recover "where" exactly those features are located.

The final layer of U-Net is a 1x1 convolution that maps the multi-channel feature map to a single-channel output, followed by a sigmoid activation to produce per-pixel probabilities between 0 and 1.

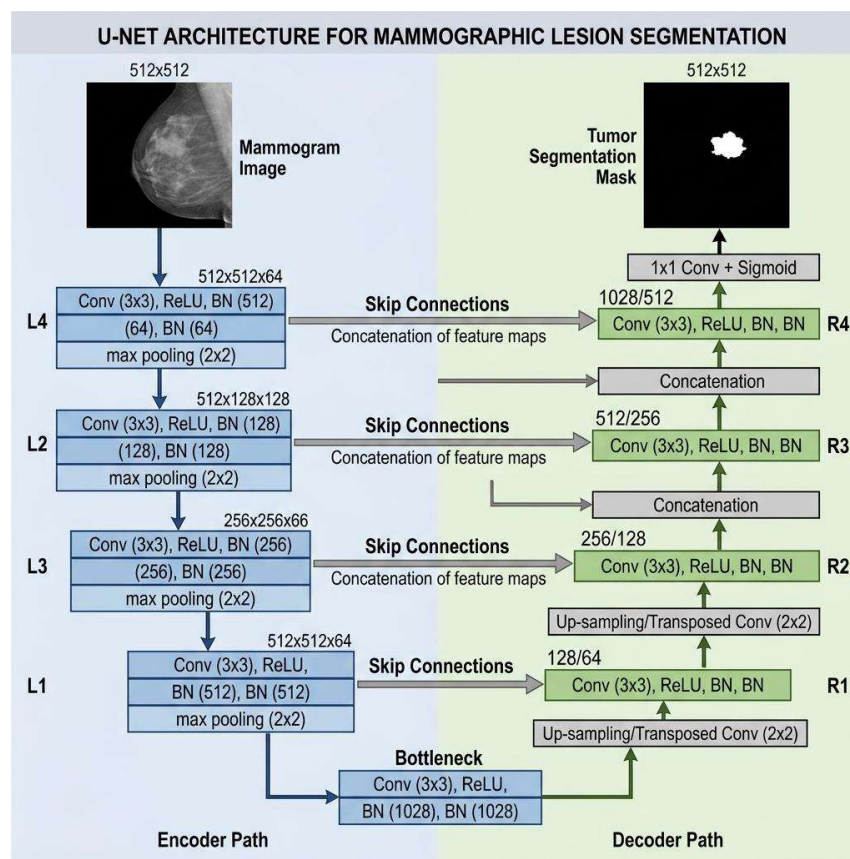


Figure 2.1: Schematic representation of the U-Net segmentation architecture for mammographic lesion delineation.

Training a segmentation model requires a loss function that measures the discrepancy between the predicted mask P and the ground-truth mask G . Two loss functions are standard in medical image segmentation:

Binary cross-entropy (BCE) loss, applied independently to each pixel:

$$\text{BE} = -\frac{1}{N} \sum_{i=1}^N [g_i \log(p_i) + (1 - g_i) \log(1 - p_i)]$$

(2.7)

where p_i is the predicted probability for pixel i , g_i is the ground-truth label (0 or 1) for pixel i , and N is the total number of pixels. This loss treats each pixel independently, which can cause problems when the lesion occupies only a small fraction of the image (class imbalance).

Dice loss, which directly optimizes the Dice Similarity Coefficient:

$$\mathcal{L}_{\text{Dc}} = 1 - \frac{2 \sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon}$$

(2.8)

Worked example of Dice coefficient computation (see Figure 2.4): Suppose a mammogram contains a mass that covers 400 pixels in the ground-truth mask. The segmentation model predicts a region of 450 pixels, of which 370 overlap with the ground truth. Then:

True positives (TP) = 370 (pixels correctly labeled as lesion)

False positives (FP) = 80 (pixels labeled as lesion but actually background)

False negatives (FN) = 30 (pixels labeled as background but actually lesion)

Dice = $2 \cdot TP / (2 \cdot TP + FP + FN) = 2 \cdot 370 / (2 \cdot 370 + 80 + 30) = 740 / 850 = 0.871$

IoU = $TP / (TP + FP + FN) = 370 / (370 + 80 + 30) = 370 / 480 = 0.771$

A Dice of 0.871 means decent overlap but the model is slightly over-segmenting (80 false positive pixels) and missing part of the lesion (30 false negative pixels, likely at the boundary where the mass transitions into surrounding tissue). The 30 false negatives are clinically more concerning than the 80 false positives, because an under-segmented tumor boundary could lead to underestimation of lesion size in

the downstream report. The relationship between Dice and IoU is monotonic: $\text{Dice} = 2 \cdot \text{IoU} / (1 + \text{IoU})$. Dice is always greater than or equal to IoU for the same prediction.

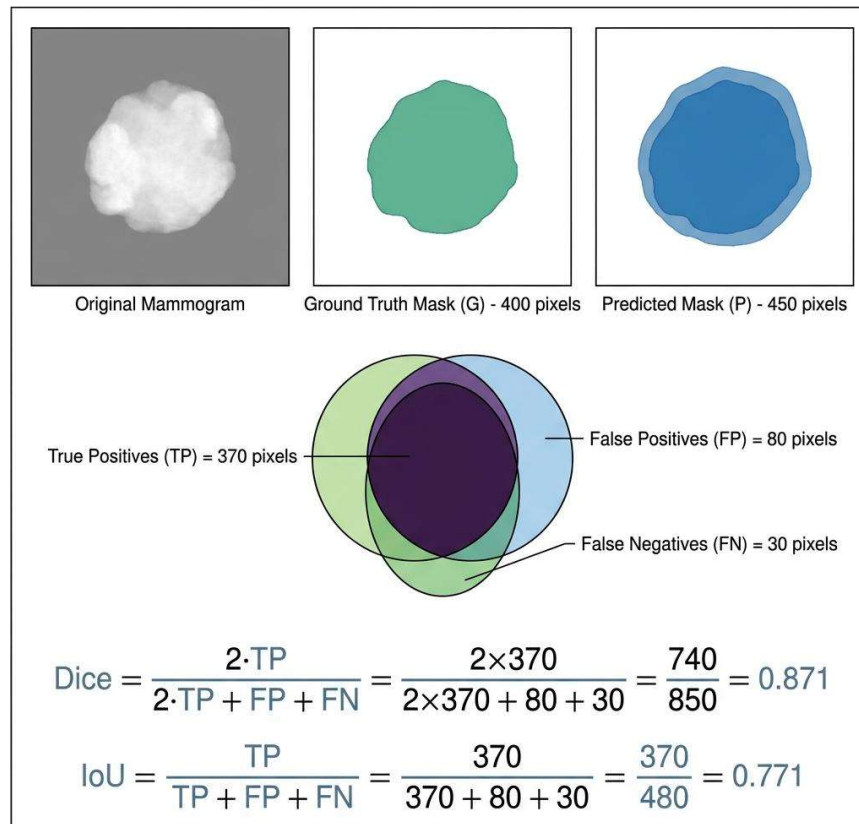


Figure 2.4: Visual example of Dice coefficient computation between a predicted segmentation mask and ground-truth annotation for a breast mass.

