

فارس عبدالغني

Breast Cancer Mammogram Classification and Segmentation using artificial intelligence-V9.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:406639751

Submission Date

Aug 10, 2026, 10:25 PM GMT+3

Download Date

Aug 14, 2026, 5:24 PM GMT+3

File Name

Breast Cancer Mammogram Classification and Segmentation using artificial intelligence-V9.docx

File Size

4.8 MB

59 Pages

11,481 Words

67,574 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

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.

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 using artificial intelligence

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 Eissa

College of Informatics, Midocean University, Comoros

Acknowledgements

We would like to thank our supervisor, Dr. Mohamed Eissa, for his continuous support and valuable feedback for this research. His expertise in artificial intelligence studies has influenced the direction of our work. His comments challenged us to think more wider and approach the research from a deeper perspective .

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

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

Abstract

Breast cancer is the common frequently diagnosed cancer among women globally, and the mammography screening is known as the primary method for detection for it. While deep learning models have reached high accuracy on mammogram classification and tumor segmentation tasks, the clinical workflow still requires a radiologist to process imaging outputs, determine the cancer type and criticality, and write a diagnostic report. These steps are consuming time and limited to inter-observer variability.

This thesis proposes a two stages that automates the full path from raw mammogram images input to structured clinical report output. At the first stage, we will use deep learning model that classifies mammographic findings (benign versus malignant, cancer subtype) and segments tumor regions found in the image. At the second stage, the classification labelling, segmentation masking, and extracting the imaging features are parsed as structured input to MedGemma, a medical vision language foundation model developed by Google. Using prompt engineering, including system prompting, few-shot examples, and chain-of-thought reasoning, MedGemma can generate a clinical report that determine the cancer type, approximate stage based on imaging features, lesion morphology, and recommended follow up actions. The system is evaluated on public published mammography datasets, including the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) and VinDr-Mammo. The classification and segmentation performance is metered using standard methods (accuracy, F1-score, Dice coefficient, IoU). Report quality is measured using both automated metrics (BLEU, ROUGE-L) and clinical accuracy evaluation versus radiologist written reports. The main aim of this research are to build the classification and segmentation pipeline, to develop prompt engineering methodologies for MedGemma-based report generation, and to utilize the clinical usefulness of the generated reports. This work adapts end to end framework that links imaging analysis to natural language reporting without model fine tuning, relying on carefully designed prompts to extract clinical reasoning from 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 common cancer worldwide. The World Health Organization reported that 2.3 million new breast cancer cases were diagnosed in 2022, making regular screening via mammograms a good and proven method of reducing mortality from this disease [1]. Diagnostic mammograms have their ability to detect cancer before any clinical signs appear, but correct readings of mammographic images and writing the reports rely on radiologists who remain in short supply within healthcare systems.

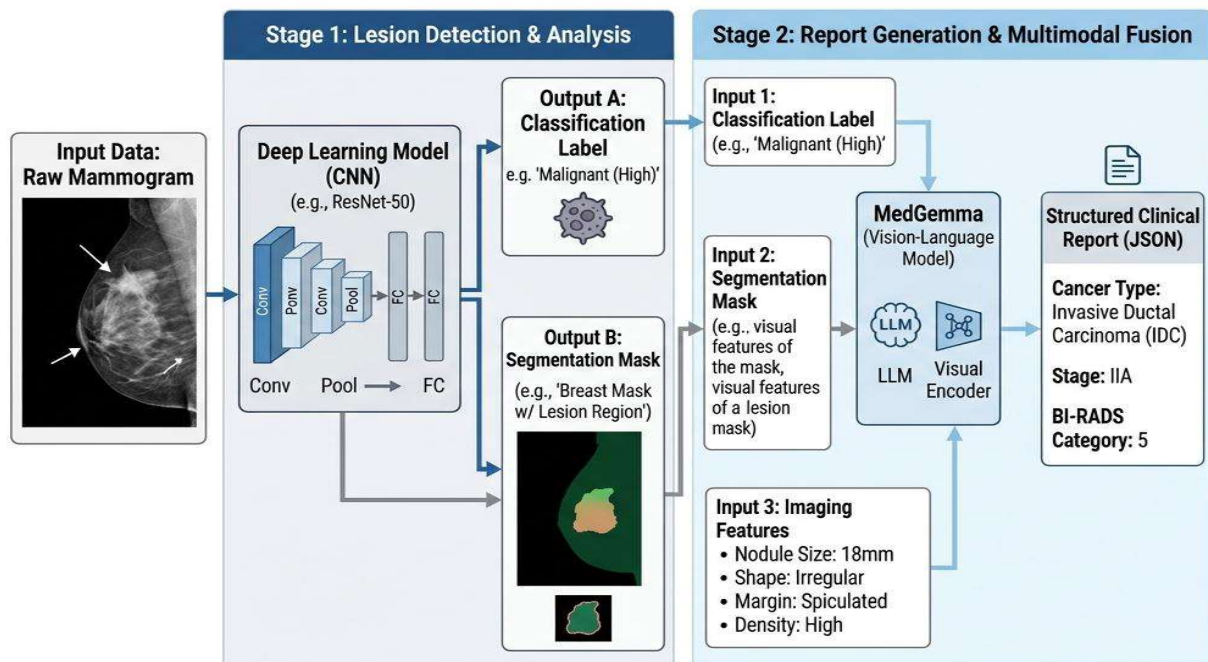
According to studies, radiologists interpret around 40 to 100 mammograms on average per day, with the concordance rate between different radiologists typically being 60%-80% depending on the complexity of the lesion [2, 3].

In the last ten years, the advent of deep learning technology has changed the way images in clinical practice are analyzed. Convolutional neural networks and their modifications have been shown to be able to detect whether a mammogram is of benign or malignant origin with over 90% of accuracy in standardized datasets [4]. Moreover, image segmentation models based on the U-Net architecture have been created, which showed great results in the tumor segmentation task, measured by the Dice coefficient which exceeds 0.90 [5]. Nevertheless, the advances in this area can only reduce a small part of the workload involved in the diagnostic process. Thus, the classification label or the segmentation mask can't substitute full clinical reports, which remain to be prepared by the radiologist. The tasks assigned to the physician include the interpretation of the results, assigning the BI-RADS category, determining the cancer stage based on available features, revealing the type of cancer if possible, and giving recommendations with regards to how to address the findings.

The developments in vision-language models (VLMs) have created a new trend in the field. For instance, Google's MedGemma which has been built on the basis of the large amount of data from medical imaging together with textual information has shown its capacity for processing visual and textual data [6]. Rather than hone either model, which utilizes using heavy computing resources and preparing a vast amount of annotated data, the idea of prompt engineering may be used in order to alter model behavior in a more light manner than the fine-tuning of the models [7].

This work suggests an approach which consists of two stages. The first stage passes raw mammographic images via a deep learning model to classify and segment them. The second stage uses the results of image classification and segmentation as well as the features that were extracted in the previous stages and writes a clinical report using them as input to MedGemma with the help of prompt engineering. The report includes cancer type, estimated stage based on the available imaging features, lesion properties, and recommendations on actions to take. The proposed approach is expected to automate the way from the original mammogram through to the clinical report, thus relieving radiologists from the need to create numerous reports and making the reporting process more consistent. [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

Breast cancer imaging AI research is currently aimed at handling the different subtasks completely independently of each other. For example, classification models only classify the mammogram as either benign or malignant, while

segmentation models only provide the outline of the lesion. In terms of reporting generation, even if there is such a model, it was developed primarily for chest X-rays and it is not compatible with the mammography terminology or BI-RADS formatting [8]. No available system allows for the connection of all the mentioned procedures within one pipeline that starts from the mammogram and finishes off in the creation of a structured report.

The research outlined in this thesis is defined by three basic problems.

First, the problem of linking clinical communication and image analysis. For example, when it is said that the segmentation mask shows a 2.3cm spiculated mass in the upper outer quadrant of the left breast, there should be an additional narrative that states which irregularity is evident and suggests a response. However, the available AI models stop at the segmentation stage.

