Hybrid Generative-Supervised AI Solution for Enhanced Breast Tumor Analysis and Clinical Decision Support

November 2024

Abstract

Generative AI is a novel subset of artificial intelligence that learns to generate new material by analysing patterns within existing datasets, enabling it to produce original outputs that reflect a similar distribution to the training data. Diffusion models have gathered special attention because of their novel generating methodology, superior quality of generated images, and comparatively simpler training process. Still, the application of such models in the medical field remains at an early stage. This paper introduces a novel hybrid generative AI model for breast cancer detection, utilizing text-to-image diffusion and diffusion inpainting pipelines to create synthetic mammograms based on specific conditions provided as text prompts. This study examines two primary tasks. (1) Generating mammograms with diverse views, breast densities and BIRADS Levels derived from textual descriptions. (2) Creating synthetic abnormality-like lesions on mammograms through diffusion inpainting, directed by text prompts and masks. Synthetic images are generated to augment the original dataset which is then utilised to train classifiers for abnormality classification and BI-RADS classification.

1 INTRODUCTION

Breast Cancer can be defined as the Uncontrolled proliferation of cells within breast tissue. The unregulated proliferation of cells can lead to the formation of tumours, which may spread throughout the body and pose a lethal risk. Breast cancer ranks as the most common cancer among women worldwide. About 13% of women, or 1 in 8, will receive a diagnosis of invasive breast cancer in their lifetime, with a median age of diagnosis at 62 years. In 2022, breast cancer was the most common cancer among women in 157 of 185 countries, with 2.3 million women diagnosed and 670,000 deaths globally.

Alessandro Carriero discussed this in [Carriero et al., 2024] despite advancements in medical technology that improve the identification and prognosis of breast cancer, the increasing workload and the risk of false positives and negatives have created a need for additional techniques to enhance diagnostic accuracy. This critical task is challenging due to its significant requirement for attention and time, which is vital for ensuring that no subtle details are missed in the daily analysis of high-resolution photographs. Exceptional concentration and consistent performance are critical competencies for medical imaging specialists; however, even the most skilled human operators are ultimately limited by factors such as fatigue, biases, and distractions.

1.1 BI-RADS Assessment & Abnormality Findings

BI-RADS(Breast Imaging Reporting and Data System) offers a standardised framework for radiologists to articulate mammogram findings and evaluate breast cancer risk. The system includes categories from 0(incomplete) to 6(confirmed biopsy-proven malignancy). Numerous studies have employed deep learning models to autonomously categorize mammographic images into BIRADS density classifications. The models are trained on extensive datasets of mammograms that are annotated with BIRADS assessments, enabling them to learn the association between image features and the corresponding BI-RADS categories.

In addition to the BI-RADS assessment, a model for improved mammography analysis must also identify specific findings in mammograms, including masses, calcifications, and architectural distortions[Carriero et al., 2024]. This necessitates the model to classify the mammogram comprehensively while also localizing and characterizing any existing abnormalities. Various AI methodologies, such as object detection and segmentation models, may be utilized for this objective. Object detection models identify the presence and location of specific objects (e.g., a mass) within an image, whereas segmentation models delineate the precise boundaries of those objects, offering detailed information regarding their shape and size.

1.2 Benefits and Limitations of AI Approaches in Breast Cancer Detection

AI algorithms have demonstrated enhanced capabilities in improving breast cancer detection through mammography, Neural networks aid in identifying patterns within breast tissue images, enabling the early detection of potential signs, However, the extent of their impact on long-term risk prediction for advanced and interval cancers remains unclear[Abhilasha, 2024] Comprehensive, well-annotated datasets are crucial for the effective training of robust AI models. Collecting and annotating datasets for mammography analysis incurs significant costs and requires considerable time investment[Abhilasha, 2024]. Furthermore there is a lack of generalizability and reproducibility; A significant number of studies utilize racially homogeneous datasets or images sourced from a single vendor and location, This restricts the applicability of the developed models to various populations and clinical environments. The efficacy of AI models can differ among various healthcare institutions, influenced by demographic variations and the types of scanning equipment utilised[Gastounioti et al., 2022]. Mammograms possess high resolution, necessitating significant computational resources for processing. Reducing image resolution to alleviate computational demands may lead to the omission of nuanced features critical for risk evaluation[Gastounioti et al., 2022].

1.3 Generative AI for Breast Cancer Detection

Diffusion models are being presented as a plausible alternative for data augmentation[Montoya-del Angel et al., 2024], these models function by incrementally introducing noise to data and subsequently learning to invert this process, thereby generating new data samples from random noise. This methodology presents multiple benefits such as High-Quality Image Generation and Stable training process. Dorjsembe et al.[Khader et al., 2023] proposed the application of denoising diffusion probabilistic models (DDPM) [Ho et al., 2020a], originally designed for computer vision, to generate high-quality MRI images of brain tumors. The initial application of diffusion models to 3D medical images achieved state-of-the-art results, surpassing baseline models that utilize 3D GANs. Recent developments in the field have resulted in latent diffusion [Rombach et al., 2022], which incorporates a latent space to achieve enhanced image resolution. Pinaya et al. [Pinaya et al., 2022] employed latent diffusion to produce high-resolution 3D brain images, enhancing the image resolution from 64 × 64 × 64 to 160 × 224 × 160 without necessitating additional GPU memory or extended training time. The Fréchet inception distance (FID) for image fidelity and the multi-scale structural similarity index measure (MS-SSIM) for generation diversity were calculated, with DM exceeding the baseline metrics of GANs in both instances. Enhancing the generation process can be accomplished by incorporating supplementary input during both training and inference phases. An example of this is stable diffusion (SD) [Rombach et al., 2022], a conditional diffusion model that utilizes text prompts for generation conditioning. Chambon et al. [Chambon et al., 2022] introduced an SD implementation for medical images, proposing a model for the generation of chest X-rays.

1.4 Our Proposal

This paper provides a hybrid generative AI model for breast cancer detection that uses text-to-image diffusion and diffusion inpainting pipelines to build synthetic mammograms based on text prompts. This study focuses on two major goals. Using textual descriptions, generate mammograms with various views, breast densities, and BIRADS levels. Using diffusion inpainting, text instructions, and masks to create synthetic lesions that resemble mammography abnormalities. Synthetic pictures are created to supplement the original dataset, which is then used to train classifiers for anomaly detection and BI-RADS classification.

Our first step is to use YOLO to detect abnormalities and generate ROI models for mammograms. Next, we use Stable Diffusion (SD) for data augmentation. Stable Diffusion allows us to control the creation process by feeding more data into the Diffusion model. We want to create specific abnormalities (microcalcifications, masses, etc.) or BIRADS, therefore we modify the SD model accordingly. This allows us to manage the output and ensure that the SD generates data of the desired quality. Finally, our discriminative component, a convolutional neural network (CNN), is trained for classification tasks such as breast birads assessment and anomaly identification using the expanded dataset, which includes both authentic and synthetic mammograms. By supplementing the original dataset with synthetic photos, the model can generalize to a greater range of scenarios, including those that were previously underrepresented.

Our contributions are:

1. Facilitating the augmentation of datasets by incorporating a broader spectrum of abnormalities, including rare or subtle legions that are frequently underrepresented in actual datasets, resulting in more robust and reliable AI systems for breast cancer detection

2. Addressing the limitations of data scarcity and diversity, while also enhancing risk stratification, with the generation of mammograms with diverse BI-RADS levels, resulting in more tailored screening approaches reducing unnecessary biopsies and anxiety for women at low risk

2 Materials and Methods

2.1 Datasets

To ensure that various patient groups and mammography unit vendors were taken into account, we opted to train our models using two datasets. Table 1 provides a synopsis of the case distribution

Table 1: Distribution of cases for both datasets

	Emory	**VinDR**	**Combined**
Healthy	12,693	18,232	30, 925
With Lesion	2,456	2,254	4,710
Total	15,149	20,486	35,635

2.1.1 EMORY

We used a subset of the Emory Breast Imaging Dataset (EMBED) hosted on the AWS Open Data Program. EMBED contains 364,000 screening and diagnostic mammographic exams for 110,000 patients from four hospitals over an 8-year period.

The dataset was composed of expert annotations, including the coordinates of a bounding box surrounding various lesions , are included in the dataset mammograms, which also include both lesions and non-lesions (masses of type benign, malignant, and tumors).

2.1.2 VinDR Mammo

A second dataset was used, which contained around 20,000 FFDM pictures with breast-level evaluation and full lesion labeling. The VinDr-Mammo dataset is a large-scale evaluation of computer-aided diagnosis in full-field digital mammography. It contains 20,000 DICOM pictures taken from 5,000 mammography tests at two major Vietnamese hospitals. Each test contains four standard views (CC and MLO in both lateralities) [Nguyen et al., 2023].