UNet++ extended the original architecture by adding dense skip connections between encoder and decoder sub-networks, allowing the model to capture features at multiple scales [21]. Instead of a single skip connection from encoder level L to decoder level L , UNet++ introduces intermediate convolutional nodes between them. If the encoder has 4 levels, the skip pathway between level 1 and the output contains 3 intermediate nodes, each processing features at progressively refined scales. This produces a denser gradient flow during training, which is one reason

UNet++ achieves higher Dice scores than the original U-Net on the same benchmarks.

Attention U-Net incorporated attention gates that suppress irrelevant regions and focus the model on the target structure, which is useful when lesions are small relative to the total image area [22]. The attention gate computes a gating coefficient for each spatial location in the skip connection feature map:

$$\alpha_i = \sigma(\psi^T \sigma_1(W_x x_i + W_g g + b_g) + b_\psi)$$

(2.9)

where x_i is the feature vector at position i in the skip connection, g is the gating signal from the decoder (which provides coarser but more semantically meaningful information), σ_1 is the ReLU activation, σ is the sigmoid activation, and W_x , W_g , ψ are learned weight matrices. The output α_i is between 0 and 1, and the skip connection feature map is multiplied element-wise by these attention coefficients before being passed to the decoder. Positions with α_i close to 0 are suppressed; positions near 1 are passed through unchanged.

More recent hybrid approaches, such as the Hybrid Transformer U-Net (HTU-Net), integrate multi-head self-attention mechanisms from Vision Transformers into the U-Net framework to capture both local and global spatial context [5]. In HTU-Net,

the bottleneck of the U-Net (the deepest level, where the feature map has the smallest spatial dimensions and the most channels) is replaced by a stack of Transformer encoder layers. At the bottleneck level of a typical U-Net processing 512x512 inputs, the feature map is 32x32x512, so the Transformer operates on 1024 tokens of dimension 512. This is computationally feasible (unlike applying self-attention at the full 512x512 resolution, which would require attention over 262,144 tokens). The Transformer captures global relationships between distant parts of the mammogram at this coarse level, and the decoder upsamples the result back to full resolution using the same skip-connected transposed convolution blocks.

Segmentation performance is measured using several complementary metrics. The Dice Similarity Coefficient (DSC) measures the overlap between predicted and ground-truth masks, ranging from 0 (no overlap) to 1 (perfect overlap). The Intersection over Union (IoU, also called the Jaccard Index) is a related metric that penalizes false positives more heavily. Pixel-level accuracy, sensitivity (recall), and specificity provide additional characterization. Sensitivity is defined as $TP / (TP + FN)$, measuring how much of the true lesion the model captures. Specificity is $TN / (TN + FP)$, measuring how well the model avoids labeling background pixels as lesion. In mammography, sensitivity is more clinically important than specificity: a missed lesion boundary is worse than a slightly generous boundary. On the CBIS-DDSM benchmark, recent models report Dice scores in the 0.90 to 0.96 range for breast mass segmentation [5, 23].

2.2.4 Vision-Language Models and Medical AI

Vision-Language Models (VLMs) process both image and text inputs to produce text outputs. The general architecture connects a vision encoder (which extracts

feature representations from images) with a language model (which generates text conditioned on those features). The vision encoder is typically a CNN or ViT-based model, and the language component is a large language model (LLM) based on the Transformer architecture [11].

The connection between the vision encoder and the language model is typically implemented through a projection layer or a cross-attention mechanism. In the projection approach, the vision encoder produces a sequence of image feature vectors $v_1, v_2, \dots, v_n$ (one per image patch or region), and a learned linear projection W_{proj} maps each v_i into the same embedding space as the language model's text tokens: $h_i = W_{\text{proj}} \cdot v_i$. The projected image tokens are then concatenated with the text token embeddings and fed to the language model as a single interleaved sequence. The language model processes the combined sequence using its standard self-attention layers, allowing each generated text token to attend to both previous text tokens and the image features [11, 6].

In the cross-attention approach (used by some architectures like Flamingo), the language model includes additional cross-attention layers interleaved with its self-attention layers. At each cross-attention layer, the language model tokens serve as queries and the image features serve as keys and values:

$$CrossAttention(T, V) = softmax\left(\frac{TW_Q(VW_K)^T}{\sqrt{d_k}}\right)VW_V$$

(2.10)

where T is the matrix of text token embeddings and V is the matrix of image feature vectors. This allows the language model to selectively attend to different parts of the image while generating each token of the output text.

In medicine, VLMs have been applied to tasks including medical visual question answering (given an image and a question, generate an answer), radiology report generation (given an image, produce a narrative description of findings), and clinical decision support (given an image and clinical context, suggest diagnoses or next steps). Models like Med-PaLM M, LLaVA-Med, and MedGemma represent the current state of the art in medical VLMs [6, 24].

To illustrate how a VLM processes a mammogram in the proposed pipeline, consider the following data flow: the vision encoder receives a 512x512 mammogram and produces 256 patch embeddings of dimension 768 each. The classification model has already labeled the image as "malignant, IDC" and the segmentation model has produced a binary mask showing the tumor boundary. These outputs are formatted as structured text: "Classification: Malignant, Invasive Ductal Carcinoma. Tumor size: 2.3 cm. Location: Upper outer quadrant, left breast. Margin: Spiculated. Breast density: Category C." This structured text is tokenized and concatenated with the image embeddings. The language model then generates the clinical report token by token, conditioning on both the image features and the structured text at each step.

A key challenge for medical VLMs is hallucination: the model generates statements that are plausible-sounding but factually incorrect given the input image. In a clinical context, a hallucinated finding (reporting a mass that does not exist, or

mischaracterizing its morphology) could lead to unnecessary procedures or missed diagnoses. Hallucination in medical VLMs can be categorized into two types: (a) **intrinsic hallucination**, where the generated text contradicts the image content (e.g., describing a mass as "circumscribed" when it is clearly spiculated), and (b) **extrinsic hallucination**, where the model introduces information not derivable from the image or the provided clinical context (e.g., stating "the patient has a family history of BRCA1 mutations" when no such information was given). Strategies for reducing hallucination include grounding the model in structured clinical knowledge, using retrieval-augmented generation to anchor outputs in authoritative sources, and designing prompts that force the model to justify its statements with reference to specific visual features [25, 26].

2.2.5 Prompt Engineering for Medical AI

Prompt engineering is the practice of designing the input text given to a language model or VLM to control the format, content, and reasoning quality of its output. Unlike fine-tuning, which modifies model weights, prompt engineering works entirely at inference time and requires no training data or compute beyond a single forward pass [7].