Second, the fact that there are no available prompt engineering strategies that apply to the creation of reports for mammograms. While prompt engineering has been investigated for chest X-ray results interpretation as well as responding to routine medical questions, prompt engineering specifically designed for mammography report creation has not been studied systematically [7] [9].

Third, the fact that fine-tuning of models is the only method applied. Fine-tuning can involve the preparation of datasets, GPU usage, and risks of catastrophic forgetting. Prompt engineering eliminates all these costs since it is iterative and does not require retraining. However, the efficacy of prompt engineering as a viable option in creating valid clinical reports remains to be seen [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

The present study makes contributions in three key areas:

- **Technical Contribution:** The two-step input pipeline architecture provided by the study facilitates the integration of automated report generation using a vision-language model with deep learning image processing. Our study sites examples of prompt engineering methodologies for clinical report generation in mammography using MedGemma.
- **Clinical Contribution:** The automatic reporting system built via MedGemma utilizes automated generation of reports that contain specific details on cancer type, estimated cancer stage, and lesion characteristics. Thus, the framework streamlines reporting tasks and ensures consistency in the report format.
- **Contribution to Research:** We have introduced a reproducible evaluation procedure based on demonstration of effectiveness of AI models in mammography reporting. The study has published prompt engineering templates that may be utilized by fellow researchers using MedGemma for automated report generation.

1.5 Organization of the Thesis

The current thesis is organized as follows :

- The next chapter provides a theoretical framework that discusses many issues surrounding breast cancer imaging, deep learning (classification and segmentation), vision-language models, prompt engineering, and automated report generation, along with relevant literature review and the gaps in the previous works utilized to justify the chosen methodology.

- In the third chapter, we provide a detailed description of the proposed methodology, which describes our pipeline from the data preparation stages all through to report generation.
- The fourth chapter presents the results of our experimental investigation as per the performance of the developed pipeline with respect to both segmentation, classification accuracy, and report generation performance of our system.
- The concluding chapter will summarize the important outcomes, discuss possible limitations, and suggest recommendations for future research activities.

Chapter 2: Theoretical Background and Related Work

2.1 Introduction

This section introduces the theoretical foundation for our two-stage automated mammography reporting pipeline and reviews the existing literature in order to establish a solid scholarly basis for the study. This infrastructure connects deep learning mammography analysis (classification and segmentation) to the automatic generation of clinical reports via prompt engineering over a medical vision-language model.

The discussion focuses on three important areas. First, it provides the clinical background in terms of breast cancer screening, mammographic views, and BI-RADS reporting lexicons. After that, it will present the technical aspect of the whole approach including convolutional neural networks for classification of imaging, U-Net encoder-decoder model for spatial segmentation, vision-language model for multimodal reasoning, and prompt engineering for steering purpose. Finally, it provides a cumulative review of fourteen studies in which above-mentioned research tasks were completed.

Chapter 2.2 presents the theoretical framework. Chapter 2.3 reviews the results of empirical studies in terms of the four categories. Chapter 2.4 provides a synthesis of these findings into a formalized gap analysis for the proposed model.

According to previous studies on multimodal biomedical artificial intelligence, the use of both vision and text modalities produces better diagnostic capabilities than using one or the other in the analysis [11].

2.2 Background

2.2.1 Breast Cancer and Mammographic Imaging

The uncontrolled growth of abnormal cells in breast tissue leads to breast cancer, which usually starts in the mammary ducts (invasive ductal carcinoma, IDC) or the lobules (invasive lobular carcinoma, ILC). Most breast cancers, 70-80% of them, are IDC, while 10-15% is ILC. Invasive types consume most mammograms, nevertheless non-invasive types, such as ductal carcinoma in situ (DCIS), describe abnormal cell growth occurring only in the ductal surface tissue without invading the basement membrane. Identifying the specific histological subtype is essential for establishing the surgery plan, choosing systemic treatment, and predicting prognosis [1]. Mammography is still the preferred type of diagnosis for early breast cancer detection. Mammograms consist of two standard views from each breast: craniocaudal (CC) and mediolateral oblique (MLO). Unlike older analog systems, full-field digital mammography (FFDM) is superior in terms of the dynamic range, soft tissue contrast, and direct usage in digital image collecting [2].

The American College of Radiology has developed the Breast Imaging Reporting and Data System (BI-RADS) for the standardization of the mammography process. BI-RADS has developed a language that consists of terms on mass shapes (round, oval, irregular), margins (circumscribed, obscured, microlobulated, indistinct, spiculated), and calcification patterns (amorphous, coarse heterogeneous, fine pleomorphic, fine linear). The variables serve for the establishment of the BI-RADS final category, where 0 means incomplete assessment, and 6 means known cancer [3, 12].

To illustrate BI-RADS' impact on the clinical approach, consider a case of an irregular mass (1.8 cm) with spiculated margins in the right breast upper half. By using BI-RADS language, spiculated and irregular characters of the mass confirm its malignancy. The radiologist using these characteristics borrows BI-RADS 5 name to categorize it. In this case, the cancer prediction value observed in BI-RADS 5 is over 95%, where core needle biopsy is prescribed [3]. Should the mass have round and circumscribed attributes, the BI-RADS category would be 3 with a maximum 2% risk of malignancy recommending further tests in interval of six months. The model used in our pipeline should be tailored to these features, as its results are based on these variations in morphological images.

In cancer staging, the TNM (Tumor-Node-Metastasis) classification is used. Of note, in pathological examination and imaging of metastasis, the tumor dimension estimated from mammograms is used for the T category. The maximum invasive diameter ranges through small (≤ 20 mm), medium (> 20 mm to ≤ 50 mm), and big (> 50 mm) categories of T stage. In addition, when there is noticeable impact on the chest wall or skin, the T category is T4 regardless of the tumor size. For example, a 5.5 cm mass with spiculated margins located on the skin is T3 by definition. All data obtained from these measurements are necessary for the report generation process [13].

2.2.2 Deep Learning for Medical Image Classification

Since AlexNet presented the ability of deep learning to learn and extract relevant features for medical imaging using Convolution Neural Networks (CNNs), CNNs have been the basis of many medical imaging methods. A convolutional layer applies its weights on an input image tensor I as follows:

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

The formula above shows the output of the convolutional layer at pixel coordinate (i, j) for a kernel of size $m \times n$. After each convolution is the nonlinear activation function, which, in most cases, will be the Rectified Linear Unit (ReLU) which is defined as $f(x) = \max(0, x)$. The ReLU function is useful in limiting the saturation of the gradient at the time of back-propagation.

In order to show how the convolution takes place, let's examine a 3×3 kernel that is applied to a grayscale mammogram image as seen in Figure 2.3.

In this case, this kernel acts as a horizontal filter. The kernel gives large and positive output value in the mammogram patch where pixel intensities changes from the dark (the fatty tissue) to light color (the dense lesion). If a pixel is in a uniform region, then the output of the filter is near zero.

The CNN learns hundreds of such filters with every layer. The early layers detect edges and textures; the mid-layers combine them to form circular shapes whereas the later layers compose the circular shapes to various diagnostic categories. [14, 15].