2.2 Data Preprocessing and Preparation

The preprocessing and preparation strategies employed for both datasets were identical. Mammograms were first stored as PNG files to optimize disk space and enhance access speed. Secondly, the images were stored in RGB format, duplicating the original gray channel across all RGB channels, thereby enabling the application of pretrained weights. The image intensities, initially represented in uint16 data types, were scaled to a range of [0, 255] using a reduced uint8 data type. Also as illustrated in Figure 2, the images were reduced to a 512 × 512 square by bilinear interpolation and center cropping to leverage the pretrained weights available for SD. Images exhibiting right laterality were horizontally inverted to guarantee that the breast region in each image originated from the same side.

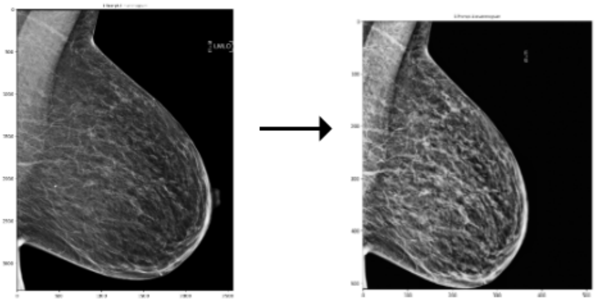


Figure 1: Resizing and Cropping of an Emory Mammogram. The same process was conducted for VinDR mammograms

2.3 Abnormality Detection and ROI Generation Models for Mammograms

Why did we chose YOLOv8?

YOLOv8 leverages the advantages of earlier YOLO iterations through significant architecture and training improvements that augment its velocity, precision, and adaptability in object identification. It comprises three primary components:

1. **Backbone:** A convolutional neural network (CNN) backbone, potentially an improved CSPDarknet, that extracts multi-scale features, encompassing both low-level and high-level representations vital for accurate object detection[Reis et al., 2024].

2. **Neck**: Employs an enhanced Path Aggregation Network (PANet) to refine and integrate characteristics across several scales, facilitating the detection of objects of diverse sizes while maximizing memory and processing efficiency[Terven et al., 2023].

3. **Head**: An anchor-free component that predicts bounding box coordinates, object scores, and class labels, reducing complexity, and addressing objects with varying aspect ratios and scales.

Key Features and Innovations

Anchor-Free Bounding Box Prediction: Eliminates the necessity for predetermined anchors, thereby streamlining the model, decreasing computational demands, and enhancing efficiency.

Loss Function Components:

Focal Loss to enhance management of class imbalance.

IoU Loss for accurate bounding box localization.

Objectness Loss emphasizes likely object regions inside photos.

Mixed Precision Training: Accelerates training and inference, diminishes memory usage, and accommodates bigger batch sizes on compatible GPUs.

Optimized CSP Backbone and Layer Aggregation: Minimizes duplication and enhances feature reutilization through an improved FPN, resulting in superior detection efficacy.

Enhanced PANet Neck: Optimizes multi-scale feature flow, augmenting detection capabilities for small or densely clustered objects and achieving superior object detection performance[Yaseen, 2024].

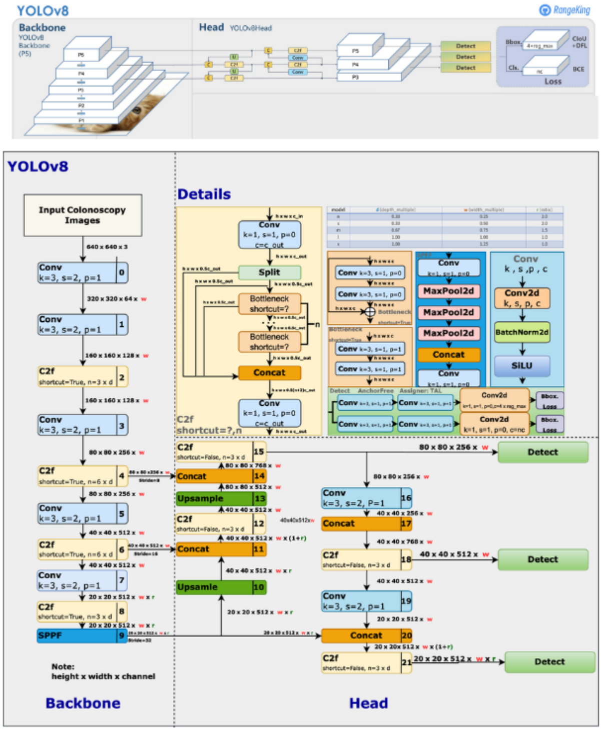


Figure 2: Yolov8 Architecture

2.4 Generative Model for Data Augmentation: Diffusion Models

Diffusion models are inspired by non-equilibrium thermodynamics. They define a Markov chain of diffusion steps to slowly add random noise to data and then learn to reverse the diffusion process to construct desired data samples from the noise. Unlike VAE or flow models, diffusion models are learned with a fixed procedure and the latent variable has high dimensionality (same as the original data). Several diffusion-based generative models have been proposed with similar ideas underneath, including diffusion probabilistic models [Sohl-Dickstein et al., 2015], noise-conditioned score network[Song and Ermon, 2020] , and denoising diffusion probabilistic models[Ho et al., 2020b].

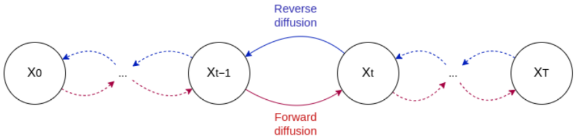


Figure 3: Forward and Reverse Diffusion Process

The Forward Process

The forward process involves sampling noise from a normal Gaussian distribution which is subsequently added to a sample image incrementally at each timestep of the Markov chain. In mathematical terms, consider an image x_0 and a noise ϵ, such that $\epsilon \sim N(\mu, I)$, For a Markov chain that operates over T timesteps, from $t = 0$ to $t = T - 1$, then the forward process q can be defined as:

$$q\left(x_{t+1} \mid x_t\right) = N\left(x_{t+1}; \left(\sqrt{1-\beta_{t+1}}\right) x_t, \beta_{t+1} I\right)$$

$q\left(x_{1:T} \mid x_0\right) = \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right)$ gradually adds noise to an initial sample x_0 from the data distribution $q\left(x_0\right)$ sampling noise from the predefined distributions $q\left(x_t \mid x_{t-1}\right)$ with variances $\{\beta_1, \ldots, \beta_T\}$.

where: β is a sequence of scheduled values defined by two hyper-parameters, $\beta_{\min}$ and $\beta_{\max}$. The values in the β sequence are representative of each timestep within the markov chain

Let us consider a Markov chain that operates over a span of 3 timesteps: at $t = 0, x = x_0$ (The original image) at $t = 1, x = x_1 = q\left(x_1 \mid x_0\right) = N\left(x_1; \left(\sqrt{1-\beta_1}\right) x_0, \beta_1 I\right)$ at $t = 2, x = x_2 = q\left(x_2 \mid x_1\right) = N\left(x_2; \left(\sqrt{1-\beta_2}\right) x_1, \beta_2 I\right)$ at $t = 3, x = x_3 = q\left(x_3 \mid x_2\right) = N\left(x_3; \left(\sqrt{1-\beta_3}\right) x_2, \beta_3 I\right)$ We can note that the subsequent timestep relies solely on the preceding timestep, with no influence from earlier timesteps. This characteristic defines the process as a Markov chain.

The generation of β values can occur through linear or cosine scheduling, transitioning from β_{min} to β_{max}.

For linear scheduling, β_t is given as:

$$\beta_t = \beta_{\text{min}} + t\frac{\beta_{\text{max}} - \beta_{\text{min}}}{T}$$

and the cosine scheduling will be given as:

$$\beta_t = \beta_{\text{min}} + 0.5\left(\beta_{\text{max}} - \beta_{\text{min}}\right)\left(1 - \cos\frac{t}{T}\pi\right)$$

Where $t \sim \text{Uniform}(0, T-1)$

A nice property of the above process is that we can sample x_t at any arbitrary time step t in a closed form using reparameterization trick. we introduce a term α, where $\alpha = 1 - \beta$. Now we can express the forward process in terms of α like so:

$$q\left(x_{t+1} \mid x_t\right) = N\left(x_{t+1}; \left(\sqrt{\alpha_{t+1}}\right) x_t, \left(1 - \alpha_{t+1}\right) I\right)$$

We can derive a term $\hat{\alpha}$, where $\hat{\alpha}$ is a sequence of the cummulative product of the terms in $\{\alpha_0, \alpha_1, \alpha_2, \alpha_3, \ldots \alpha_T\}$,

$$\hat{\alpha} = \prod_{t=0}^{T-1} \alpha_t$$

With this, the forward diffusion sampling process can be expressed by

$$x_t := \sqrt{\hat{\alpha}_t}.x_0 + \sqrt{1 - \hat{\alpha}_t} \cdot \epsilon$$

The Reverse Diffusion Process

Reversing the diffusion process to sample from $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$, we will be able to recreate the true sample from a Gaussian noise input, $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. It is important to note that β_t is small enough, $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ will also be Gaussian. Unfortunately, we cannot easily estimate $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ because it needs to use the entire dataset and therefore we need to learn a model p_θ to approximate these conditional probabilities which is essential for executing the reverse diffusion process.

$$p_\theta\left(\mathbf{x}_{0:T}\right) = p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) \quad p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right)$$

$p\left(x_{0:T}\right) = \prod_{t=1}^{T} p\left(x_{t-1} \mid x_t\right)$ gradually denoises a latent variable $x_T \sim q\left(x_T\right)$ and allows generating new data samples from $q\left(x_0\right)$.