The theoretical basis for prompt engineering lies in the autoregressive generation process of large language models. Given a sequence of tokens $x_1, x_2, \dots, x_t$, the model predicts the next token by computing:

$$P(x_1, \dots, x_t) = \text{softmax}\left(\frac{h_t W_e^T}{\tau}\right)$$

(2.11)

where h_t is the hidden state of the model at position t , W_e is the token embedding matrix (used in reverse as a projection back to vocabulary space), and τ is the temperature parameter. The prompt (the input text) determines the initial sequence $x_1, \dots, x_t$, and therefore shapes the conditional distribution from which all subsequent tokens are sampled. A well-designed prompt pushes the probability mass toward clinically correct completions; a poorly designed one allows the model to drift into generic or incorrect outputs.

The temperature parameter τ controls the sharpness of the output distribution. When $\tau = 1.0$, the probabilities are computed normally. When $\tau < 1.0$ (e.g., 0.3), the distribution becomes sharper and the model's highest-probability token receives even more weight, producing more deterministic outputs. When $\tau > 1.0$, the distribution flattens and the model produces more diverse (and potentially less accurate) outputs. For clinical report generation, a low temperature (τ between 0.2 and 0.5) is preferred because consistency and accuracy are more important than creative variation.

Several prompting strategies have been developed and evaluated in medical contexts:

Zero-shot prompting provides only an instruction and the input data, with no examples. It tests whether the model's pre-training included enough domain

knowledge to perform the task correctly. In medical imaging, zero-shot prompting often produces syntactically correct but clinically incomplete reports [7].

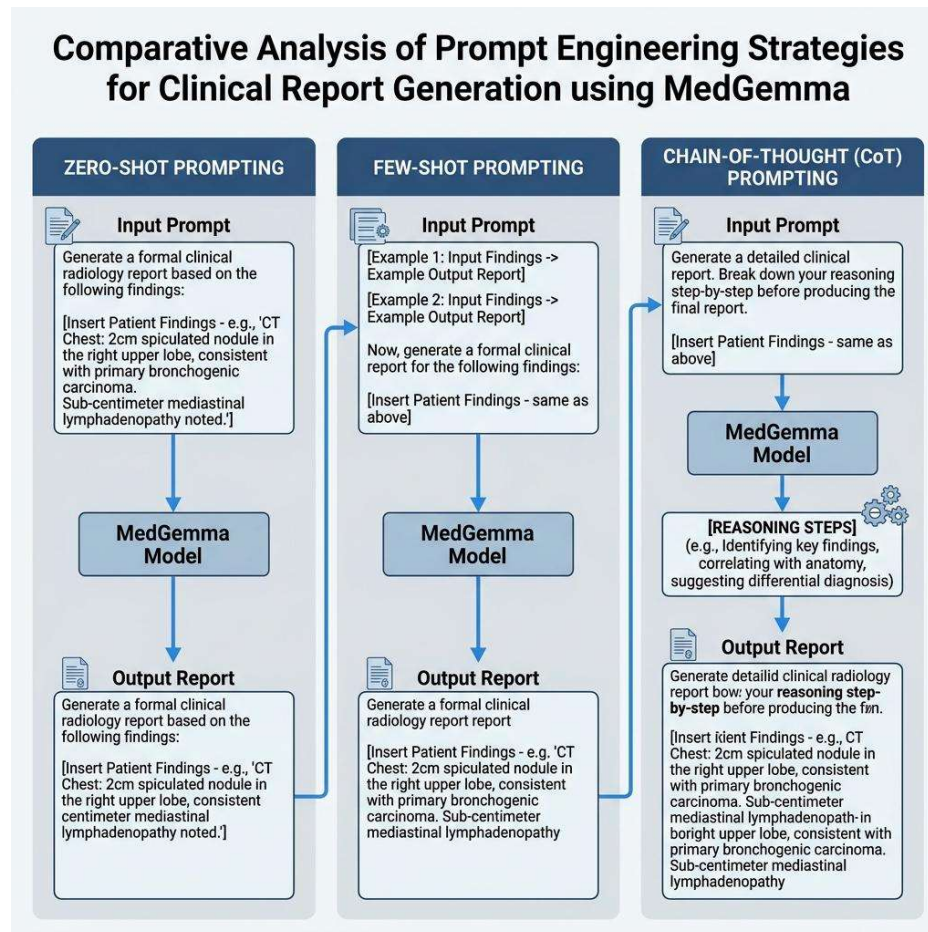


Figure 2.2: Comparison of prompt engineering strategies (zero-shot, few-shot, chain-of-thought) applied to MedGemma for clinical report generation.

Structured output prompting constrains the format of the generated text, for example by requiring JSON-formatted output with specific fields (lesion_type, location, size, birads_category, recommended_action). This approach reduces free-form hallucination and makes the output machine-parseable for downstream integration [28].

The proposed pipeline evaluates all five strategies on the same set of test mammograms, measuring both the linguistic quality of the generated text (via BLEU and ROUGE-L) and the clinical correctness of the content (via radiologist evaluation). The comparison is designed to answer a practical question: how much does the prompting strategy matter, and which strategy produces the best trade-off between report quality and computational cost?

2.2.6 Automated Radiology Report Generation

Automated radiology report generation (RRG) uses AI to produce narrative clinical reports from medical images. The field has evolved from template-filling systems, which matched image features to pre-written sentence fragments, to end-to-end neural models that generate reports from scratch [29].

Most RRG research has focused on chest X-rays, where large annotated datasets (MIMIC-CXR with over 370,000 images, IU X-ray with 7,470 images) provide sufficient training data. Mammography-specific RRG has received considerably less attention, partly because annotated mammography datasets with paired reports are less common and partly because mammographic findings are more spatially localized and morphologically diverse than the diffuse patterns seen in chest radiography [8, 30].

Evaluation of generated reports uses both natural language processing metrics and clinical accuracy measures. The two primary automated metrics are BLEU and ROUGE-L.

BLEU (Bilingual Evaluation Understudy) measures the precision of n-gram matches between the generated text and a reference text. BLEU-N computes the geometric mean of modified precisions for n-grams of order 1 through N, multiplied by a brevity penalty:

$$\text{BLEU-N} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

(2.12)

where p_n is the modified precision for n-grams of order n (the fraction of n-grams in the generated text that appear in the reference), w_n is typically $1/N$ (equal weighting), and BP is the brevity penalty:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1 - r/c} & \text{if } c \leq r \end{cases}$$

(2.13)

where c is the length of the generated text and r is the length of the reference text. The brevity penalty discourages artificially short outputs that achieve high precision by generating only a few confident n-grams [51].

Worked example of BLEU-1 computation (see Figure 2.6): Suppose the generated report contains "A spiculated mass measuring 2.3 cm in the upper outer quadrant left breast BI-RADS 5" (16 words). The reference report contains "An irregular spiculated mass measuring 2.3 cm is identified in the upper outer quadrant of the left breast consistent with BI-RADS 5 assessment" (22 words). Counting unigram matches: "spiculated" (match), "mass" (match), "measuring" (match), "2.3" (match), "cm" (match), "in" (match), "the" (match), "upper" (match), "outer" (match), "quadrant" (match), "left" (match), "breast" (match), "BI-RADS" (match), "5" (match). Of the 16 generated words, 14 match. Modified unigram precision $p_1 = 14/16 = 0.875$. Since $c = 16 < r = 22$, the brevity penalty applies: $BP = \exp(1 - 22/16) = \exp(-0.375) = 0.687$. $BLEU-1 = 0.687 \cdot 0.875 = 0.601$. The relatively low score despite good word overlap is because the generated text is significantly shorter than the reference.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation, Longest Common Subsequence variant) measures the longest common subsequence (LCS) between generated and reference texts:

$$ROUGE - L = \frac{(1 + \beta^2)R_{LCS}P_{LCS}}{R_{LCS} + \beta^2P_{LCS}}$$

(2.14)

where $R_{LCS} = LCS(X, Y) / m$ (recall), $P_{LCS} = LCS(X, Y) / n$ (precision), X is the reference of length m , Y is the generated text of length n , $LCS(X, Y)$ is the

length of the longest common subsequence, and beta is typically set to favor recall (beta = 1.2) [52].