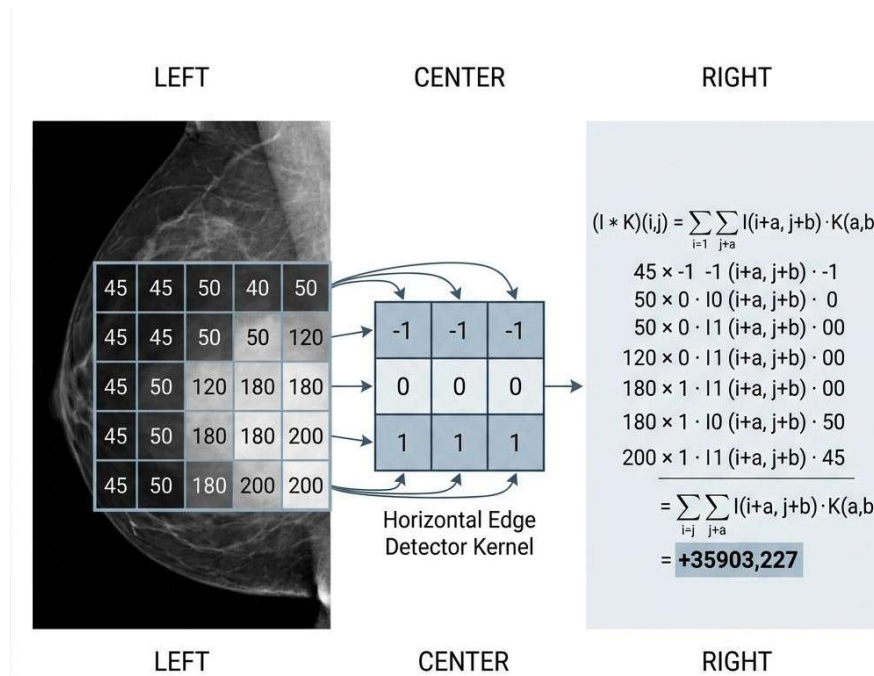


Figure 2.3: Convolution operation on a mammogram patch showing a 3x3 kernel applied to a grayscale input region.

Many derived versions of CNNs have appeared with various innovations. One of them is ResNet that has used skip connections in order to train networks with more than 100 layers. The skip connection is known to be an identity mapping obtained through the following equation:

$$y = F(x, \{W_i\}) + x \quad (2.2)$$

Where x is the input, $F(x, W_0)$ is obtained through the above formula.

In this case, it is noted that F retains the form of the identity mapping when there is no advantage of further deepening the model. [16].

For mammogram diagnosis, softmax is used for the very last layer of the network :

$$P(z) = \frac{e^{z_c}}{\sum_{k=1}^C e^{z_k}} \quad (2.3)$$

To minimize the cross-entropy loss based on the predicted probability distribution and the true label, we have the equation below :

$$L_c = -\log \log P(z) \quad (2.4)$$

When working with a dataset that comprises N labeled mammograms, the total training loss can be expressed as the mean of individual losses. The optimizer (generally either Adam or SGD with momentum) adjusts the weights of the network based on the gradients of the loss with respect to all parameters computed through backpropagation. [49].

In the case of classification of mammograms, transfer learning has been found to be efficient. Pre-trained models using ImageNet can be fine-tuned for mammography datasets, thus employing the advantage of low-level feature representations (for example, edges and textures) which can be transferred from natural images.

Transfer learning creates an opportunity to use low-level feature representation because the initial convolutional layers learn common visual features like edge orientations, color gradients, and simple textures among different images. While using a pre-trained model and deploying fine-tuning to classify mammograms, it will only be required to update the top-layer and any other layer that acquires domain-specific features. Therefore, the labeled data would be needed up to an order of magnitude less as opposed to training from scratch [17].

In addition, various data augmentation techniques are available for implementing this approach. A few of such techniques include rotation, flipping, and Contrast Limited Adaptive Histogram Equalization (CLAHE). In the CLAHE technique, which subsists by dividing the image into small tiles and applying histogram equalization to each of them independently, it is ensured that there are no outbursts of noise.

Recently, Vision Transformers have been introduced in medical imaging. Unlike other CNN architectures which create hierarchy using local features maps, ViTs divide the images into patches and embed them in the self-attention structures through the use of attention which enables ViTs to focus on long distances within the images. In light of this fact, the self-attention is able to determine how various patches interact with each other. (see figure 2.5)

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (2.5)$$

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].

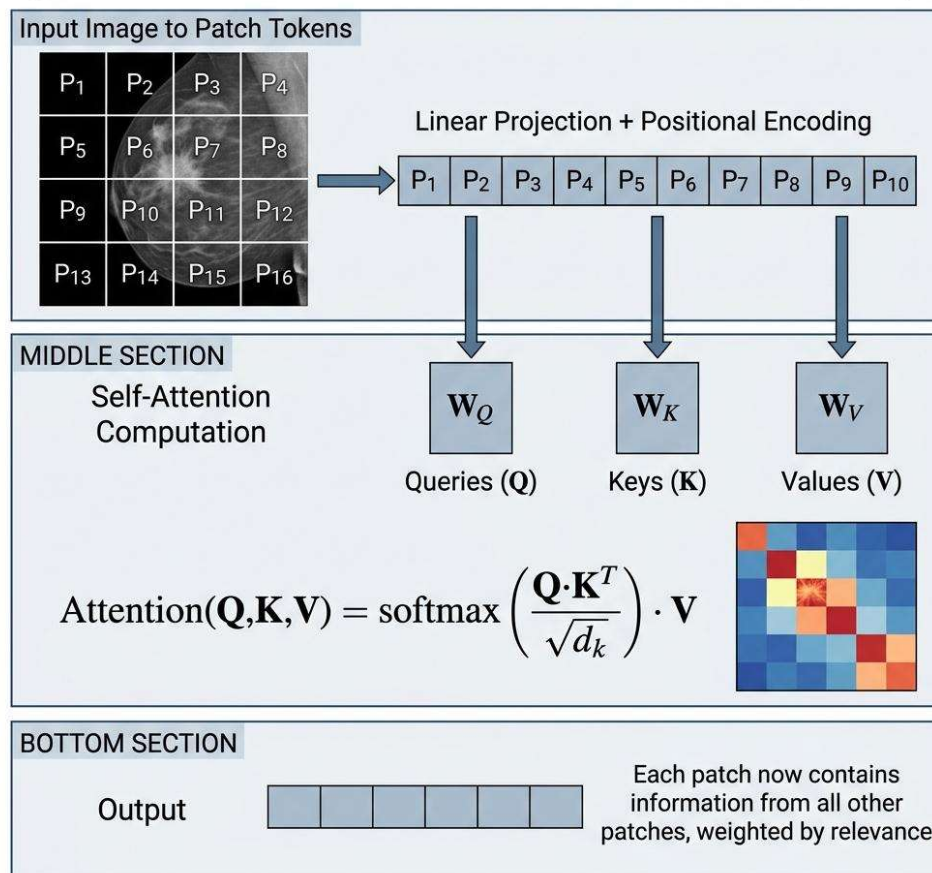


Figure 2.5: Illustration of the self-attention mechanism in Vision Transformers, showing query-key-value computation for image patch tokens.

The clinical application of artificial intelligence models continues to create issues regarding interpretable results. Grad-CAM uses the weighted activations of the last convolutional layer, with weights being the gradients of the score of the respective target class. Specifically, the weight given to channel k is given by:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y_c}{\partial A_k(i,j)} \quad (2.6)$$

where the score for class c is denoted by y_c , and $A_k(i,j)$ is the activation of the k -th channel in spatial location (i,j) , and Z indicates the number of spatial locations. The heatmap is then the ReLU of the weighted projection: $L_{\text{Grad-CAM}} = \text{ReLU}(\sum_k \alpha_k A_k)$, where ReLU disregards any negative values to only consider features that positively contribute to the models' prediction [19]. Therefore, if radiologists see that the heatmap is focusing on the correct site for the lesion, then classifications can be considered as based on relevant morphological properties, otherwise, radiologists should take classification with skepticism.

2.2.3 Medical Image Segmentation

Segmentation, on the other hand, provides a label to every pixel of the image. With respect to mammography, the representative structures are solid lesions and clusters of calcifications. Segmentation can also be considered in breast region segmentation.

Finally, whereas classification predicts one label for the whole image, segmentation provides an output map identical to the input with each pixel assigning either "lesion" (1) or "background" (0).

U-Net architecture, defined by Ronneberger et al. in 2015, remains the most widely adopted model in medical images [20]. U-Net uses an encoder-decoder approach with the help of skip connection transferring the details of the fine-grain spatial information from encoder layer to the other corresponding encoder block. Because of this property, the network is able to have precise boundary delineations, with little training data as is often the case for medical imaging.