Distributions $p\left(x_{t-1} \mid x_t\right)$ are typically not known and are estimated using a neural network with parameters θ. These parameters are learned from the data by optimizing a variational lower bound:

$$\log q\left(x_0\right) \geq E_{q\left(x_0\right)}[\underbrace{\log p_\theta\left(x_0 \mid x_1\right)}_{L_0} - \underbrace{KL\left(q\left(x_T \mid x_0\right) \mid q\left(x_T\right)\right)}_{L_T} - \sum_{t=2}^{T} \underbrace{KL\left(q\left(x_{t-1} \mid x_t, x_0\right) \mid p_\theta\left(x_{t-1} \mid x_t\right)\right)}_{L_t}]$$

Gaussian diffusion models operate in continuous spaces ($x_t \in R^n$) where forward and reverse processes are defined by Gaussian distributions:

$$\begin{aligned} & q\left(x_t \mid x_{t-1}\right) := \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right) \\ & q\left(x_T\right) := \mathcal{N}\left(x_T; 0, I\right) \\ & p_\theta\left(x_{t-1} \mid x_t\right) := \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \end{aligned}$$

Ho et al. [Ho et al., 2020b] suggest using diagonal $\Sigma_\theta\left(x_t, t\right)$ with a constant σ_t and computing $\mu_\theta\left(x_t, t\right)$ as a function of x_t and $\epsilon_\theta\left(x_t, t\right)$:

$$\mu_\theta\left(x_t, t\right)=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta\left(x_t, t\right)\right)$$

where $\alpha_t:=1-\beta_t, \bar{\alpha}_t:=\prod_{i \leq t} \alpha_i$ and $\epsilon_\theta\left(x_t, t\right)$ predicts a "groundtruth" noise component ϵ for the noisy data sample x_t. In practice, the objective (1) can be simplified to the sum of mean-squared errors between $\epsilon_\theta\left(x_t, t\right)$ and ϵ over all timesteps t :

$$L_t^{\text{simple}}=E_{x_0, \epsilon, t}\left\|\epsilon-\epsilon_\theta\left(x_t, t\right)\right\|_2^2$$

Figure 44 illustrates the components of the denoising UNet employed in the reverse process. The UNet output is the noise, ϵ_θ, which must be eliminated from the input image to approximately restore the original noise-free image[[Montoya-del Angel et al., 2024]]

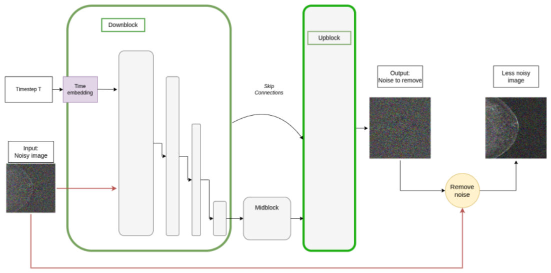


Figure 6. Reverse diffusion process using a denoising UNet. The upblock layers are a mirror of the downblock layers.

Figure 4: Reverse diffusion process using a denoising UNet. The upblock layers are a mirror of the downblock layers.

Latent Variable Space

The Latent Diffusion Model (LDM), introduced by Rombach et al [Rombach et al., 2022], conducts the diffusion process in a compressed latent space instead of the pixel space. This approach minimizes computational expenses and enhances inference speed. The procedure entails:

1. Perceptual Compression: An autoencoder compresses the input image $\mathbf{x}$ into a latent representation $\mathbf{z}=\mathcal{E}(\mathbf{x})$. The decoder $\mathcal{D}$ can then reconstruct the image from $\mathbf{z}$.

2. Diffusion in Latent Space: The diffusion model operates on the latent representation $\mathbf{z}$, allowing for manipulation and generation of images based on semantic content.

Regularization techniques, including KL divergence regularization (KL-reg) and vector quantization (VQ-reg), are utilized in autoencoder training to ensure stability and mitigate high variance in the latent space.

StableDiffusion Model: By conducting the diffusion process in a compressed latent space, StableDiffusion achieves efficient text-to-image generation with improved semantic control. Classifier-free guidance[Ho and Salimans, 2022] is employed for more adaptable conditioning, providing balanced fidelity and diversity without relying on an external classifier, making StableDiffusion more scalable and versatile for generating high-quality, contextually aligned images based on input text prompts.

Synthetic Image Generation using Stable Diffusion

Phase One:

We selected the Dreambooth technique to fine-tune a diffusion model at a foundational level, aiming to accurately capture the distinct structural and visual characteristics of mammograms. The main objective of fine-tuning the base diffusion model with the DreamBooth technique on the target mammogram dataset is to equip the model with essential prior knowledge and representations pertinent to the specific class of mammograms. With the aim of using the class-specific prior preservation loss for subsequent finetuning, This initial fine-tuning was crucial, as diffusion models, despite their advantages from extensive and varied training data, do not possess the necessary representational capacity to effectively model specialized medical images, like mammograms, without focused pre-training.

The class-specific prior preservation loss is essential as it mitigates overfitting and promotes diverse sample generation, depending on the model's accurate initial understanding of the target class. The effectiveness of the prior preservation loss in a base diffusion model is contingent upon the presence of suitable class-level representations; without them, the model will face challenges in producing plausible and diverse samples for the target class[Ruiz et al., 2023].

In the absence of an initial fine-tuning step to internalize unique features, the model's prior-preservation mechanisms, intended to uphold subject diversity and fidelity, would prove ineffective, as the model would not possess a solid prior understanding of mammographic structures. Fine-tuning the base model on the mammogram dataset with DreamBooth enables the model to learn the specific visual characteristics, patterns, and representations critical for accurately identifying the mammogram class. This fine-tuned model provides a robust basis for the application of the class-specific prior preservation loss, enhancing the model's ability to retain class-level knowledge and generate diverse, high-quality mammogram samples during the fine-tuning stage.

Following the same training procedure proposed in [Montoya-del Angel et al., 2024], The VAE weights were frozen, allowing only the CLIP text encoder and UNet weights to be trained. We trained our model with a combination of both ViniDR and Emory dataset.

Phase 2

Training a Diffusion Model for each BIRADS Class, We train a unique model for each unique BIRADS Levels, allowing us to capture the intricacies of each BIRADS Levels, we found this to be more effective than an overall training of the entire dataset mixed with all BIRADS Levels. We finetune our foundational base model using two methods. A) Training with Low-Rank Adaptation of Large Language Models (LoRA) and prior preservation loss

Low-Rank Adaption of Large Language Models was first introduced by Microsoft in [Hu et al., 2021] LoRA allows to adapt pretrained models by adding pairs of rank-decomposition matrices to existing weights and only training those newly added weights. This has a couple of advantages:

Previous pretrained weights are kept frozen so that the model is not prone to catastrophic forgetting Rank-decomposition matrices have significantly fewer parameters than the original model, which means that trained LoRA weights are easily portable. LoRA attention layers allow to control to which extent the model is adapted towards new training images via a scale parameter.

When using LoRA we can use a much higher learning rate compared to vanilla dreambooth. Here we use 1e-4 instead of the usual 2e-6. We also train additional LoRA layers for the text encoder.

B) Training with Dreambooth and prior preservation loss

We use a unique instance prompt for each image, instead of a generic instance prompt, because we want to condition our diffusion model on the breast density and birads level of each mammogram.

Phase 3:

The data flow from the Region of Interest (ROI) Generator to the Inpainting Pipeline is essential in our hybrid generative-supervised AI model for breast cancer detection, facilitating accurate detection and augmentation of mammographic abnormalities. The initial phase of this data flow entails identifying Regions of Interest (ROIs) within the mammogram images. The YOLOv8 model, enables precise detection of abnormalities including masses, calcifications, and asymmetries. The mAP@50 score, which measures mean average precision at an IoU threshold of 50%, serves as a critical metric for assessing the performance of the YOLOv8 model. The improved detection capability is essential, as it guarantees that the generated ROIs align with precisely localised areas of the mammogram that are likely to exhibit abnormalities. Upon detection of abnormalities, the model produces bounding boxes that delineate the regions of interest within the image. The ROIs are extracted and subsequently input into the next phase of the pipeline, which is the inpainting process.

Inpainting of abnormality lesions, Similar to phase 2, We train a unique diffusion inpainting model for each unique Abnormality Lesion, allowing us to capture the intricacies of each abnormality lesion, we found this to be more effective too, Following the same training procedure proposed in [Montoya-del Angel et al., 2024], During the training phase, for each mammogram containing a lesion, two new components are introduced per example: the mask and a masked version of the original image. The masked version of the original image is a replica in which the pixel values within the bounding box are assigned a value of zero. During training, both the image and the masked image are initially encoded into the latent space utilizing the VAE encoder. The mask is subsequently adjusted to correspond with the latent representation size of the images. This procedure is conducted for each dataset, resulting in the training of each unique inpainting models, one corresponding to each abnormality lesion.