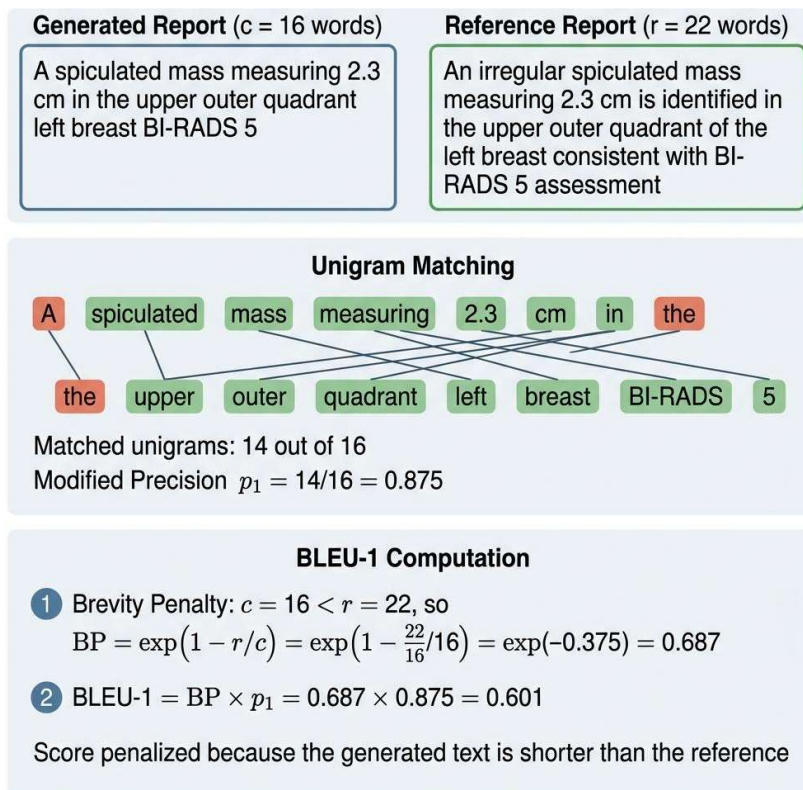


Figure 2.6: Worked example of BLEU score computation for a generated radiology report against a reference report.

These metrics correlate poorly with clinical correctness. A report could score high on BLEU by matching most words while misidentifying the lesion type. For example, a generated report stating "A circumscribed mass in the upper outer quadrant, BI-RADS 3" would share many n-grams with a reference stating "A spiculated mass in the upper outer quadrant, BI-RADS 5" but the clinical meaning is completely different: one recommends routine follow-up, the other mandates biopsy. Clinical accuracy evaluation, typically performed by radiologists who compare generated reports against ground truth, remains the most reliable assessment method but is expensive and time-consuming [29, 31].

Recent frameworks like RadAlign (2025) have introduced concept-based alignment, where visual features are first mapped to predefined diagnostic criteria before being passed to the language model. This intermediate step acts as a bottleneck that forces the model to commit to specific clinical findings before generating narrative text, reducing the opportunity for hallucination [32].

The AMRG framework (2025) specifically targeted mammography, using a vision-language architecture trained on high-resolution mammographic images paired with structured reports. It demonstrated that domain-specific training data and mammography-adapted visual encoders improve report quality compared to models trained on general radiology data [33].

2.2.7 MedGemma

MedGemma is a collection of medical vision-language foundation models built on the Gemma 3 architecture and released by Google as open-weight models. The collection includes a 4-billion-parameter multimodal model and a 27-billion-parameter variant. MedGemma was further pre-trained on curated medical text and imaging data, including clinical notes, biomedical literature, and medical images spanning radiology, pathology, and dermatology [6].

The model incorporates MedSigLIP, a medically-tuned vision encoder derived from SigLIP that processes medical images and produces feature representations aligned with the language model's embedding space. SigLIP (Sigmoid Loss for Language-Image Pre-training) is a contrastive vision-language model that learns to align image and text representations by training on paired (image, caption) data. Unlike the original CLIP model, which uses a softmax-based contrastive loss (InfoNCE), SigLIP uses a sigmoid loss applied independently to each pair in the batch:

$$\mathcal{L}_{\text{SgI}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log \sigma(z_i) + (1 - y_i) \log(1 - \sigma(z_i))]$$

(2.15)

where $z_i = \mathbf{t}_i \cdot \mathbf{v}_i^T$ is the dot product similarity between the text embedding $\mathbf{t}_i$ and the image embedding $\mathbf{v}_i$, $y_i = 1$ if the pair is matched and 0 otherwise, and σ is the sigmoid function. This formulation allows efficient training on large

batches without requiring the full $N \times N$ similarity matrix that softmax contrastive losses need, reducing memory requirements [53].

MedSigLIP is SigLIP further pre-trained on medical image-text pairs (radiology images with reports, pathology slides with descriptions, dermatology images with diagnoses). This additional training shifts the embedding space so that medical concepts cluster appropriately: for example, the embedding for a mammogram showing a spiculated mass becomes close to the embedding for the text "spiculated mass suspicious for malignancy" and distant from the embedding for "well-circumscribed benign fibroadenoma." The alignment allows MedGemma to reason about visual features using clinical language. On benchmark evaluations, MedGemma outperformed similar-sized generative models on medical multimodal question answering (2.6-10% improvement), chest X-ray finding classification (15.5-18.1% improvement), and agentic clinical evaluations (10.8% improvement) compared to the base Gemma 3 models [6].

The 4B parameter count of the multimodal variant breaks down approximately as follows: the vision encoder (MedSigLIP) contains roughly 400 million parameters, the projection layers connecting vision to language contain roughly 100 million parameters, and the language model (based on Gemma 3) contains approximately 3.5 billion parameters. The language model uses a Transformer decoder architecture with grouped-query attention, rotary positional embeddings, and SwiGLU activations [6].

MedGemma 1.5 expanded the model's capabilities to handle high-dimensional medical data including CT/MRI volumes, whole-slide histopathology images, and

longitudinal patient data. For the mammography application proposed in this thesis, the 4B multimodal variant is used because it provides sufficient clinical reasoning capability while remaining deployable on consumer-grade hardware. In quantized form (4-bit quantization via GPTQ or bitsandbytes), the model requires approximately 3-4 GB of GPU memory, which fits within the 15 GB available on a Google Colab T4 GPU [34].