In essence, the encoder is similar to a typical CNN structure where each block includes two 3×3 convolutions (with a ReLU activation following each convolution), and two pooling operations allowing to replace the width and height by half every single time. For instance, if the mammogram image has a resolution of 512×512 , then its representation on the encoder can be observed in feature map resolutions of 256×256 , 128×128 , and lastly, 64×64 . Conversely, the decoder takes the results of the reverse process, applying the 2×2 up-convolution (transposed convolution) that doubles the sizes of the map every time, concatenating it with the current encoder map through using skip connection, and two further convolutions with 3×3 kernels.

The final layer for U-Net is 1*1 convolution producing a multi-channel output for a single-channel representation, and then followed by a sigmoid activation to create per pixel probabilities ranging from 0 and 1.

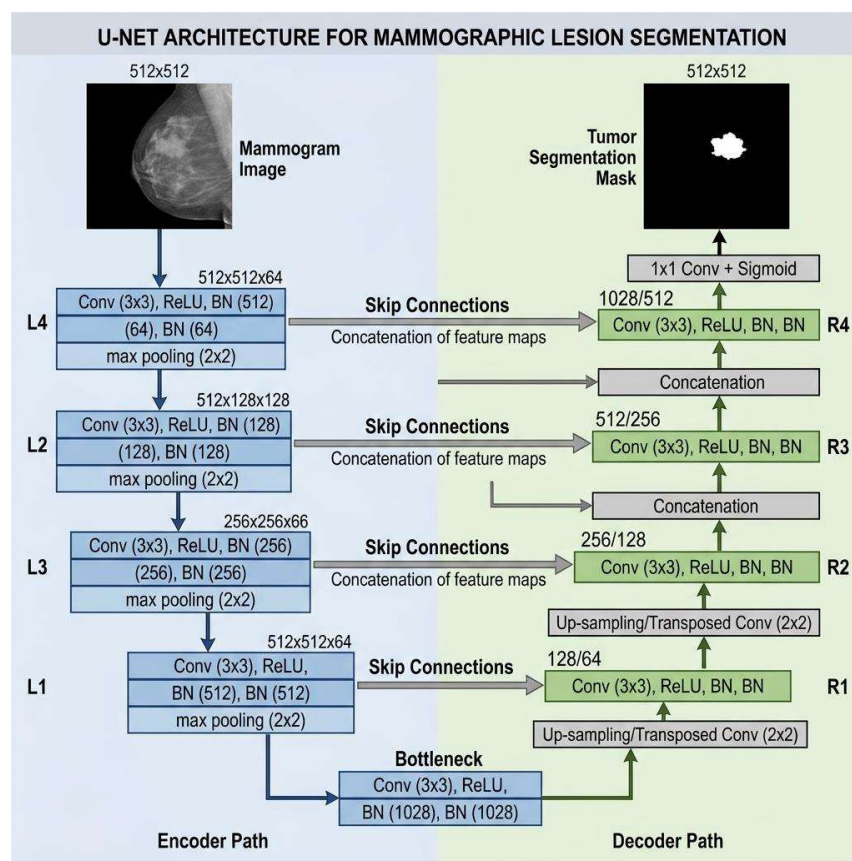


Figure 2.1: Schematic representation of the U-Net segmentation architecture for mammographic lesion delineation.

In order to train a segmentation model, a loss function is needed to quantify the difference between the predicted mask P and the ground-truth mask G . In the medical imaging field, there are two standard loss functions used for image segmentation:

Binary cross-entropy (BCE) loss function, which is applied to every single pixel

$$BE = -\frac{1}{N} \sum_{i=1}^N [g_i \log(p_i) + (1 - g_i) \log(1 - p_i)] \quad (2.7)$$

And dice loss function, which directly maximizes the Dice Similarity coefficient :

$$\mathcal{L}_{Dc} = 1 - \frac{2 \sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \quad (2.8)$$

Dice coefficient example (see Figure 2.4): Let us assume a mammogram is being analyzed that contains a tumor covering 400 pixels in the ground-truth mask. The segmentation model predicts a mask that covers 450 pixels, out of which 370 are matching with the ground truth.

TP = 370

FP = 80

FN = 30

The Dice coefficient = $2 \cdot TP / (2 \cdot TP + FP + FN) = 740/850 = 0.871$
and IoU = $TP / (TP + FP + FN) = 370/480 = 0.771$.

A Dice coefficient equal to 0.871 represents an acceptable overlap; however, it indicates the model is slightly over predicting the tumor size due to 80 false-positive pixels and that it has also missed a part of the tumor due to 30 false-negative pixels which mostly happen at the boundaries of the tumor with the surrounding tissue. The presence of 30 false-negative pixels can be a matter of concern as when a full tumor boundary is not segmented it might lead to an

underestimation of the tumor size in the report. The relationship between Dice and IoU is monotonic, and therefore Dice would always be greater than or equal to IoU in cases of the same prediction.

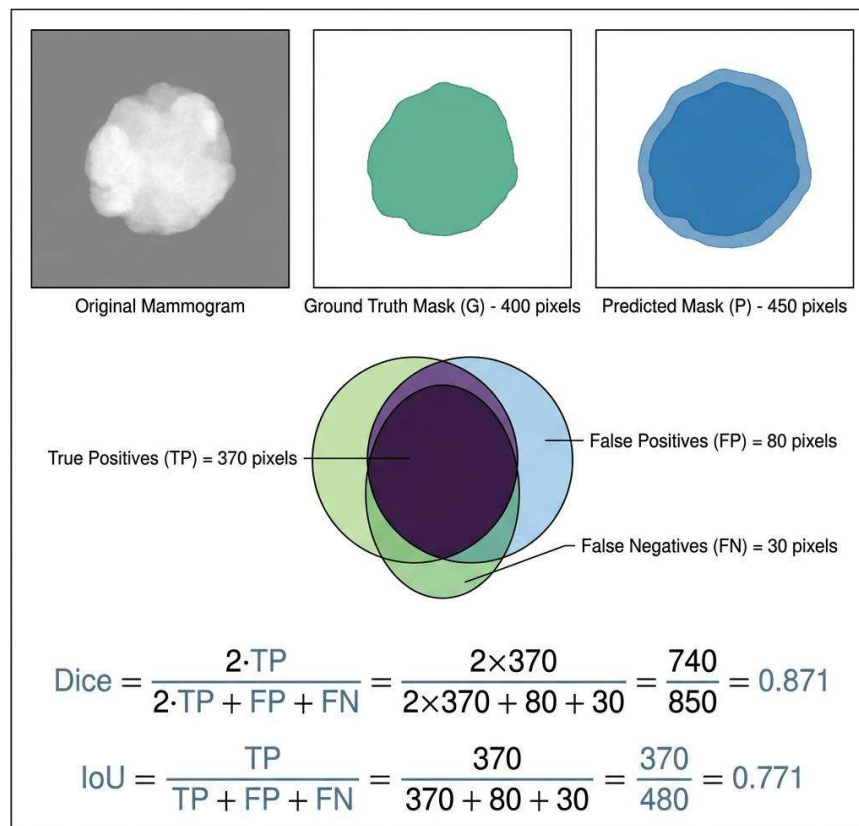


Figure 2.4: Visual example of Dice coefficient computation between a predicted segmentation mask and ground-truth annotation for a breast mass.

The original U-Net framework was outperformed in terms of Dice scores by the modified U-Net framework.

Attention U-Net developed the concept of the use of attention gates to eliminate regions of lack of interest and direct the model to concentrate on the target region, which has a lot of utility, particularly when it comes to cases where the lesions are smaller than the total image size [22]. The use of attention gate is to calculate the gating coefficient for the different spatial channels in the skip connection feature map:

$$\alpha_i = \sigma(\psi^T \sigma_1(W_x x_i + W_g g + b_g) + b_\psi) \quad (2.9)$$

Where x_i is the input feature vector in the skip connection and g is the gating function generated from the decoder (which is able to give less precise data, but is more semantically meaningful). Output α_i is between 0 and 1, with those with α_i around 0 being eliminated while those near to 1 being transmitted without any change.