During inference, unlike a standard SD inference pipeline, two additional inputs are required: an image onto which the lesion will be inpainted and a mask indicating the designated region for inpainting.

```
The main training hyperparameters explored throughout all these phases are:

pretrained_model: stable-diffusion-v1-5/stable-diffusion-v1-5
resolution: 512
instance_prompt: "A mammogram"
batch_size: 4, 8, 16
learning_rate: 1e-6, 2e-6, 1e-5
train_text_encoder: True
max_train_steps: 1k - 20k
validation_prompt: "A mammogram"
num_validation_images: 4
validation_steps: 500
```

2.5 Multiview classification using EfficientNet

Deep Learning-based BI-RADS Classification and Abnormality-based Classification.

Classification is performed via a pretrained EfficientNet model. EfficientNet refers to convolutional neural networks that employ a compound scaling technique to optimize model depth, width, and input resolution, hence attaining high accuracy while utilizing resources efficiently.

EfficientNet's architecture is founded on the inverted bottleneck residual blocks of MobileNetV2, augmented with squeeze-and-excitation blocks. This combination enables the model to discern complex patterns while preserving computational efficiency[Tan and Le, 2020].

Multi-view Feature Extraction

The EfficientNet Model was employed for feature extraction, with the fully connected layer at the end of the architecture removed to obtain hidden representations from the input photos. The concealed representation is a tensor with dimensions C × H × W that has been down-sampled from the original dimensions. Specifically, we used EfficientNet-B0 for the feature extraction step. To train EfficientNetB0 extractor, all input images were resized to H × W = 224 × 224. Furthermore, each mammography has been replicated from the original grayscale channel into three channels. Upon processing via the trained extractor, we acquired the feature tensor with dimensions C × H × W = 1280 × 7 × 7.. Finally, this hidden representation was subjected to average pooling across the spatial dimensions to provide a 1280-dimensional vector. Following the Feature extractor layer, a weight component in the form of a linear layer is applied to the MLO-view feature dimensional vector. This is subsequently concatenated along the same channel dimension, after which the concatenated features are processed through a linear layer for averaging to obtain the predicted classes. The primary concept of multi-view architecture was to integrate multiple concealed data from four perspectives to build an effective classifier. The EfficientNet backbones during the feature extraction phase are trained independently for each dataset view. In this combinatorial technique, we computed the average of the L-CC and LMLO feature vectors, as well as the R-CC and R-MLO features.

Data Preprocessing We extracted breast regions to ensure a uniform scale across photos while preventing distortions by implementing the following measures: :

- Defining a target aspect ratio and a consistent target size for the breast region.
- Ensuring that the extracted breast region fits within a predefined box size, which includes padding to add space around the breast region.
- Use padding instead of direct resizing to maintain the aspect ratio and avoid distortions.

We pad only on one side (where the image intensity is lower)

Other Data Preprocessing methods can be viewed in Figure 5

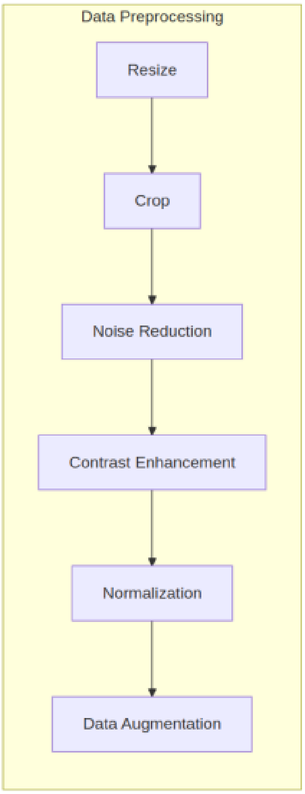


Figure 5: Data Preprocessing pipeline

Data Augmentation

Let $\mathbf{x}_{\text{orig}}$ represent the original image, and $\mathbf{x}_{\text{synth}}$ represent a synthetic image generated by the pre-trained diffusion model. The data augmentation process is as follows:
1. We pre-generate a fixed set of n synthetic images, denoted as $\{\mathbf{x}^{(1)}_{\text{synth}}, \mathbf{x}^{(2)}_{\text{synth}}, \ldots, \mathbf{x}^{(n)}_{\text{synth}}\}$, using the pre-trained diffusion model, we then concatenate both original and synthetic images. but only concatenating with a certain proportion of synthetic images.
2. During the training phase, we keep track of the indices of the synthetic images in the batch, denoted as $\mathcal{I}_{\text{synth}}$.
3. After obtaining the model's predictions, we extract the predictions corresponding to the synthetic images, represented by $\hat{\mathbf{y}}_{\text{synth}} = \{\hat{\mathbf{y}}^{(i)}_{\text{synth}} \mid i \in \mathcal{I}_{\text{synth}}\}$.
4. To define the synthetic loss coefficient $\alpha(\mathbf{p})$ as a function of the proportion of synthetic images $\mathbf{p}$, we need a function that satisfies the following conditions:

1. $\alpha(p) < 1.0$ for all p, ensuring that the original data is always prioritized.
2. $\alpha(p)$ should increase with p, meaning that as the proportion of synthetic data increases, the influence of the synthetic data on the model's loss should also increase.
3. The rate of increase should be controlled in a reasonable manner, ensuring balance between the synthetic and original data.

We use a sigmoid-life function, but scaled and shifted to meet our requirements:

$$\alpha(p) = k * (\frac{p}{1+p})$$

Where:

- k is a constant controlling the steepness of the function's increase. This allows us to adjust how quickly $\alpha(p)$ increases as p grows.
- p is the proportion of synthetic images, ranging from 0 to 1.

For our specific cases:

- 25% synthetic: $\alpha \approx 0.16$
- 50% synthetic: $\alpha \approx 0.27$
- 75% synthetic: $\alpha \approx 0.34$
- 100% synthetic: $\alpha \approx 0.40$

5. The total loss is then calculated as a weighted sum of the loss on the original images and the loss on the synthetic images:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{orig}} + \lambda \mathcal{L}_{\text{synth}}$$

where λ is the weight assigned to the synthetic loss term, controlling the impact of synthetic images on the overall loss.

The overall pipeline for our system can be seen in Figure6

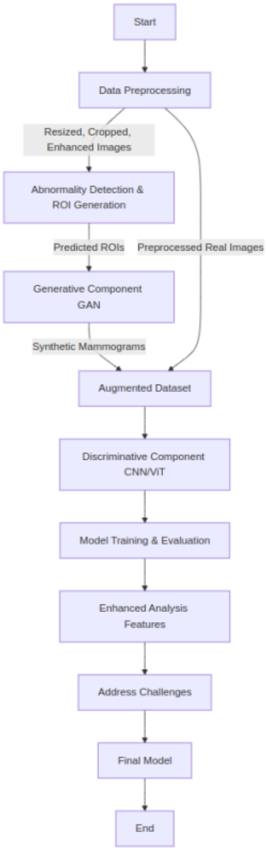

Start
Data Preprocessing
Resized, Cropped, Enhanced Images
Abnormality Detection & ROI Generation
Predicted ROIs
Preprocessed Real Images
Generative Component GAN
Synthetic Mammograms
Augmented Dataset
Discriminative Component CNN/ViT
Model Training & Evaluation
Enhanced Analysis Features
Address Challenges
Final Model
End

Figure 6: Overall System Pipeline

3 Results

Dataset Statistics

It is essential to emphasize our intention to implement data augmentation through the use of synthetic mammograms to address the imbalances in our dataset.

During our abnormality detection and ROI prediction procedure, we identified a comparable imbalance in the abnormalities within our mammography dataset. Consequently, we narrowed it to the three most prevalent categories: "MASS," "SUSPICIOUS CALCIFICATION," and "ASYMMETRY."

Phase 1

As Initially stated, our Synthetic Data Generation can be split into two parts.

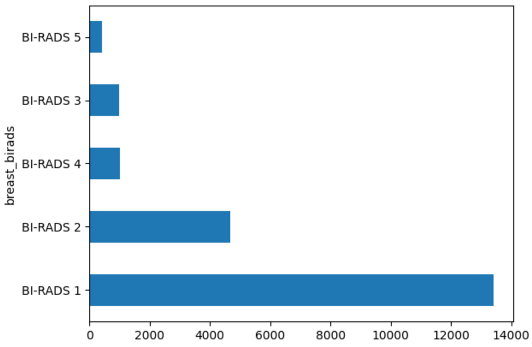


Figure 7: ViniDr Birads Distribution

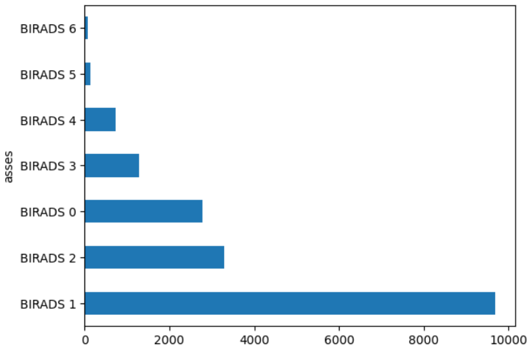


Figure 8: Emory Birads Distribution

Figure 9: Datasets Birads Distribution

In the first section we will generate several mammograms based on the textual cues. This is employed in the production of mammograms that align with specific Breast BIRADS Levels, Breast Density, or Vendor Type requirements.