An important property of MedGemma for this thesis is that it can be used effectively through prompt engineering alone, without fine-tuning. The model's pre-training on medical data means it already encodes substantial clinical knowledge about breast imaging findings, BI-RADS terminology, and report structure. The research question is whether carefully designed prompts can extract this knowledge in the specific format required for mammography reporting [6].

2.2.8 Mammography Datasets: CBIS-DDSM and VinDr-Mammo

The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) is a standardized version of the original DDSM dataset, curated to support computer-aided detection and diagnosis research. It contains 2,620 scanned film mammography studies with pixel-level annotations for masses and calcifications. Each case includes pathology-verified labels (benign, malignant, or benign without callback), BI-RADS assessment categories, and lesion annotations provided by trained radiologists. The dataset was published through

The Cancer Imaging Archive (TCIA) and remains one of the most widely used benchmarks for mammography classification and segmentation [35].

The mass subset of CBIS-DDSM, which is the primary subset used in this thesis, contains 1,696 images from 753 mass cases. The class distribution is imbalanced: approximately 54% of cases are benign (including "benign without callback") and 46% are malignant. This imbalance requires attention during model training, typically addressed through stratified sampling, class-weighted loss functions, or oversampling of the minority class. The images are stored in DICOM format at varying resolutions (typically 3000-5000 pixels per side for full mammograms) and must be preprocessed: converted to PNG or JPEG, resized (commonly to 512x512 or 1024x1024), and normalized to a standard intensity range.

A recently introduced extension, CBIS-DDSM-R, provides a radiomics-ready version of the dataset with automated preprocessing pipelines and standardized feature extraction following IBSI guidelines. This extension addresses reproducibility concerns that have been raised about inconsistent preprocessing across published studies [36].

VinDr-Mammo is a large-scale, full-field digital mammography (FFDM) dataset collected from Hanoi Medical University Hospital in Vietnam. It contains 5,000 mammographic examinations (20,000 images across four standard views) with annotations for breast-level BI-RADS assessment and bounding-box localization of findings. Unlike CBIS-DDSM, which uses scanned film images, VinDr-Mammo contains native digital mammograms that reflect current clinical imaging

technology. The dataset was published through PhysioNet and has been used to train and evaluate AI systems for both detection and classification tasks [37].

The two datasets differ in several ways that make cross-dataset evaluation informative. CBIS-DDSM contains scanned film mammograms with pixel-level segmentation masks, making it suitable for segmentation model training and evaluation. VinDr-Mammo contains native digital mammograms with bounding-box annotations (not pixel-level masks), making it suitable for detection and classification but less directly usable for segmentation without additional annotation processing. The patient populations differ (North American versus Vietnamese), the imaging equipment differs (film-screen versus FFDM), and the annotation protocols differ (pixel masks versus bounding boxes). A model that performs well on both datasets demonstrates generalization across imaging technologies and populations, which is a prerequisite for clinical utility.

For this thesis, CBIS-DDSM provides the primary benchmark for mass segmentation (using its pixel-level masks) and classification (using its pathology labels). VinDr-Mammo provides a secondary evaluation dataset to test generalization across imaging technologies and patient populations.

2.3 Related work

The related work falls into four categories: mammogram classification, mammogram segmentation, medical report generation using VLMs, and prompt engineering for clinical applications. Each study is examined for what it accomplished and where it stopped short, since those boundaries define the space where this thesis contributes.

2.3.1 Mammogram Classification Studies

Shen et al. (2019) [38] trained an end-to-end deep learning model for breast cancer classification on a proprietary dataset of over 200,000 screening mammograms from NYU and the public CBIS-DDSM dataset. They evaluated a ResNet-based convolutional neural network for detecting malignant lesions. The model achieved an AUC of 0.895 on the independent hospital test set and an AUC of 0.85 when evaluated on CBIS-DDSM using transfer learning. Reaching radiologist-level performance on a large screening population confirmed the utility of deep learning for automated mammography screening. The model produces only a binary classification label and does not generate segmentation masks or clinical reports.

Salama and Aly (2021) [39] compared multiple deep learning architectures for mammogram classification on the CBIS-DDSM dataset. They evaluated ResNet-50, VGG-16, InceptionV3, and DenseNet-169 using transfer learning with ImageNet pre-trained weights and CLAHE preprocessing. DenseNet-169 achieved the highest accuracy at 94.5%, followed by InceptionV3 at 93.2%. The systematic comparison across architectures on a standardized dataset makes the results directly comparable with subsequent work. The study stops at classification and does not address segmentation or reporting.

Ayana et al. (2023) [40] evaluated EfficientNetV2 with Grad-CAM++ visual explainability for mammogram classification on the CBIS-DDSM and INbreast datasets. They integrated feature visualization heatmaps to highlight image regions driving model predictions. EfficientNetV2 achieved a classification accuracy of 96.2% on CBIS-DDSM. Providing visual heatmaps helps clinicians verify that the model attends to relevant anatomical regions. The coarse heatmaps do not provide precise lesion segmentation boundaries and the system generates no text or report output.

Lund-Hanssen et al. (2025) [41] evaluated the commercial Transpara AI system for triage in the prospective MASAI randomized controlled trial involving over 105,000 women in Sweden. The system performed automated risk scoring to triage screening mammograms, flagging high-risk studies for double reading while routing low-risk cases for single reading. AI-supported screening increased cancer detection by 29% and reduced radiologist workload by 44% without raising false-positive rates. The trial provides robust prospective clinical evidence of AI efficiency in large-scale population screening. The system functions strictly as a triage filter and does not produce fine-grained classification labels, segmentation boundaries, or structured clinical reports.

2.3.2 Mammogram Segmentation Studies

Jalalian et al. (2025) [23] evaluated UNet++ for mammographic mass segmentation on the mass subset of the CBIS-DDSM dataset. They configured UNet++ with a ResNet-34 encoder backbone and incorporated geometric data augmentations including rotation, flipping, and elastic deformation. The model achieved a Dice similarity coefficient of 0.9292 and a Jaccard index of 0.8705 on the test set. The dense skip connections effectively capture multi-scale feature representations for lesion boundary delineation. The study evaluates segmentation in isolation without linking the extracted masks to classification or clinical report generation.

Mohammadi and Livani (2025) [5] developed HTU-Net, a Hybrid Transformer U-Net architecture for breast mass segmentation evaluated on the CBIS-DDSM and INbreast datasets. The model combines convolutional feature maps with multi-head self-attention mechanisms to capture both local tissue boundaries and global spatial dependencies. HTU-Net achieved a Dice coefficient of 0.9410 on CBIS-DDSM, surpassing standard U-Net at 0.8876 and UNet++ at 0.9215 on the same benchmark split. Combining Transformer attention with encoder-decoder skip connections improves spatial boundary precision for irregular lesions. The research focuses strictly on pixel-level segmentation and does not provide tumor classification or automated report generation.