In more recent hybrid models like Hybrid Transformer U-Net (HTU-Net), we see a combination of self-attention mechanisms of Vision Transformers with the original U-Net framework to achieve local and global spatial context [5]. In this instance, the deepest part of the U-Net model (where the input data are of the smallest spatial size but having maximum number of channels) is replaced with the use of a set of Transformer encoder layers. For example, the bottleneck of the U-Net model, which processes input data of 512×512 is using a feature map with dimensions $32 \times 32 \times 512$, allowing it to use 1024 tokens having the feature dimension of 512, something that is computationally feasible, unlike self-attention using full dimensions of 512×512 , which would require computations on 262,144 tokens. This way, the Transformer will capture global relations between isolated portions of the mammogram, while decoder will again upsample the result to the original dimension using the skip-connected transposed convolution blocks.

The performance of segmentation is assessed using different metrics: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), pixel accuracy, sensitivity, and specificity. The DSC gives the percentage of overlapping between predicted and actual masks, where 0 refers to no overlap while 1 is for total overlap. IoU also calculates the percentage but gives more weight to false positives. Sensitivity is calculated as $TP / (TP + FN)$, while specificity is described as $TN / (TN + FP)$. With mammography images, sensitivity is of more significance than

specificity since a loss of lesion region is more severe than having an extra incorrect region. At the CBIS-DDSM benchmark, the recent models achieved at least 0.90 – 0.96 Dice scores for breast mass segmentation [5, 23].

2.2.4 Vision-Language Models and Medical AI

Vision-Language Models (VLMs) perform processing tasks using two types of input data, namely images and text. A typical VLM architecture consists of a vision encoder that extracts features from the images and a language model that generates text through feature conditioning. Resolving usually entails the use of CNN or ViT-based vision encoders and a large language model, which is based on the Transformer architecture. [11].

In VLM models, the vision encoder sends data to the language model through the use of either a projection or a cross-attention method. In the projection method, the vision encoder produces a series of image feature vectors $v_1, v_2, \dots, v_n$ for each of the image patches or areas, and with the aid of a learned projection matrix, W_{proj} , the vector v_i is mapped into the same embedding space as the text tokens of the language model, so that the generated image token $h_i = W_{proj} \cdot v_i$. The concatenated image token and the text token embeddings are fed into the language model as an intertwining sequence of data, which is processed by each of the text token generation self-attention layers so that attention is distributed to both the previously generated text token and to the image features. [11, 6].

For the cross-attention approach which is used in architectures such as Flamingo, the language model uses specific cross-attention layers, which are intermixed with

the self-attention layer such that the queries consist of the language model tokens while the keys and the values consist of the image feature vectors.:

$$CrossAttention(T, V) = softmax\left(\frac{TW_Q(VW_K)^T}{\sqrt{d_k}}\right)VW_V \quad (2.10)$$

In the field of medicine, some examples of VLMs include medical visual question answering, which entails generating answers from a given image and a clinical question, and diagnostic report generation through producing descriptions of findings from an image. With respect to the latter task, the processed images are prepared in their respective structured text formats, which contain text types such as diagnosis of the images and tumor size, which are tokenized and merged with the image tokens. Thus, the clinical report is generated through the language model by the generation of the text tokens.

2.2.5 Prompt Engineering for Medical AI

Prompt engineering is the process of designing the input text provided to a language model or visual language model (VLM) to control the shape, content, and quality of the reasoning in its results. In the case of fine-tuning, which involves adjusting the model's weights, prompt engineering operates entirely at the inference step and requires no training data or more than a one-time forward pass [7].

The theory behind prompt engineering is based on the self-regressive 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 calculating:

$$P(x_1, \dots, x_t) = \text{softmax} \left(\frac{h_t W_e^T}{\tau} \right) \quad (2.11)$$

where h_t is the hidden state of the model at time t , W_e is the code embedding matrix (which is used in reverse as a return projection to the vocabulary space), and τ is the degree of temperature. The prompt (input text) defines the initial sequence $x_1, \dots, x_t$, and thus forms the distribution from which all subsequent tokens are sampled. A carefully designed prompt moves the probability cluster towards output that is clinically valid, but poorly designed prompting allows the model to shift towards generic or invalid outputs.

The value of the “tau” coefficient has an impact on the distribution of results intensity. If $\tau=1.0$, we can expect to have the probability calculated as per its usual methods. If $\tau < 1.0$, i.e., τ equal to 0.3, the distribution of the probabilities sharpens, and the resulting symbol with the highest probability receives more weight in the model. If $\tau > 1.0$, the distribution flattens and thus more differentiating outputs are obtained. In the case of the medical reporting the recommended method is to use lower temperature as τ of 0.2-0.5 should be used since it is more important to have results precise than diverse.

Different prompting methods were developed and tested.

Zero-shot prompting provides only an instructional prompt and input data, without any samples. It tests the extent to which the model's prior training has ensured sufficient domain knowledge to perform the task correctly. In medical imaging, zero-shot prompting often produces reports that are grammatically correct but medically incomplete [7].

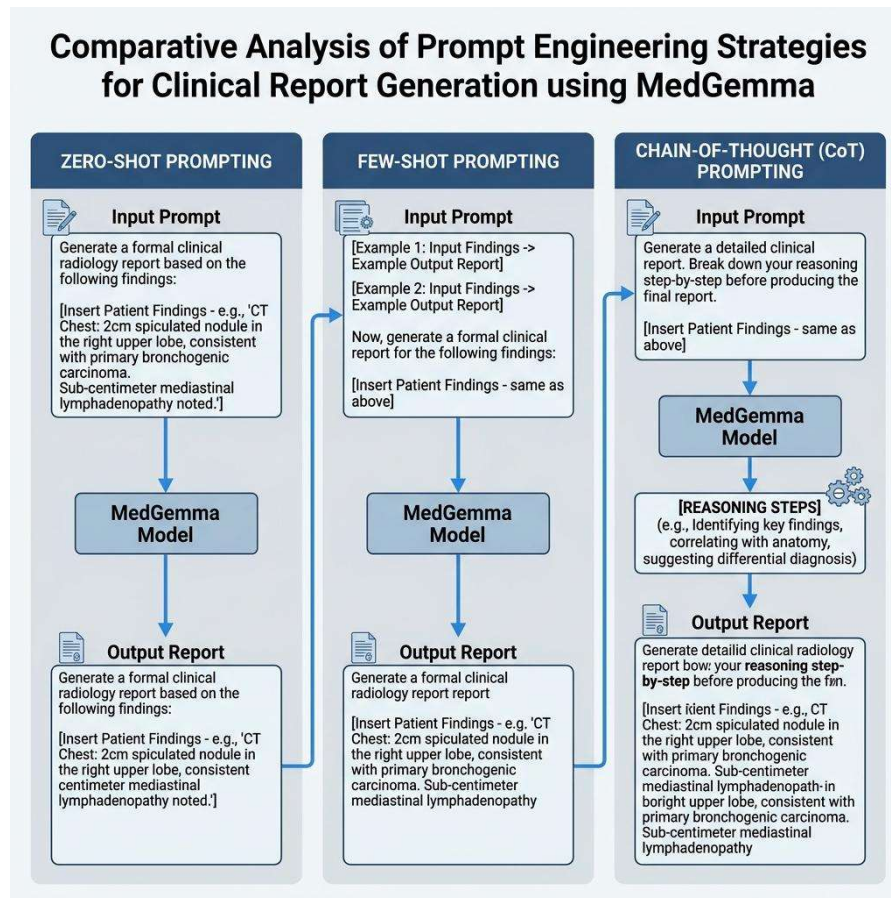


Figure 2.2: Comparison of prompt engineering strategies (zero-shot, few-shot, chain-of-thought) applied to MedGemma for clinical report generation.

The structure of the output is set by the output guidelines, for example, by requiring that the output be in JSON with specific fields (lesion_type, location, size, birads_category, recommended_action). The use of structured output reduces the number of free-form “hallucinations” and makes the output suitable for further automated processing in future stages [28].