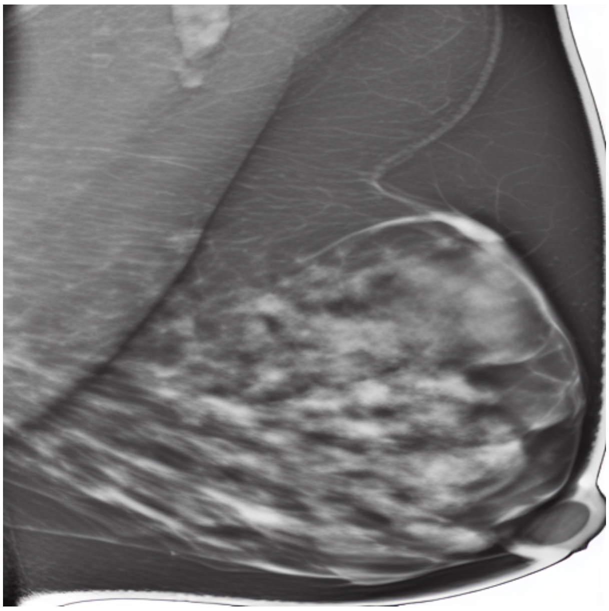
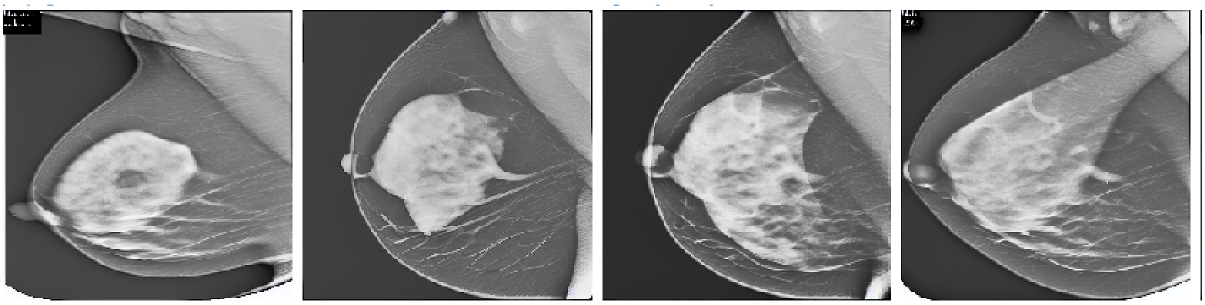


Figure 10: A synthetic mammogram



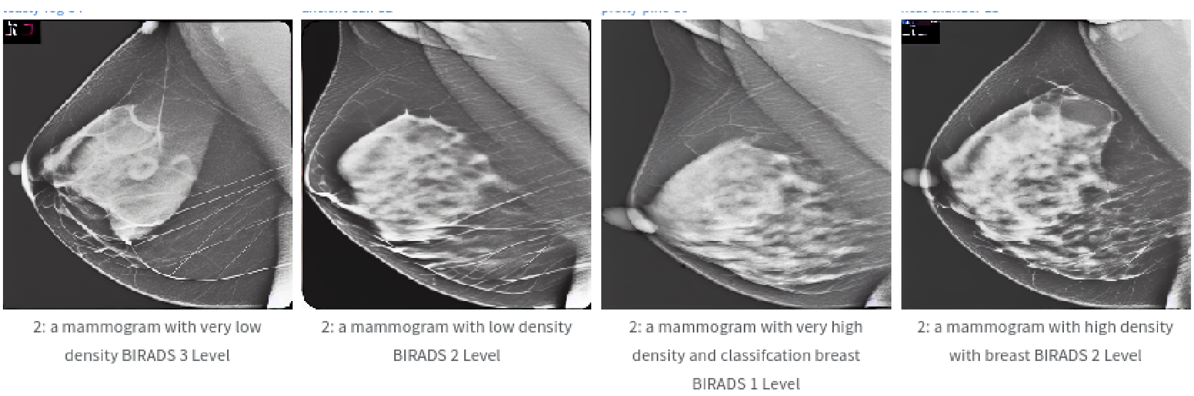
2: a mammogram with low density BIRADS 5 Level

2: a mammogram with very low density BIRADS 4 Level

2: a mammogram with high density BIRADS 3 Level

2: a mammogram with low density BIRADS 4 Level

Figure 11: Synthetic Mammograms of the various BIRADS Levels



2: a mammogram with very low density BIRADS 3 Level

2: a mammogram with low density BIRADS 2 Level

2: a mammogram with very high density and classifcation breast BIRADS 1 Level

2: a mammogram with high density with breast BIRADS 2 Level

Figure 12: Synthetic Mammograms of the various BIRADS Levels

In the second section, utilising a technique called "Model Inpainting," we can create breast abnormalities in specific locations by identifying potential areas of interest. Subsequently, we utilise these sites as masks to include the desired anomalies into our diffusion model.

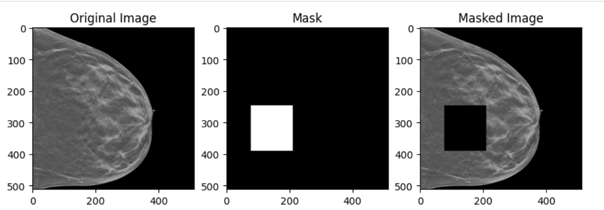


Figure 13: Inpainting of a Mammogram

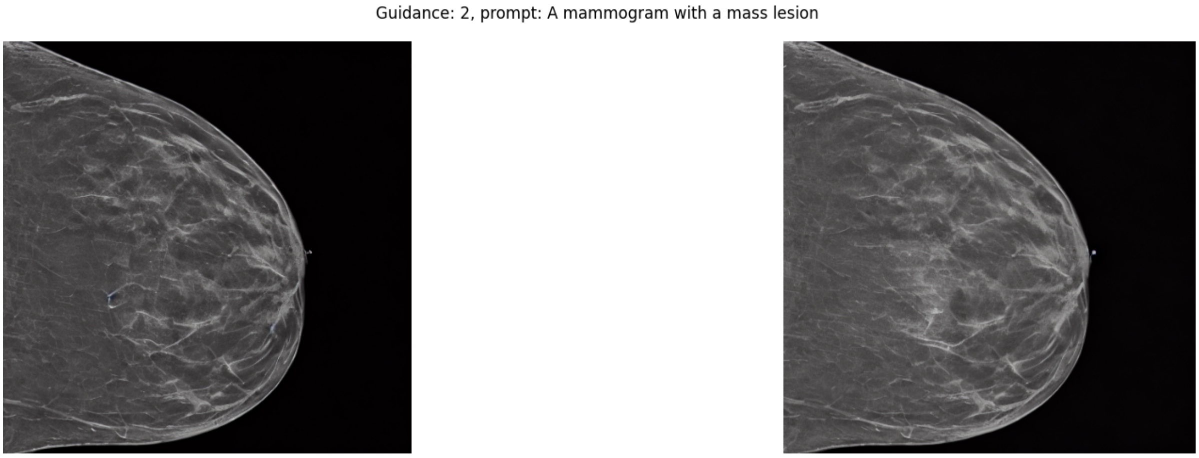


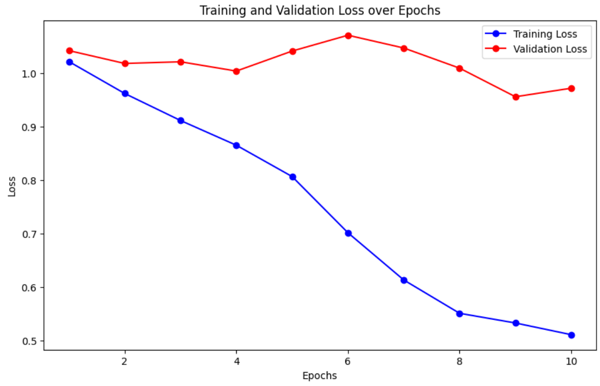
Figure 14: A sample Mammogram inpainted with a lesion of choice

To assess the impact of our synthetic data on the outcomes, we analyse many train-validation plots that incorporate both synthetic and actual data, as well as those containing solely real data. Furthermore, we illustrate the distance between training and validation as the quantity of synthetic data fluctuates throughout the procedure.

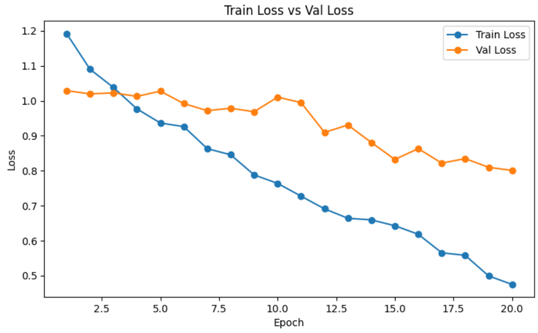
3.1 Abnormality classification using Multi-view Extractor CNNs

To improve our understanding of the the impact of synthetic data augmentation on abnormality classification, we employed a multi-view CNN extractor that accounts for spatial correlations across various views of the breast. The architecture facilitates simultaneous analysis of craniocaudal(CC) and mediolateral oblique (MLO) views, thereby enhancing feature representation and classification accuracy.