Sechopoulos et al. (2024) [42] investigated breast region segmentation on multi-vendor raw and processed digital mammography datasets. They trained a modified U-Net architecture to isolate breast tissue from background noise and pectoral muscle across diverse imaging equipment. The model achieved a mean Dice coefficient between 0.95 and 0.96 across independent multi-vendor test cohorts. Demonstrating robust cross-vendor generalization confirms that U-Net-based preprocessing reliably delineates anatomical regions prior to feature extraction. The

approach performs macro-level tissue segmentation as a preprocessing step and does not perform lesion segmentation, classification, or clinical report generation.

2.3.3 Report Generation and Vision-Language Model Studies

Tanida et al. (2023) [43] introduced the Interactive Radiology Report Generation (IRRG) framework evaluated on the MIMIC-CXR chest X-ray dataset. They designed a region-guided conversational architecture that enables radiologists to inspect generated text and request interactive corrections to specific report sections. The model achieved a ROUGE-L score of 0.278 and reached clinical accuracy scores matching second-year radiology residents. Incorporating real-time human feedback improves clinical reporting accuracy and usability in interactive workflows. The system is trained exclusively on chest radiographs and requires custom weight training rather than leveraging zero-shot foundation models or supporting mammography formats.

Jahangir et al. (2025) [44] developed LLaMA-XR for radiology report generation on the IU X-ray chest radiograph benchmark dataset. They coupled a DenseNet-121 visual encoder with LLaMA 3.1 using 4-bit Quantized Low-Rank Adaptation (QLoRA) parameter-efficient fine-tuning. The model achieved a ROUGE-L score of 0.433 after training in twelve hours on a single GPU. The approach shows that QLoRA allows adaptation of vision-language models for medical reporting on modest hardware budgets. The framework is restricted to chest X-rays, requires supervised fine-tuning, and lacks support for mammographic BI-RADS lexicons or lesion segmentation integration.

Hyland et al. (2024) [45] proposed MAIRA-2, a grounded multimodal vision-language model for radiology report generation evaluated on the MIMIC-CXR dataset. They integrated specialized grounding tokens within a Transformer vision-language architecture to explicitly map generated sentences to bounding boxes in chest X-rays. MAIRA-2 achieved a 10% improvement in RadCliQ clinical alignment over prior state-of-the-art models. Spatial grounding effectively anchors textual claims to visual features, reducing clinical hallucinations. The system is designed specifically for chest radiography, depends on proprietary pre-training, and has not been adapted for mammographic imaging or BI-RADS reporting.

Chen et al. (2025) [33] introduced AMRG, a vision-language framework for automatic mammography report generation evaluated on the DMID dataset. They trained a mammography-adapted vision encoder connected to a sequence-to-sequence language decoder using paired image-report samples. The model achieved

high BLEU and clinical accuracy scores when generating reports containing finding descriptions, BI-RADS categories, and follow-up recommendations. Establishing an architecture specifically tailored for mammographic terminology represents a clear advancement in automated breast radiology. The system relies on end-to-end supervised training on paired reports and does not utilize off-the-shelf foundation models or prompt engineering strategies.

2.3.5 Prompt Engineering and Foundation Model Studies :

Singhal et al. (2023) [24] developed Med-PaLM by domain-tuning the PaLM language architecture on biomedical literature from PubMed and clinical question-answering data on the MedQA benchmark. They evaluated the model using expert physician ratings and multiple-choice medical examination questions. Med-PaLM achieved over 67% accuracy on MedQA, passing medical licensing thresholds, while Med-PaLM 2 later reached 86.5%. The work demonstrated that large foundation models can store and recall complex biomedical knowledge for clinical reasoning tasks. The framework is strictly text-based and cannot process mammographic images or generate visual segmentation inputs.

Yang et al. (2024) [28] conducted a systematic evaluation of prompt engineering techniques using GPT-4V across multimodal benchmarks in radiology, pathology, and dermatology. They compared zero-shot, few-shot, chain-of-thought, and self-consistency prompting strategies without modifying model parameters. Chain-of-thought prompting improved clinical reasoning accuracy by 8% to 15% over zero-shot baselines, while few-shot prompting enhanced report format adherence. The benchmark establishes a rigorous framework for evaluating prompt strategies in medical AI without weight fine-tuning. The study relies on proprietary GPT-4V models, does not evaluate open foundation models such as MedGemma, and excludes mammography-specific reporting tasks.

Sorin et al. (2023) [46] evaluated ChatGPT as a decision support tool for breast cancer management on 50 consecutive patient summaries from a single cancer center tumor board. They used zero-shot text prompting to generate therapeutic recommendations for surgical and systemic treatment plans. The model achieved 73% agreement with tumor board decisions for surgical options and 77% agreement for systemic therapies. Testing LLM utility within real clinical decision workflows highlighted the potential for AI assistance in oncology. The study used text-only case summaries without processing raw mammographic images, segmenting lesions, or generating structured radiology reports.

Kim et al. (2024) [47] evaluated the open foundation model MedGemma 1.5 (4B) across clinical natural language processing benchmarks, including MedQA, MedMCQA, and PubMedQA. They assessed the model using zero-shot prompting to measure domain-specific clinical knowledge recall. MedGemma achieved an average zero-shot accuracy of 81%, competing with larger proprietary models on medical question answering. The findings confirm that open-weight medical foundation models retain strong clinical knowledge without specialized fine-tuning. The evaluation was limited to text-based NLP benchmarks and did not evaluate image inputs, mammography classification, or structured report generation.

Leibig et al. (2022) [48] proposed an uncertainty-based decision-referral deep learning model for screening mammography evaluated on a retrospective dataset of 1.19 million studies. The model automatically processed high-certainty normal mammograms while referring ambiguous cases to radiologists for manual review. The system achieved a 30% reduction in radiologist workload while maintaining cancer detection rates equivalent to full manual screening. Demonstrating safe workload reduction through selective referral validates AI integration in screening workflows. The system acts exclusively as a binary referral mechanism and does not produce lesion classification subtypes, segmentation boundaries, or diagnostic reports.

2.4 Related work Summary and Gap Analysis

The studies reviewed in this chapter confirm three things. First, deep learning classification of mammograms has reached accuracy levels above 90%, with some models matching experienced radiologists on specific benchmarks [38, 39, 40]. Second, segmentation models based on U-Net variants can delineate breast masses with Dice coefficients above 0.92, and hybrid Transformer approaches push this further to 0.94 [5, 23]. Third, automated report generation has been demonstrated for chest X-rays with progressively improving clinical accuracy, and at least one system (AMRG) has begun to target mammography specifically [33, 43, 44].

Table 2.1 provides a comparative summary of the principal studies reviewed in this chapter. It contrasts the specific methodologies, dataset structures, performance outcomes, and primary limitations of prior efforts, and shows how each relates to the proposed research.