This research tested all five generation strategies on the same set of mammography images. The assessment was based on two dimensions: linguistic quality (measured using BLEU and ROUGE-L scores) and accuracy of content, as determined by the evaluation of radiologists. The goal was to explore the role of guidance strategies and to define the type of strategy that can balance report quality and computational cost most effectively.

2.2.6 Automated Radiology Report Generation

Automated radiology report generation (RRG) is an artificial intelligence-based technology that generates narrative clinical reports from medical images. It has evolved from early template-filling systems to today's end-to-end neural models [29]

Most research on radiological representation graphs (RRGs) has been focused on chest X-ray images, where the large annotated datasets (such as MIMIC-CXR, which contains more than 370,000 images, and IU X-ray, which contains 7,470 images) offer sufficient training data. In the case of mammography-specific RRGs, however, there has been much less attention, partly because annotated mammography datasets accompanied by paired reports are less available, and partly because mammography findings are more spatially specific 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:

$$\text{BLEU-}N = \text{BP} \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (2.12)$$

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:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1 - r/c} & \text{if } c \leq r \end{cases} \quad (2.13)$$

The length denoted by c is related to the length of the generated text whereas r refers to the length of the reference text. The brevity penalty discourages artificially short texts that have very high precision because only a limited number of n-grams are confidently produced by them [51].

An example of a BLEU-1 count (see Figure 2.6): Let's assume that the created report includes the following statement: “A spiky mass measuringAccording to the report, “A spiculated, irregular mass measuring 2.3 cm was located in the upper outer quadrant of the left breast and was categorized category 5 in BI-RADS” (23 words). The taken reference report says “An irregular, spiculated mass of 2.3 cm was noted in the outer upper quadrant of the left breast that is aligned to the severity of a BI-RADS 5 category” (27 words). Single-word 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 words created, 14 match. The modified single-unit accuracy is $p_1 = 14/16 = 0.875$. Since $c = 16 < r = 22$, the brevity penalty applies: $BP = \exp(1 - 22/16) = \exp(-0.375) = 0.687$. $BLEU-1 = 0.687 \times 0.875 = 0.601$. The reason for the relatively low score, despite the good overlap between words, is that the generated text is much shorter than the reference text.

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

$$ROUGE - L = \frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2P_{LCS}} \quad (2.14)$$

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 β is typically set to favor recall ($\beta = 1.2$) [52].

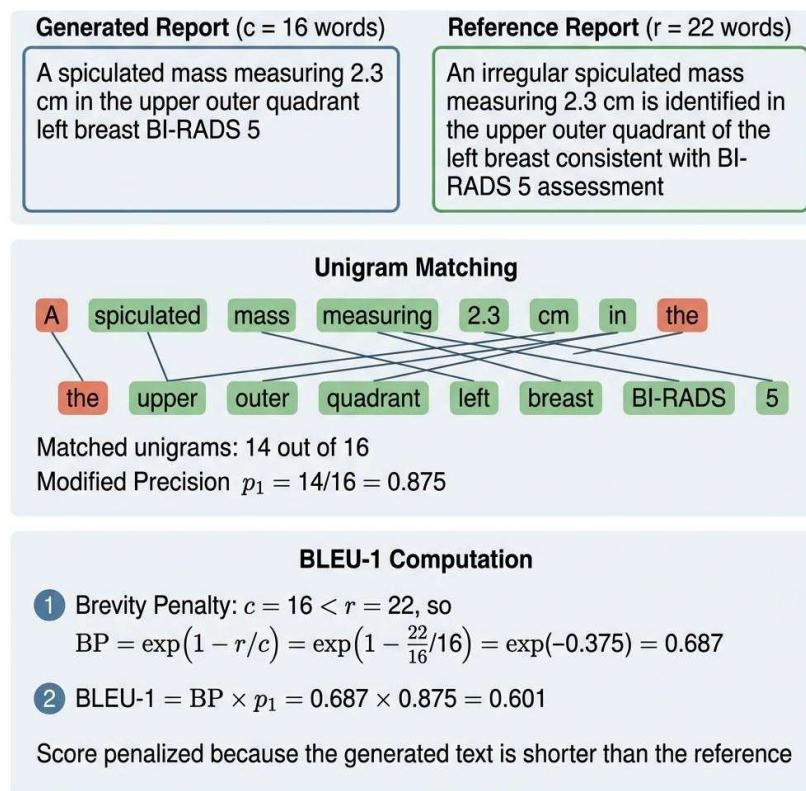


Figure 2.6: Worked example of BLEU score computation for a generated radiology report against a reference report.

These metrics correlate poorly with clinical correctness. A report could score high on BLEU by matching most words while misidentifying the lesion type. For These metrics are not closely related to clinical health. A given report may score high on the BLEU metric by matching most of the words, yet still incorrectly identify the type of lesion. For instance, a created report stating “a well-defined mass in the upper outer quadrant, BI-RADS 3” may share many n-grams with a report stating “a spiny mass in the upper outer quadrant, BI-RADS 5,” but the clinical meaning is completely different: one advises routine follow-up, while the other mandates a biopsy. Clinical accuracy assessment, typically performed by radiologists who compare the reports generated with real-world facts, is still the most reliable evaluation method, but it is expensive and time-consuming [29, 31].

More recent approaches such as RadAlign (2025) have proposed the idea of concept-based alignment, in which features are first related to a set of predetermined diagnostic criteria before being fed into the language model. This middle step acts as a bottleneck that requires the model to adhere to certain clinical outcomes before creating the text, thereby reducing the chance of “hallucinations” [32].

The AMRG (2025) pipeline specifically targeted breast radiography, using a “vision-language” architecture trained on high-resolution mammography radiography images paired with structured radiology reports. This model demonstrated that domain-specific training data and visual encoders adapted to

mammography improve the quality of report generation compared to models trained on general radiology data [33].

2.2.7 MedGemma

MedGemma is a group of basic medical vision and language models built on the Gemma 3 architecture, which Google launched as open-weight models. The group includes a 4 billion-parameter multimodal model and a modified version with 27 billion parameters. MedGemma has also been pre-trained on additional carefully selected medical text and imaging data, such as medical notes, biomedical literature, and medical images spanning the fields of radiology, pathology, and dermatology [6].

The model includes MedSigLIP, which is a medical editing vision encoder based on SigLIP, and this encoder analyses medical images producing related feature representations with the embedding space of the language model. SigLIP (Sigmoid Loss for Language-Image Pre-training) is a vision-language model of contrast type which gets trained on pairs of (image and caption) data for correctly aligning images with text. 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:

$$\mathcal{L}_{\text{Sgl}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log \sigma(z_i) + (1 - y_i) \log(1 - \sigma(z_i))] \quad (2.15)$$

Where $z_i = t_i \cdot v_i^T$ is the similarity of the dot product between the text representation t_i and the image representation v_i , $y_i = 1$ if the pair is matched and 0 in other cases, and σ is the sigmoid function. This formulation allows for efficient training on large datasets without requiring the full $N \times N$ similarity matrix needed by softmax losses, thereby reducing memory requirements [53].

MedSigLIP is a SigLIP-based model that has been additionally training on medical image-text pairs (radiology images with reports, pathology slides with descriptions, and dermatology images with diagnoses). This additional training adjusts the embedding space so that Medical concepts are grouped appropriately: for example, the embedding of a mammogram showing a spiculated mass becomes closer to the embedding of the text “mass suspected to be malignant” and farther from the embedding of “well-defined benign fibroadenoma.” This correspondence allows MedGemma to reason through visual attributes using medical language. In benchmark assessments, MedGemma outperformed generative models of similar size in the areas of answering multimodal medical queries (improvement of 2.6–10%), classifying chest X-ray findings (improvement of 15.5–18.1%), and active clinical assessments (10.8% improvement) compared to the baseline Gemma 3 models [6].