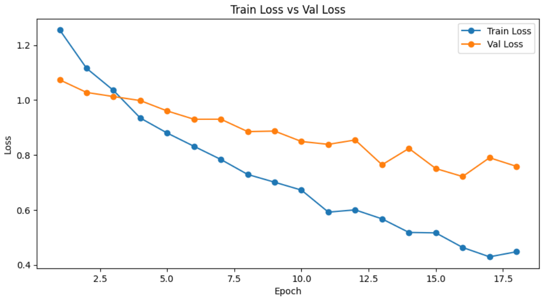
We adopted the same synthetic data proportions(0%, 0.25%, 0.5%, 0.75%, and 1.0%) as in the YOLOv8 detection experiments. The specified proportions facilitated a systematic assessment of the influence of synthetic data on multi-view classification performance. The model was trained for 20 epochs using a learning rate of 1e-4 and a batch size of 64. Performance was evaluated based on precision, recall, and the weighted F1-score for each class of abnormalities: ”MASS,” ”SUSPICIOUS CALCIFICATION,” and ”ASYMMETRY.”



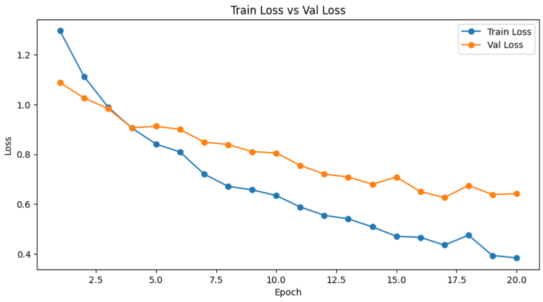
0%



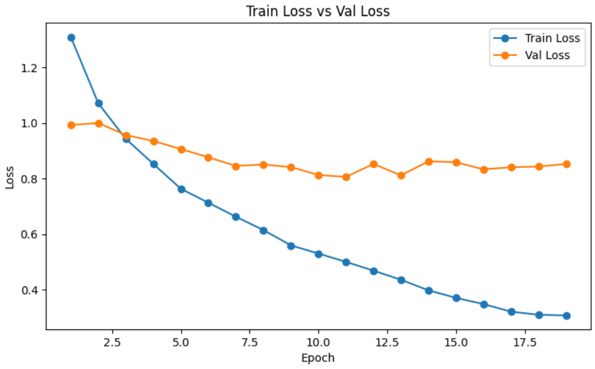
0.25%



0.5%



0.75%



1%

Findings and Observations

0% Synthetic Baseline Data: The training loss decreases consistently, albeit at a lower rate than that of synthetic data. In contrast to training loss, validation loss remains consistently elevated and exhibits minimal variation. The model exhibits inadequate learning in the absence of synthetic data, particularly concerning under-represented anomalies, leading to suboptimal generalisation.

Both 25% and 50% Synthetic data demonstrate a consistent reduction in training loss, the training loss consistently decreases across epochs, signifying effecting learning with the validation loss consistent and declining, however overfitting in training data remains a concern. At 75% Synthetic Data which is the optimal proportion, the ideal equilibrium is represented by validation loss, which decreases alongside training loss, results indicate that this fraction mitigates overfitting and enhances generalization. The model's remarkable drop in training loss with 100% synthetic data shows it can memorize patterns, where the validation loss stabilizes at a higher value and fluctuates, indicating model inability to generalize to real data. Although training loss is minimized, the complete reliance on synthetic data leads to a lack of diversity, which adversely affects generalisation.

Table 2: Evaluation Metrics Summary

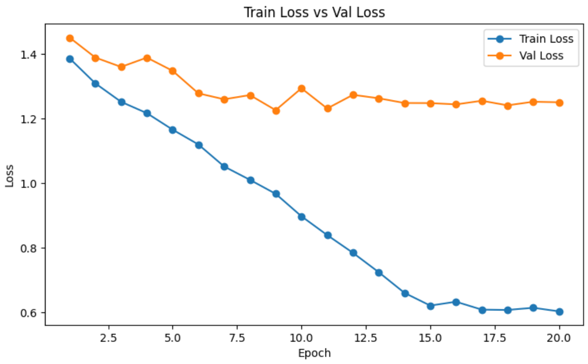
Synthetic Data proportion	A F1 Score	S-C F1Score	M F1-Score	Sensitivity	Specificity	AUC Score
0	0.42	0.56	0.67	0.55	0.62	0.68
25	0.63	0.7	0.78	0.8	0.7	0.73
50	0.71	0.79	0.84	0.82	0.78	0.78
75	0.79	0.81	0.85	0.85	0.84	0.83
100	0.79	0.78	0.85	0.84	0.80	0.80

Where A, S-C and M, represent the 3 abnormalities considered, Asymmetry, Suspicious-Calcification and Mass.. The F1-score for "Asymmetry" increased considerably from 0.42(0% synthetic data) to 0.79(100% synthetic data), showing that the model's ability to detect asymmetry is significantly improved by the use of synthetic data. The "Suspicious Calcification" and the "Mass" class showed the highest F1-Scores (0.81) and (0.85) at 75% synthetic data. This classes benefited the most from the introduction of synthetic data due to well-defined features.

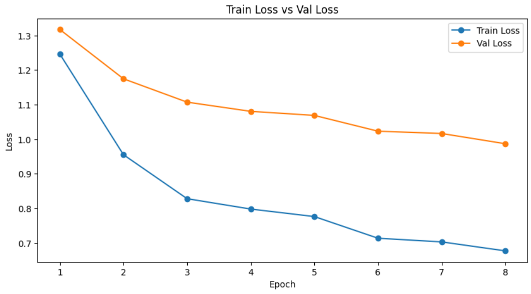
Because of the imbalance between the classes, the generalisability of the model is considerably limited at 0%. Through the addition of synthetic data at a 25% level, the generalisation of the model is boosted; still, the challenges that minority classes are encountering are not totally overcome. Because performance hits a high for specific classes around 50%, the efficacy of synthetic data is dependant on optimising the proportions. Performance reaches a plateau at 50%. The sweet spot for real and synthetic data gives considerable increases without resorting to overfitting artificial patterns, and it is located at 75%. Whenever we rely disproportionately on synthetic data, we fail to take into account the diversity that exists in the real world and may not be able to generalise as effectively.

3.2 Birads classification using Multiview Extractor CNNs

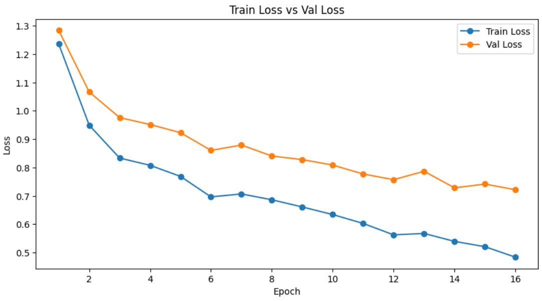
We also adopted the multi-view CNN extractor that accounts for spatial correlations across various views of the breast to improve our understanding of the the impact of synthetic data augmentation on BIRADS classification. Performance was evaluated based on precision, recall, and the weighted F1-score for four classes of BIRADS: "BIRADS1", "BIRADS2" "BIRADS3" and "BIRADS4".



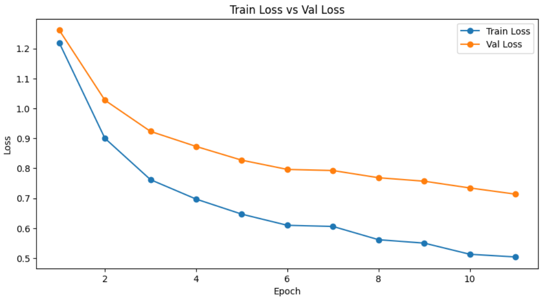
0%



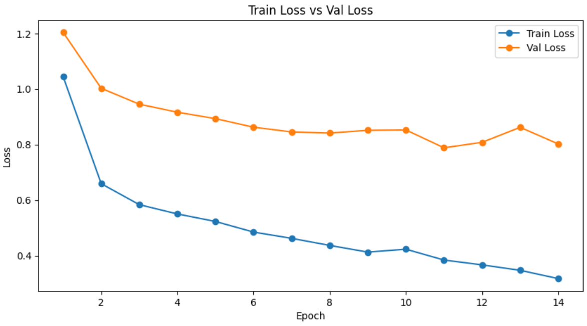
0.25%



0.5%



0.75%



1%

Findings and Observations

At 0% Synthetic Data, the training loss decreases gradually, indicating effective learning from the data, but there is still a noticeable gap between training and validation losses, the model is not generalizing fully, at 25% the model performs better than 0% synthetic data, but over-fitting persists, the gap between training and validation loss is still present, suggesting a problem of generalization on unseen data, at 50% , 75% the train loss curve starts much lower than the validation loss curve, with the validation loss smoothing out overtime indicating the model is generalizing much better, at 100%, the gap becomes more significant, at this proportion the model is likely memorizing the training data and not learning generalizable features.