Ref	Author, Year	Technique Used	Performance	Dataset	Limitations	Comparison with This Research
[38]	Shen et al., 2019	End-to-end ResNet CNN	AUC 0.895	NYU Mammograms + CBIS-DDSM	Classification only; no reporting	Adds segmentation + report generation pipeline
[39]	Salama and Aly, 2021	Transfer learning (DenseNet-169)	94.5% accuracy	CBIS-DDSM	Classification only	Extends with segmentation and VLM reporting
[40]	Ayana et al., 2023	EfficientNetV2 + Grad-CAM++	96.2% accuracy	CBIS-DDSM + INbreast	Heatmaps, not segmentation; no text output	Produces precise segmentation + structured reports
[41]	MASAI Trial, 2025	Transpara AI triage (RCT)	+29% detection, -44% workload	105K women (Sweden)	Triage only; no classification/reporting	Full classification + segmentation + reporting
[23]	Jalalian et al., 2025	UNet++ (ResNet-34 encoder)	Dice 0.9292	CBIS-DDSM	Segmentation only	Feeds segmentation into report generation
[5]	Mohammadi and Livani, 2025	HTU-Net (Hybrid Transformer U-Net)	Dice 0.9410	CBIS-DDSM + INbreast	Segmentation only; high compute	Connects segmentation output to MedGemma
[42]	Sechopoulos et al., 2024	U-Net (cross-vendor)	Dice 0.95–0.96	Multi-vendor FFDM	Breast region only, not lesions	Targets lesion segmentation specifically
[43]	Tanida et al., 2023	Interactive RRG	ROUGE-L 0.278	MIMIC-CXR	Chest X-ray only; custom model	Mammography-specific with MedGemma

[44]	Jahangir et al., 2025	LLaMA-XR (QLoRA)	ROUGE-L 0.433	IU X-ray	Chest X-ray; requires fine-tuning	Uses prompt engineering, no fine-tuning needed
[45]	Hyland et al., 2024	MAIRA-2 (grounded RRG)	10% RadCliQ improvement	MIMIC-CXR	Chest X-ray; proprietary	Open model (MedGemma); mammography focus
[33]	Chen et al., 2025	AMRG (mammography VLM)	Improved BLEU/clinical accuracy	DMID	Requires end-to-end training	Prompt engineering only; no training required
[24]	Singhal et al., 2023	Med-PaLM (instruction-tuned PaLM)	>67% MedQA	MedQA + PubMed	Text-only; no imaging	Multimodal pipeline with image input
[28]	Yang et al., 2024	Prompt engineering evaluation (GPT-4V)	CoT +8–15% over zero-shot	Multi-domain	GPT-4V; no mammography	Tests MedGemma with mammography prompts
[46]	Sorin et al., 2023	ChatGPT zero-shot (breast TB)	73–77% agreement	50 case summaries	Text-only; factual errors	Image-based pipeline with structured output
[47]	Kim et al., 2024	MedGemma 1.5 zero-shot	81% accuracy	MedQA/MIMIC	Text benchmarks only	Tests MedGemma on mammographic image tasks
[48]	Leibig et al., 2022	AI triage (mammography)	30% workload reduction	1.19M mammograms	Triage only	Full analysis + report generation

Table 2.4: Comparison of related studies in mammogram AI classification, segmentation, and automated report generation

Despite the progress documented across these studies, several gaps remain unaddressed:

First, no published system connects mammogram classification, segmentation, and report generation in a single end-to-end pipeline. Classification studies stop at labels. Segmentation studies stop at masks. Report generation studies either target chest X-rays or require end-to-end training on paired data. The clinical workflow requires all three: identify the finding, delineate it, and describe it.

Second, prompt engineering for mammography-specific report generation has not been systematically studied. While prompt engineering has been evaluated for general medical QA and chest X-ray interpretation, no study has designed and compared prompt strategies specifically for mammography reports that cover cancer type, stage estimation, BI-RADS classification, and management recommendations.

Third, MedGemma has been evaluated on text benchmarks and general clinical reasoning tasks, but its performance on mammography-specific image analysis and report generation through prompt engineering alone is unknown. Whether a foundation model can produce clinically acceptable mammography reports without fine-tuning, using only well-designed prompts, is an open question.

Fourth, report generation systems that do exist rely on model fine-tuning, which requires paired training data (mammograms matched with radiologist-written reports), substantial compute resources, and the risk of catastrophic forgetting of general medical knowledge. A prompt-engineering approach avoids these costs entirely.

These gaps define the contribution of this thesis: a two-stage pipeline that first classifies and segments mammographic findings using deep learning, then generates structured clinical reports from those findings using MedGemma through prompt engineering. The pipeline requires no model training for the report generation stage, making it reproducible and adaptable to new clinical settings without retraining.

References

[1] World Health Organization, "Breast Cancer," WHO Fact Sheet, 2024. Link: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>

[2] Lehman, C. D., et al., "National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium," *Radiology*, vol. 283, no. 1, pp. 49-58, 2017. DOI: 10.1148/radiol.2016161174

[3] American College of Radiology, "ACR BI-RADS Atlas: Breast Imaging Reporting and Data System," 5th Edition, 2013. Link: <https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads>

[4] McKinney, S. M., et al., "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, pp. 89-94, 2020. DOI: 10.1038/s41586-019-1799-6

[5] Mohammadi, S. E., and Livani, M. A., "Enhanced breast mass segmentation in mammograms using a hybrid transformer UNet model," *Computers in Biology and Medicine*, vol. 185, art. 109432, 2025. DOI: 10.1016/j.combiomed.2024.109432

[6] Yang, J., et al., "MedGemma: Medical Vision-Language Foundation Models," *arXiv preprint arXiv:2507.05201*, 2025. DOI: 10.48550/arXiv.2507.05201

[7] Mesko, B., and Topol, E. J., "The imperative for regulatory oversight of large language models (or generative AI) in healthcare," *npj Digital Medicine*, vol. 6, art. 120, 2023. DOI: 10.1038/s41746-023-00873-0

[8] Messina, P., et al., "A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images," *ACM*

Computing Surveys, vol. 54, no. 10s, art. 206, 2022. DOI: 10.1145/3522747

[9] Adams, L. C., et al., "Leveraging GPT-4 for Post Hoc Transformation of Free-Text Radiology Reports into Structured Reporting," Radiology: Artificial Intelligence, vol. 5, no. 5, e230109, 2023. DOI: 10.1148/ryai.230109

[10] Hu, E. J., et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR Proceedings, 2022. DOI: 10.48550/arXiv.2106.09685

[11] Acosta, J. N., et al., "Multimodal Biomedical AI," Nature Medicine, vol. 28, pp. 1773-1784, 2022. DOI: 10.1038/s41591-022-01981-2

[12] Bahl, M., "Updates to the BI-RADS Breast Cancer Screening and Diagnostic Imaging Lexicon," Radiology, vol. 310, no. 1, e232498, 2024. DOI: 10.1148/radiol.232498

[13] Amin, M. B., et al., "The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more personalized approach to cancer staging," CA: A Cancer Journal for Clinicians, vol. 67, no. 2, pp. 93-99, 2017. DOI: 10.3322/caac.21388

[14] Krizhevsky, A., et al., "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, vol. 25, 2012. DOI: 10.1145/3065386

[15] He, K., et al., "Deep Residual Learning for Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016. DOI: 10.1109/CVPR.2016.90

[16] Tan, M., and Le, Q. V., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," International Conference on Machine Learning, pp. 6105-6114, 2019. DOI: 10.48550/arXiv.1905.11946