The number of parameters in the 4B model for the multimodal version is distributed approximately as follows: The vision encoder (MedSigLIP) has approximately 400 million parameters, the projection layers that link vision to language have approximately 100 million parameters, and the language model (based on Gemma 3) has approximately 3.5 billion parameters. The language model uses a Transformer architecture with grouped query attention, Rotary positional embeddings, and SwiGLU activations [6].

The MedGemma 1.5 version expanded the model's abilities to process high-dimensional medical data, including computed tomography (CT) and magnetic resonance imaging (MRI) sets, full-section histopathological images, and longitudinal patient data. For the mammography applications proposed in this study, the 4B multimodal version is used as it offers sufficient clinical inference capabilities while being deployable on consumer-grade devices. In the embedded form (4-bit embedding via GPTQ or bitsandbytes), the model requires approximately 3–4 gigabytes of GPU memory, which is consistent with the 15 gigabytes of memory available on the T4 GPU in Google Colab [34].

One important feature of MedGemma in the context of this study is that it can be used successfully with claim data alone, without the need for fine-tuning. The model's pre-training on radiology data means that it already possesses great clinical knowledge about mammography findings, BI-RADS terminology, and reporting structure. The question of this research is whether it is possible for carefully

designed prompts to extract this information in the specific format needed for mammography reporting [6].

2.2.8 Mammography Datasets: CBIS-DDSM and VinDr-Mammo

The Breast Imaging Coordinated Subset of the Digital Database for Mammography Screening (CBIS-DDSM) is a unified version of the original DDSM dataset, coordinated to support research in computer-aided detection and diagnosis. It includes 2,620 mammographic studies on scanned films, with pixel-level annotations for masses and calcifications. Every case includes classifications that have been verified by pathologists (benign, malignant, or benign without the need for further evaluation), BI-RADS assessment categories, and annotations on lesions provided by trained radiologists. This dataset was published through the Cancer Imaging Architecture (TCIA) and is still one of the most widely used benchmarks for breast radiograph classification and segmentation [35].

The tumor subset of the CBIS-DDSM database—the main subset used in this thesis—contains 1,696 images from 753 tumor cases. The class distribution is imbalanced: approximately 54% of cases are benign (including “benign cases that do not require re-examination”) and 46% are malignant. This unbalanced distribution requires attention during model training and is usually addressed through stratified sampling, class-weighted loss-based functions, or oversampling of the minority class. Images are saved in the DICOM format with different resolutions (which in the case of full-field mammograms are generally 3,000 to 5,000 pixels on each side) and should be preprocessed by transforming them into either PNG or JPEG formats, resizing them (most often to the sizes 512×512 or 1024×1024), and normalizing them according to a certain density range.

A newly proposed extension called CBIS-DDSM-R enables the usage of a radiomics-ready version of the dataset by including preprocessing processes as well as feature extraction that adheres to the IBSI guidelines.

VinDr-Mammo is a wide-ranging dataset of digital mammography (FFDM) images collected from Hanoi Medical University Hospital in Vietnam. It includes 5,000 mammography screening exams (20,000 images with four standard projection positions) with BI-RADS breast-level assessment annotations and lesion localization using the “envelope box.” Unlike the CBIS-DDSM dataset which has been formed from images scanned from the films, the VinDr-Mammo dataset has images of real digital mammograms that use modern clinical imaging techniques. The dataset has been published in PhysioNet, and is being used to train AI models for detection and classification tasks [37].

Despite a few differences present in these datasets, they have something in common that makes comparing models worth it. In contrast to the CBIS-DDSM dataset which contains pixel-wise segmentation masks good for training strategies in image segmentation, the VinDr-Mammo dataset is consisting of raw digital mammograms together with bounding boxes. As such, the VinDr-Mammo dataset focuses on detection and classification tasks and requires further processing of the identification markers to adapt them for direct segmentation. In addition, these two datasets have stark differences in terms of patient demographics (North America vs. Vietnam), imaging technology (film-screen radiography vs. direct digital mammography (FFDM)), and imaging protocols. Demonstrating strong performance on both datasets confirms the robustness of the AI model and its generalizability across diverse imaging modalities and different clinical populations—a critical hurdle before real-world deployment.

In this thesis, CBIS-DDSM serves as the main reference standard for segmenting masses (using pixel-level masks) and categorizing them (using their pathological classifications). VinDr-Mammo, on the other, serves as a secondary assessment dataset to test how well the model generalizes across different imaging techniques and patient cohorts.

2.3 Related work

The related work is grouped into four categories: mammogram classification, mammogram segmentation, medical reporting using VLM models, and claims engineering for clinical applications. Every study is reviewed for its contributions and limitations, as these limitations define the scope of this thesis's contribution.

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. Salamah and Aly (2021) [39] compared several 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 pre-trained weights from ImageNet and CLAHE pretreatment. The DenseNet-169

network had the highest accuracy at 94.5%, followed by the InceptionV3 network at 93.2%. The systematic evaluation of different architectures on a common dataset allows for easy comparison of results with later research. The work is focused on classification and does not deal with segmentation or grouping.

Ayana et al. (2023) [40] assessed the EfficientNetV2 model using the Grad-CAM++ tool to highlight visual results for classifying mammograms in the CBIS-DDSM and INbreast datasets. They integrated thermal maps to visualize features, with the aim of highlighting image regions that affect model predictions. The EfficientNetV2 model reached a classification accuracy of 96.2% on the CBIS-DDSM dataset. The utilization of heat maps allows health care practitioners to confirm that the model focuses on the pertinent parts of the human body. The early stages of heat mapping do not clearly define the edges of lesions or create any report of text.

Lund-Hansen et al. (2025) [41] evaluated the commercial artificial intelligence (AI)-powered “Transpara” system for screening purposes in the prospective, randomized, controlled “MASAI” clinical trial, which included more than 105,000 women in Sweden. The system performed an automated risk assessment to triage mammography images, identifying high-risk examinations for double reading while directing low-risk cases to single reading. AI-assisted screening led to a 29% increase in cancer detection and a 44% reduction in the workload for radiologists without increasing false-positive rates. The tests indicate a good potential for the application of artificial intelligence in terms of the efficacy of mass public screening. The system serves only as an initial assessment tool, producing no accurate classification, segmentation boundaries, or formal medical reports.

2.3.2 Mammogram Segmentation Studies

Jalalian et al. did a study on their 2025 paper where they tested UNet++ on the mass segment of the CBIS-DDSM dataset concerning mammographic mass segmentation. Using the ResNet-34 encoder, they carried out a UNet++ modification with geometric augmentation (flipping, elastic deformation, etc.). 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 efficiently obtain the multi-scale feature representations that are utilized in obtaining lesion boundaries. In this research, segmentation was evaluated independently without connecting the generated masks to any other processes such as classification or the generation of clinical reports.

The hybrid architecture of HTU-Net, a Hybrid transformer U-Net, is presented by Mohammadi and Livani (2025). In this study, HTU-Net was trained and tested on CBIS-DDSM and INbreast datasets. The architecture highlights the ability of the convolutional neural network (CNN) to capture both local tissue edges and global spatial dependencies through the integration of the CNN with the multi-head self-attention mechanism. HTU-Net achieved a Dice coefficient of 0.9410 on CBIS-DDSM, surpassing standard U-Net at 0.8876 and UNet++ at 0.9215. Based on the same benchmark splitting, the application of Transformer attention along with encoder-decoder skip connections enhances the localization accuracy for having spatial boundaries of irregular lesions. However, the work is solely limited to pixel segmentation of lesions and does not include any tumor type identification or automatic reporting.