Table 3: Evaluation Metrics Summary

Synthetic Data Proportion	1 F1 Score	2 F1 Score	3 F1 Score	4 F1 Score	5 F1 Score	Sensitivity	Specificity	AUC Score
0	0.53	0.41	0.52	0.59	0.32	0.46	0.44	0.47
25	0.65	0.47	0.58	0.64	0.44	0.53	0.49	0.56
50	0.73	0.68	0.70	0.69	0.48	0.60	0.66	0.66
75	0.67	0.58	0.68	0.76	0.51	0.65	0.6	0.64
100	0.68	0.65	0.57	0.65	0.47	0.59	0.56	0.6

Where the number ranging from 1-5, represent the BIRADS classses considered, from BIRADS 1 - BIRADS 5, The optimal performances for BIRADS 1, BIRADS 2, and BIRADS 3 are observed at 50% synthetic data, beyond which a slight decline in benefit is noted. This indicates that the incorporation of synthetic data enhances the model's capacity to identify BIRADS 1, 2, and 3 cases. Conversely, the best performance for BIRADS 4 & BIRADS 5 is achieved at 75% synthetic data. The optimal proportion of synthetic data differs among the various BIRADS classes, suggesting that a uniform approach may not be effective; however, the ideal proportion appears to be approximately 50-75%. The utilization of entirely synthetic data seems to negatively impact classification performance across all BIRADS categories.

Ensembling

Noting that the best model's performance occurs at both 50 and 75 percents, we sought to improve performance across both abnormality classification and BIRADS classifciation by ensembling the results of both models, using the logits gotten from both models, we find the average of the logits, and use that to get our final results using the max function. The ensemble integrated predictions from individual models trained with synthetic and real data proportions of 50-75%, as this range demonstrated optimal performance.

Table 4: Ensemble Metrics Summary

Results	Sensitivity	Specificity	AUC Score
Abnormalities	0.85	0.91	0.9
BIRADS	0.72	0.75	0.75

Recent research trained DL models on large screening cohorts of the general population using normal mammograms collected at least one year before breast cancer diagnosis or BIRADS 1 or 2 follow-up, These study designs better measure breast cancer risk by identifying high-risk patients before diagnosis. These models have shown promising results, with AUCs ranging from 0.60 to 0.84, often surpassing current breast cancer risk models. Dembrower et al [Dembrower et al., 2020] found that their FFDM-driven DL risk score beat auto-mated breast density measurements (odds ratios of 1.6 and 1.3, respectively). Yala et al [Yala et al., 2019] found that a mammographic DL risk score beat the clinically utilised Tyrer-Cuzick model (AUC of 0.68 vs. 0.62).

Table 5: Comparison with existing Literature

Study	No Images	Model Architecture	Model performance
Dembrower et al	150,502 images (1188 cases; 10,563 controls)	Inception-ResNet	AUC = 0.65
Yala et al	88,994 images (1821 cases; 38,284 controls)	ResNet-18	AUC = 0.68 for image only DL AUC = 0.70 for hybrid
Arefan et al	452 images (113 cases; 113 controls)	GoogleLeNet	AUC=0.68, CC AUC=0.60, MLO AUC=0.72 CC + MLO
Maleika Heenaye-Mamode Khan et al	9395 images(6775 from the CBIS-DDSM dataset)	ResNet50	AUC on Abnormalities = 0.87
Refat Khan et al	830	Mult-Headed CNN	AUC on risk assessment = 0.73
Ensemble Model	35635 images	EfficientNetB0	AUC on BIRADS Classification=0.75 AUC on Abnomality Classification=0.9

As seen in **Table 5**, while some studies report good AUC scores, our model shows a competitive performance in both abnormality detection and BIRADS classification, with the added advantage of handling false positives more effectively through the ensemble method.

3.3 Abnormality and ROI detection using YOLOv8

The Yolov8 model is trained for fifty epochs using an integrated dataset that includes both the Emory Dataset and the Vini Dataset. The image dimensions employed are 640 × 640, and the learning rate applied is 0.001.

There is a persistent decline in train box_loss, train cls_loss, and train dfl_loss, alongside a decrease of val box_loss, val cls_loss, and val dfl_loss. Our model effectively learns to capture bounding boxes and predict lesions in the images. For this Object Detection Model while further developments are required, we can observe good results with mAP@50 score in the range of 0.4-0.6 indicating significant potential for enhancement. The mass abnormality class is receiving the greatest support relative to other classes. Incorporating synthetic data into our dataset is an avenue we can explore to determine if it enhances the outcomes.

We generate a total of 6,000 synthetic images for both the VINI and EMORY datasets, with 3,000 images allocated to each abnormality class. A total of 6,000 synthetic images are distributed across the ”MASS,” ”SUSPICIOUS CALCIFICATION,” and ”ASYMMETRY” classes due to this practice being conducted twice. To sample two separate distributions from the EMORY and VINI data sets, the sample size is reduced by fifty percent. This enables us to sample various distributions. An illustration of a potential 25% data allocation is as follows: 12.5% would consist of synthetic EMORY data, while the remaining 12.5% would comprise synthetic VINI data.

The mAP@50 for detecting masses generally remains higher compared to other classes across all synthetic data proportions. We observed that the model finds it easier to learn patterns for masses, potentially due to better-defined features. The overall mAP@50 peaks at a synthetic image proportion of 0.5. This indicates that while synthetic data is helpful to a certain extent, excessive reliance on it (as seen at 1.0) may degrade performance, due to the synthetic data failing to fully replicate the complexity or diversity of real-world data. At 0.5 and 0.75, there appears to be a balance between the contribution of synthetic and original images.

Combined Sample Assessments

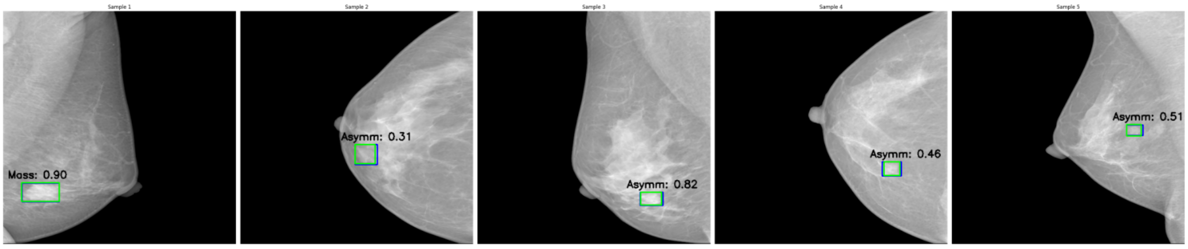


Figure 25: Sample Mammograms and Results

Table 5: Samples Assessment Results

Sample	target birad	predicted birad	predicted abnormality	target abnormality	predicted roi	target roi
1	BIRADS 4	BIRADS 2	Mass	Mass	(40, 377, 123, 418)	(41, 378, 123, 419)
2	BIRADS 4	BIRADS 4	Asymmetry	Asymmetry	(252, 293, 296, 335)	(252, 293, 301, 337)
3	BIRADS 3	BIRADS 1	Asymmetry	Asymmetry	(358, 398, 404, 426)	(357, 398, 407, 427)
4	BIRADS 4	BIRADS 4	Asymmetry	Asymmetry	(372, 331, 407, 361)	(368, 331, 410, 362)
5	BIRADS 4	BIRADS 4	Asymmetry	Asymmetry	(384, 250, 420, 273)	(384, 249, 416, 274)

The sample assessments demonstrate the potential of the clinical decision support system by accurately classifying abnormalities and BIRADS levels in most cases, the system is designed to assist in the decision-making processes of experts, identifying subtle lesions along with precise predictions of BI-RADS levels in cases 2,4 and 5, this shows potential to decrease diagnostic inaccuracies and enhance clinical evaluations. ROIs are identified to improve diagnostic accuracy, this method assists physicians in detecting concerning tumors and enhances evaluation efficiency by pinpointing problematic mammography locations. The program enhances diagnostic workflows and assists radiologists in prioritizing high-risk cases. The system successfully completed these examinations using training synthetic mammograms. The report posits that synthetic data addresses issues of data scarcity and unpredictability, thereby exposing the model to a broader range of imaging conditions and tumor characteristics. Incorporating synthetic images into the training set has the potential to enhance diagnostic accuracy and reliability. Combined sample assessments indicate that the hybrid generative-supervised AI model possesses significant potential for clinical decision assistance. Enhancing the training dataset and generalization through synthetic images enhances diagnostic accuracy, professional decision-making, and patient care.