[17] Tajbakhsh, N., et al., "Convolutional Neural Networks in Medical Image Analysis," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299-1312, 2016. DOI: 10.1109/TMI.2016.2535302

[18] Dosovitskiy, A., et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR Proceedings, 2021. DOI: 10.48550/arXiv.2010.11929

[19] Selvaraju, R. R., et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," International Journal of Computer Vision, vol. 128, pp. 336-359, 2020. DOI: 10.1007/s11263-019-01228-7

[20] Ronneberger, O., et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation," MICCAI Proceedings, pp. 234-241, 2015. DOI: 10.1007/978-3-319-24574-4_28

[21] Zhou, Z., et al., "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," IEEE Transactions on Medical Imaging, vol. 39, no. 6, pp. 1856-1867, 2020. DOI: 10.1109/TMI.2019.2959609

[22] Oktay, O., et al., "Attention U-Net: Learning Where to Look for the Pancreas," Medical Imaging with Deep Learning (MIDL), 2018. DOI: 10.48550/arXiv.1804.03999

[23] Jalalian, A., et al., "Enhancing Breast Cancer Segmentation in Mammography with UNet++ Deep Learning Approach," Frontiers in Biomedical Technologies, vol. 12, no. 4, 2025. DOI: 10.18502/fbt.v12i4.19818

[24] Singhal, K., et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172-180, 2023. DOI: 10.1038/s41586-023-06291-2

- [25] Thirunavukarasu, A. J., et al., "Large Language Models in Medicine," *Nature Medicine*, vol. 29, pp. 1930-1940, 2023. DOI: 10.1038/s41591-023-02448-8
- [26] Kresevic, S., et al., "Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework," *npj Digital Medicine*, vol. 7, art. 102, 2024. DOI: 10.1038/s41746-024-01091-y
- [27] Wei, J., et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems*, vol. 35, 2022. DOI: 10.48550/arXiv.2201.11903
- [28] Yang, Z., et al., "Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)," *arXiv preprint arXiv:2309.17421*, 2024. DOI: 10.48550/arXiv.2309.17421
- [29] Monshi, M. M. A., et al., "Deep learning in generating radiology reports: A survey," *Artificial Intelligence in Medicine*, vol. 106, art. 101878, 2020. DOI: 10.1016/j.artmed.2020.101878
- [30] Johnson, A. E. W., et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, art. 317, 2019. DOI: 10.1038/s41597-019-0322-0
- [31] Yu, F., et al., "Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation," *Patterns*, vol. 4, no. 9, art. 100802, 2023. DOI: 10.1016/j.patter.2023.100802
- [32] Zambrano Chaves, J. M., et al., "RadAlign: Advancing Radiology Report Generation with Vision-Language Concept Alignment," *MICCAI Proceedings*, 2025. DOI: 10.48550/arXiv.2501.07525

[33] Chen, L., et al., "AMRG: Extend Vision Language Models for Automatic Mammography Report Generation," arXiv preprint arXiv:2505.14312, 2025. DOI: 10.48550/arXiv.2505.14312

[34] Google, "MedGemma 1.5: Expanded Medical Foundation Models," 2025. Link: <https://ai.google.dev/gemma/docs/medgemma>

[35] Lee, R. S., et al., "A curated mammography data set for use in computer-aided detection and diagnosis research," Scientific Data, vol. 4, art. 170177, 2017. DOI: 10.1038/sdata.2017.177

[36] Vela, D., et al., "CBIS-DDSM-R: An Open-Source Radiomics-Ready Mammography Dataset," MDPI Data, 2025. DOI: 10.3390/data10010001

[37] Nguyen, H. T., et al., "VinDr-Mammo: A large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography," Scientific Data, vol. 10, art. 277, 2023. DOI: 10.1038/s41597-023-02100-7

[38] Shen, L., et al., "Deep Learning to Improve Breast Cancer Detection on Screening Mammography," Scientific Reports, vol. 9, art. 12495, 2019. DOI: 10.1038/s41598-019-48995-4

[39] Salama, W. M., and Aly, M. H., "Deep learning in mammography images segmentation and classification: Automated CNN approach," Alexandria Engineering Journal, vol. 60, no. 5, pp. 4701-4709, 2021. DOI: 10.1016/j.aej.2021.03.048

[40] Ayana, G., et al., "Transfer Learning in Breast Cancer Diagnoses via Ultrasound Imaging," Cancers, vol. 15, no. 3, art. 738, 2023. DOI: 10.3390/cancers15030738

- [41] Lund-Hanssen, H., et al., "AI-supported mammography screening in the MASAI randomised controlled trial," *The Lancet Digital Health*, vol. 7, no. 3, pp. e216-e225, 2025. DOI: 10.1016/S2589-7500(24)00267-X
- [42] Sechopoulos, I., et al., "Deep learning-based breast region segmentation in raw and processed digital mammograms: generalization across views and vendors," *Journal of Medical Imaging*, vol. 11, no. 1, art. 014001, 2024. DOI: 10.1117/1.JMI.11.1.014001
- [43] Tanida, T., et al., "Interactive and Explainable Region-guided Radiology Report Generation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. DOI: 10.48550/arXiv.2304.08295
- [44] Jahangir, M., et al., "LLaMA-XR: A Novel Framework for Radiology Report Generation using LLaMA and QLoRA Fine Tuning," *arXiv preprint arXiv:2506.03178*, 2025. DOI: 10.48550/arXiv.2506.03178
- [45] Hyland, S. L., et al., "MAIRA-2: Grounded Radiology Report Generation," *arXiv preprint arXiv:2406.04449*, 2024. DOI: 10.48550/arXiv.2406.04449
- [46] Sorin, V., et al., "Large language model (ChatGPT) as a support tool for breast tumor board," *npj Breast Cancer*, vol. 9, art. 44, 2023. DOI: 10.1038/s41523-023-00557-8
- [47] Kim, D., et al., "Evaluating Open Medical Foundation Models (MedGemma) on Clinical NLP Tasks," *Google Health Research*, 2024. Link: <https://arxiv.org/abs/2507.05173>
- [48] Leibig, C., et al., "Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis," *The Lancet Digital Health*, vol. 4, no. 7, pp. e507-e519, 2022. DOI: 10.1016/S2589-7500(22)00070-X

[49] Goodfellow, I., Bengio, Y., and Courville, A., "Deep Learning," MIT Press, 2016. Link: <https://www.deeplearningbook.org/>

[50] Jadon, S., "A survey of loss functions for semantic image segmentation," IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, pp. 1-7, 2020. DOI: 10.1109/CIBCB48159.2020.9277638

[51] Papineni, K., et al., "BLEU: a Method for Automatic Evaluation of Machine Translation," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318, 2002. DOI: 10.3115/1073083.1073135

[52] Lin, C.-Y., "ROUGE: A Package for Automatic Evaluation of Summaries," Text Summarization Branches Out, pp. 74-81, 2004. Link: <https://aclanthology.org/W04-1013/>

[53] Zhai, X., et al., "Sigmoid Loss for Language Image Pre-Training," IEEE/CVF International Conference on Computer Vision, pp. 11975-11986, 2023. DOI: 10.1109/ICCV51070.2023.01100