Sechopoulos et al. (2024) [42] investigated breast region segmentation on multi-vendor raw and processed digital mammography datasets. Their trained modified U-Net design was utilized to separate breast tissue from surrounding background noise and pectoral muscle in images taken from different imaging devices. The tested model obtained Dice coefficients ranging from 0.95 to 0.96 based on various test datasets. The fact that U-Net-based preprocessing shows good performance in cross-vendor testing proves that this model is capable of detecting anatomical areas accurately before any features are extracted from them. 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] presented the “Interactive Radiology Report Generation” (IRRG) framework, which was evaluated on the MIMIC-CXR chest X-ray dataset. They created a region-based dialog architecture that allows radiologists to review the report generated and request interactive corrections to specific sections of the report. The model achieved a ROUGE-L score of 0.278 and clinical accuracy scores comparable to those achieved by second-year radiology residents. The real-time integration of human feedback improves the clinical reporting accuracy and ease of use in the interactive workflow. The system was trained exclusively on chest X-ray images and requires specialized training on weights rather than relying on “zero-shot” base models or supporting mammography formats.

Jahangir et al. (2025) [44] developed the LLaMA-XR model to generate radiology reports on the “IU X-ray chest radiograph” benchmark dataset. They fused a

DenseNet-121 visual coder with the LLaMA 3.1 model using 4-bit “Quantum Low-Rank Adaptation” (QLoRA), a parameter-efficient fine-tuning technique. The model reached a ROUGE-L score of 0.433 after 12 hours of training on a single graphics processing unit (GPU). This study demonstrates that QLoRA can be used to fine-tune vision and language models for radiology reporting with limited device resources. This model is limited to chest X-ray images, requires supervised fine-tuning, and lacks support for BI-RADS breast imaging dictionaries or lesion segmentation integration.

Hiland et al. (2024) [45] suggested the MAIRA-2 model, a spatially-anchored multimodal vision-language model designed for radiology report generation, which was tested on the MIMIC-CXR dataset. They have integrated custom-made grounding tokens into a vision-language “Transformer” architecture to explicitly link the generated sentences to the boxes in chest X-ray images. The MAIRA-2 model showed a 10% improvement in clinical agreement according to the RadCliQ metric as compared to the most recent state-of-the-art models. Location anchoring effectively connects textual claims to visual features, which reduces clinical hallucinations. The system is designed specifically for chest radiography, relies on custom training data from the company, and has not been adjusted for mammography or BI-RADS report generation.

Chen et al. (2025) [33] introduced the AMRG model, a vision-language framework for automatically generating mammography reports, which was evaluated on the DMID dataset. They trained a vision model adapted for mammography, connected to a “sequence-to-sequence” language model, using paired samples of images and reports. The model achieved high scores on the BLEU metric and clinical accuracy when generating reports containing descriptions of findings, BI-RADS categories, and follow-up recommendations. The creation of a structure specifically designed for mammography terminology represents a clear advance in the field of automated breast radiology. The system relies on end-to-end supervised training on paired reports and does not use off-the-shelf base models or claim engineering strategies.

2.3.5 Prompt Engineering and Foundation Model Studies :

Singhal et al. (2023) [24] developed the Med-PaLM system by adapting the PaLM language architecture to the biomedical field, using biomedical literature from PubMed and medical question-answering data from the MedQA benchmark. They assessed the model using expert physician reviews and multiple-choice medical test questions. The Med-PaLM model achieved an Accuracy score >67% on the

MedQA benchmark, thereby exceeding the thresholds for medical licensure, while the subsequent Med-PaLM 2 model reached an Accuracy of 86.5%. This study showed that large foundational models are capable of storing and retrieving complex medical knowledge vital for clinical reasoning tasks. This approach is text-based and can't process breast imaging or generate visual segmentation inputs.

Yang et al. (2024) [28] carried out a thorough review of prompt engineering methods using GPT-4V based on various multimodal benchmarks in fields such as radiology, pathology, and dermatology. They used zero-shot prompting, few-shot prompting, chain-of-thought prompting, and self-consistency prompting without changing the parameters of the model. Chain-of-thought prompting showed an improvement of clinical reasoning accuracy when compared with zero-shot method of 8-15%. The benchmark provides a robust system for assessing prompt engineering in medical AI without any weight fine-tuning. The research is based on proprietary GPT-4V models of which it does not study open foundation models like MedGemma, also, the research does not involve reporting tasks specifically dealing with mammography.

Soren et al. (2023) [46] evaluated ChatGPT as a decision-support tool in breast cancer management, using 50 consecutive case summaries from the oncology committee at a single cancer treatment center. They used “zero-shot text prompting” to generate treatment recommendations for surgical and systemic treatment plans. The model achieved 73% agreement with the oncology committee’s decisions regarding surgical options and 77% agreement regarding systemic therapies. This test of the utility of large language models (LLMs) within real-world clinical decision-making workflows highlights the potential of AI-assisted support in the field of oncology. The study used only textual case summaries without processing raw mammography images, segmenting lesions, or generating structured radiology reports.

Kim et al. (2024) [47] evaluated the open foundation model MedGemma 1.5 (4B) over clinical NLP benchmarks like MedQA, MedMCQA, and PubMedQA. The model was evaluated using zero-shot prompting to assess domain-specific recall of clinical knowledge. MedGemma achieved a zero-shot accuracy of 81% on average, which is on par with larger proprietary models for the medical question-answering task. The results indicate that open-weight models in the medical domain are able to retain sufficient clinical knowledge without undergoing any specialized fine-tuning. The evaluation was limited to text-based NLP benchmarks and did not evaluate image inputs, mammography classification, or structured report generation.

According to the findings of Leibig et al. (2022) [48], a deep learning model has been introduced for screening mammography, which is based on uncertainty decision recommendations and is based on the analysis of a large retroactive dataset that comprised 1.19 million studies. The deep learning model automatically analyzed those mammography scans that have a low level of ambiguity and forwarded the more unclear results to radiologists for further reviewing. An advantage of this system is that it helps decrease a workload of radiologists by 30% without worsening rates of cancer detection in comparison with full manual screening. It proves that the introduction of AI can positively influence screening procedures by demonstrating its safe workload efficiency. The system mentioned above functions as a simple binary referral tool without providing much more complex lesion classifications and segmentation borders or diagnosis reports.

2.4 Related work Summary and Gap Analysis

This section of the document has summarized three findings. Firstly, the results of the analysis have indicated that deep learning technology used for breast X-ray evaluation has achieved a 90% accuracy level, with some methods producing results similar to trained medical specialists [38, 39, 40]. Secondly, it has been established that the modified U-Net system can perform breast area segmentation with over 0.92 Dice coefficients, and also that the effect can be further improved by the use of hybrid Transformer techniques to 0.94 [5, 23]. Finally, it should be noted that automated reporting algorithm development for chest x-rays is in progress, and even though the accuracy level is on the rise, there is one system (AMRG) that already works for mammograms [33, 43, 44].

The comparative overview of the major studies examined in this chapter can be found in Table 2.1. It shows differences and similarities in the methods used, nature of the data employed, results achieved, and the main limitations of each previous study. Finally, it states how the previous work is related to the current 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

While the progress seen in these studies are great, some gaps remain unaddressed: To start with, there is no existing system that combines all the three components of mammogram classification, segmentation and report generation into a single integrated pipeline. Classification research reaches up to obtaining the results in terms of labels. Segmentation research ends with obtaining the results as masks. Research in report generation is either targeting the chest X-rays or requires full scale training on paired data. In the clinical workflow, all three components are needed since one has to find the finding, segregate it and describe it.

Secondly, prompt engineering has not been implemented for the generation of mammography-specific reports. Whereas prompt engineering has worked in areas like medical QA and chest X-ray reading, there has been no research work designed for mammography reports regarding cancer type, stage, BI-RADS grade, and management recommendations.

Another interesting point is the fact that although MedGemma has been evaluated through benchmarks regarding its text performance and clinical reasoning, we can not be sure about its performance regarding mammography image assessment and reporting through prompt engineering.

Moreover, the current report-generation systems require the model to be re-trained based on paired training sets that connect mammograms to radiologists’

conclusions. This process is resource-consuming and can lead to the loss of some other medical knowledge. In contrast, prompt engineering does not require such costs.

The gaps stated above present the basis of this dissertation's contribution, which is the two-stop pipeline that uses deep learning for the recognition and segmentation of abnormalities and employs MedGemma for the generation of clinical reports based on these findings through prompt engineering.

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.compbiomed.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