4 Discussion

This study revealed that hybrid generative AI models, such as text-to-image diffusion with diffusion inpainting, are capable of detecting breast cancer. To mitigate the deficiencies in medical imaging databases and the issue of underrepresentation, it is essential to generate synthetic mammograms that encompass various BI-RADS levels and abnormalities. Our findings suggest that these models have the potential to transform clinical and academic environments. Synthetic data improved the representation of underrepresented anomalies, such as asymmetry and suspicious calcifications, thereby enhancing categorization and model generalization across diverse patient profiles and imaging conditions, the optimal combination of 50-75% synthetic and real-world data indicates that the variability of data aligns with the strengths of synthetic augmentation. This approach improves the categorization of anomalies, identification of lesions, diagnostic precision, and overall system performance.

Practical Applications and Impact

False positives in breast cancer diagnoses represent a significant clinical challenge, particularly in the context of AI-driven diagnostic tools. These errors may lead to unnecessary biopsies, increased patient anxiety, and elevated healthcare costs. This hybrid generative supervised AI approach employs diffusion based generative models alongside supervised classifiers to minimise false positives. Below are examples illustrating how our model addresses this issue:

- Generative data augmentation: Improved training data minimises false positives in AI systems. Conventional machine learning algorithms face challenges when dealing with imbalanced datasets that inadequately represent breast abnormalities such as microcalcifications and asymmetry. Abnormalities that are under-represented are difficult to identify, leading to the potential for models to inaccurately classify them as benign, Diffusion models facilitate generational data augmentation by producing synthetic mammograms for various breast disorders. Datasets are deficient in rare diseases and atypical breast tissue patterns, while diffusion-based synthetic data includes these elements. Incorporating synthetic samples into the training dataset enhances the AI model's exposure to various abnormalities, including subtle and rare lesions, thereby improving its capacity to differentiate between benign and malignant characteristics.

- High-Risk Class Threshold Optimisation: A notable strategy for reducing false positives is threshold optimisation. BI-RADS 4 suggests the presence of potentially malignant breast lesions requiring further assessment, whereas BI-RADS 2 denotes benign tumours. Excessive sensitivity to anomalies, including benign ones, may lead to misdiagnosis of low-risk malignancies. This issue is resolved by modifying the criteria for the high-risk class within our algorithm. The BI-RADS 4 decision boundary now categorises positive cases as having a higher likelihood of malignancy. The threshold is reduced for suggestive yet non-malignant lesions to minimise false positives. The model's thresholds differ, as lower-risk classes such as BI-RADS 2 or 3 require greater confidence to classify lesions as malignant. To minimise unnecessary biopsies and false positive results.

- Ensemble learning and multi-view validation: Craniocaudal (CC) and mediolateral oblique (MLO) mammograms demonstrate breast tissue characteristics. A singular perspective on minor abnormalities may result in incomplete or misinterpreted information. A lesion that is observable in one angle but not in another may lead to misclassifications and false positives. The model employs multi-view validation to select based on CC and MLO predictions. The model, utilising data from various perspectives, enhances its comprehension of lesions and reduces the likelihood of misclassifying benign anomalies as malignant. This technique effectively identifies small or low-contrast defects that may be present in one image but absent in another during the final diagnosis. Ensemble learning combines predictions from multiple models to enhance performance in complex situations. Anomaly detection and BIRADS classification models are employed with both real and synthetic data to enhance forecasting accuracy. The ensemble model can improve predictions and reduce false positives by utilising classifier findings. The utilisation of multiple models and their varying strengths enhances ensemble classification and generalisation.

4.1 Limitations

The research, while promising, encountered several challenges, revealing issues within AI-powered breast cancer detection domains. The constraints render generative models in medical imaging both challenging and innovative.

Limits of Data Size and Variability:
The annotated samples within the bounded box were inadequate for optimal training, even with the presence of two large datasets, also the absence of segmentation masks hindered localization and recognition processes, because of the availability of only bounding boxes. The model exhibits difficulties in detecting objects based on the training data. The absence of diverse data hinders the model's ability to learn comprehensive patterns, leading to classification overfitting, particularly in the case of rare anomalies.

Generalization across clinical settings and vendors:
A further concern was the generalizability of the model's imaging source and clinical scenarios. Generative models trained on specific imaging equipment and clinical settings demonstrated inconsistency in results across various vendors and scenarios. Diversity is essential in imaging equipment, patient demographics, and clinical practice training datasets. Clinical imaging across various technologies and methodologies necessitates model generalizability.

Text prompt diffusion and inpainting present a trade-off in lesion generation:
Comparative analysis of text-diffusion inpainting reveals lesions. Inpainting of masks identified mammographic abnormalities, whereas text-prompted diffusion failed to do so. This contradiction suggests that diffusion models are inadequate for modeling complex text descriptions. Images generated from text prompts may fail to exhibit mammographic lesions. Diffusion models can generate realistic lesions from text descriptions in extensive and diverse mammography datasets.

4.2 Future Outlook

Enhanced text prompts for text-to-image diffusion may facilitate better control over lesion properties. Text-to-image diffusion models exhibit rapid improvement when provided with more precise and condition-specific textual prompts. Synthetic lesions may exhibit greater realism and diversity when accompanied by enhanced language descriptions. The model may replicate underrepresented diseases using tailored and contextually appropriate concepts for specific or unusual abnormalities, such as discrete calcifications or subtle tumor types.

Transformer-Based Advanced Task Models Transformer-based segmentation and other sophisticated models facilitate breast cancer diagnosis. Transformers effectively capture long-range data relationships, rendering them highly suitable for mammography and other complex regional patterns. Transformers demonstrate superior capability in detecting global pixel correlations compared to CNNs, particularly for subtle or diffuse breast tissue abnormalities. Transformer-based segmentation architectures enable models to identify and characterize moderate or irregular lesions. Improved segmentation may enhance health outcomes through earlier and more precise detection.

References

[Abhilasha, 2024] Abhilasha, Ashima Narang, P. V. (2024). A-review-of-ai-in-breast-cancer-detection. *International Journal of Innovative Research in Computer Science and Technology (IJIRCST)*, 12(2):126–129.

[Carriero et al., 2024] Carriero, A., Groenhoff, L., Vologina, E., Basile, P., and Albera, M. (2024). Deep learning in breast cancer imaging: State of the art and recent advancements in early 2024. *Diagnostics*, 14(8).

[Chambon et al., 2022] Chambon, P., Bluethgen, C., Delbrouck, J.-B., der Sluijs, R. V., Połacin, M., Chaves, J. M. Z., Abraham, T. M., Purohit, S., Langlotz, C. P., and Chaudhari, A. (2022). Roentgen: Vision-language foundation model for chest x-ray generation.

[Dembrower et al., 2020] Dembrower, K., Liu, Y., Azizpour, H., Eklund, M., Smith, K., Lindholm, P., and Strand, F. (2020). Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology*, 294(2):265–272.

[Gastounioti et al., 2022] Gastounioti, A., Desai, S., Ahluwalia, V. S., Conant, E. F., and Kontos, D. (2022). Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. *Breast Cancer Research*, 24(1):14.

[Ho et al., 2020a] Ho, J., Jain, A., and Abbeel, P. (2020a). Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

[Ho et al., 2020b] Ho, J., Jain, A., and Abbeel, P. (2020b). Denoising diffusion probabilistic models.

[Ho and Salimans, 2022] Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance.

[Hu et al., 2021] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.

[Khader et al., 2023] Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., Stegmaier, J., Kuhl, C., Nebelung, S., Kather, J. N., and Truhn, D. (2023). Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(1):7303.

[Montoya-del Angel et al., 2024] Montoya-del Angel, R., Sam-Millan, K., Vilanova, J. C., and Martí, R. (2024). Mam-e: Mammographic synthetic image generation with diffusion models. *Sensors*, 24(7).

[Nguyen et al., 2023] Nguyen, H. T., Nguyen, H. Q., Pham, H. H., Lam, K., Le, L. T., Dao, M., and Vu, V. (2023). Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography.

[Pinaya et al., 2022] Pinaya, W. H. L., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., and Cardoso, M. J. (2022). Brain imaging generation with latent diffusion models. In Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., and Yuan, Y., editors, *Deep Generative Models*, pages 117–126, Cham. Springer Nature Switzerland.

[Reis et al., 2024] Reis, D., Kupec, J., Hong, J., and Daoudi, A. (2024). Real-time flying object detection with yolov8.

[Rombach et al., 2022] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA. IEEE Computer Society.

[Ruiz et al., 2023] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.

[Sohl-Dickstein et al., 2015] Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics.

[Song and Ermon, 2020] Song, Y. and Ermon, S. (2020). Generative modeling by estimating gradients of the data distribution.

[Tan and Le, 2020] Tan, M. and Le, Q. V. (2020). Efficientnet: Rethinking model scaling for convolutional neural networks.

[Terven et al., 2023] Terven, J., Córdova-Esparza, D.-M., and Romero-González, J.-A. (2023). A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716.

[Yala et al., 2019] Yala, A., Lehman, C., Schuster, T., Portnoi, T., and Barzilay, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1):60–66.

[Yaseen, 2024] Yaseen, M. (2024). What is yolov8: An in-depth exploration of the internal features of the next-generation object detector.