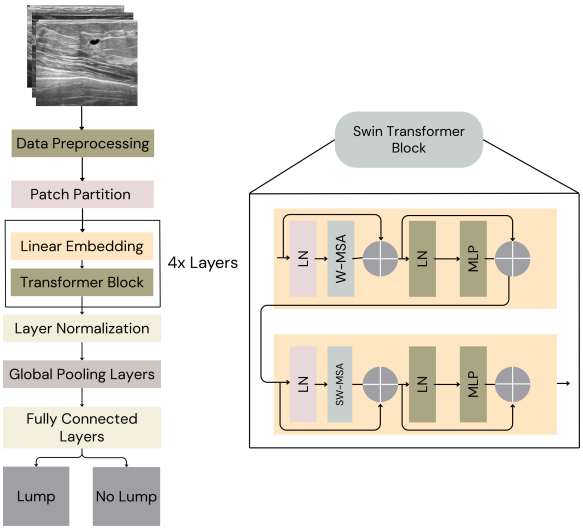
ROBT310 Final Project: Breast Cancer Detection

Madiyar Kairolla, Dosbolat Adekenov,
Aibar Akimbayev, Aldiyar Sagat

***Abstract*— Early and reliable detection of breast cancer in ultrasound (US) imaging remains challenging due to low contrast, speckle noise, and wide variability in tissue appearance. We study two complementary deep learning approaches for two-class US classification (lump, no lump): (i) a lightweight convolutional neural network (CNN) baseline and (ii) a Swin Transformer. The dataset (Hugging Face source, masks removed) was curated into class-balanced train/validation/test splits. Image processing techniques included resizing inputs to 224 by 224 pixels, Z-score normalization using ImageNet statistics, and PARULA pseudocoloring to enhance visual contrast. We also applied extensive on-the-fly augmentation, such as affine transformations, multiplicative noise, and Gaussian blur. The CNN (three convolutional blocks + dropout) yields test f1-score in the range between 0.7 and 0.8, with most errors concentrated in the *no lump* class, reflecting class imbalance and the subtlety of no lump parenchyma on US. The Swin Transformer, trained with mixed precision and tuned learning-rate and weight-decay, demonstrated consistently lower validation loss and improved precision and recall relative to the CNN under comparable data preprocessing, though compute constraints limited exhaustive ablations. We report implementation details, training dynamics, and error analyses, and we discuss practical remedies (loss re-weighting, sampling, calibration) that reduce false negatives without sacrificing specificity. Additionally, we explore the potential of pretrained medical Vision-Language Models (VLMs) through preliminary zero-shot testing on select samples. Our results highlight the value of hybrid or Transformer-based pipelines for capturing both local textural cues and global context in breast US, while emphasizing dataset design and evaluation choices that materially affect clinical utility.**



***Index Terms*— Breast ultrasound, medical image classification, Swin Transformer, convolutional neural networks, transfer learning, data augmentation, class imbalance, calibration.**

I. Introduction

Breast cancer is one of the most common and dangerous diseases that affect women around the world. Detecting it early is very important because it increases the chance of successful treatment. Doctors usually use medical images such as mammograms or microscopic images to look for signs of cancer. However, checking these images by hand can take a long time and sometimes leads to mistakes such as false positives (detecting cancer when it is not there) or false negatives (missing cancer that is actually there) [10]. In recent years, deep learning has become a popular and powerful tool for analyzing medical images. One of the most common architectures used is the Convolutional Neural Network(CNN). CNNs can automatically learn useful features from medical images and perform binary classification identifying between cancerous and non-cancerous cases. However, CNNs mainly focus on small, local parts of an image and sometimes miss the larger picture, such as overall tissue structure or texture [1]. To solve this problem, researchers developed Vision Transformers (ViTs). These architectures can learn both local details and global patterns in an image by using a mechanism called self-attention, which allows the model to look at different areas of the image at once [2]. One of the most advanced versions is the Swin Transformer. It divides an image into small windows and shifts them slightly at each layer. This helps the model learn both fine and broad image features more efficiently, without using too much computer power [10]. According to Tanimola et al. (2024), the Swin Transformer showed very high accuracy - around 99.9% - in the process of performing binary classification on mammographic breast cancer images,

Submitted December 2025 and revised month year.
Corresponding authors: Madiyar Kairolla, Dosbolat Adekenov, Aibar Akimbayev, Aldiyar Sagat.
Madiyar Kairolla is with Nazarbayev University, Nur-Sultan, Kazakhstan (e-mail: madiyar.kairolla@nu.edu.kz).
Dosbolat Adekenov is with Nazarbayev University, Astana, Kazakhstan (e-mail: dosbolat.adekenov@nu.edu.kz).
Aibar Akimbayev is with Nazarbayev University, Astana, Kazakhstan (e-mail: aibar.akimbayev@nu.edu.kz).
Aldiyar Sagat is with Nazarbayev University, Astana, Kazakhstan (e-mail: aldiyar.sagat@nu.edu.kz).

separating cancerous from non-cancerous cases. It performed better than well-known CNN architecture such as ResNet50 and VGG16 [10]. Similarly, Chen and Martel (2025) combined the Swin Transformer and CNN in a hybrid model for mammogram analysis. Their system achieved an AUC of 0.889 and showed that using both architecture together helps capture both global and detailed features more effectively [1]. In this project, we compare a Regular CNN and a Swin Transformer for breast cancer detection. We also use tuning (adjusting model settings for better results) and pruning (removing less important parts to make the model smaller and faster). We test both architectures using evaluation metrics such as accuracy, precision, recall, F1-score, and a confusion matrix to measure how well they classify breast cancer images. Our goal is to build an accurate and efficient deep learning model for breast cancer detection that combines the strong local feature learning of CNNs and the global image understanding of Transformers, following recent successful research in this area [1], [2], [10].

II. Model Description

A. Convolution Neural Network

Convolutional Neural Networks are deep learning models made of several stages that process data step by step. Input and output stages takes a set of arrays called feature maps as input and produces another set of feature maps as output. For example, a color image has three 2D feature maps - one for each color channel. Audio signals use 1D feature maps, while video uses 3D feature maps. After passing through a stage, each map represents some learned visual pattern.

A typical stages includes three main layers: a filter bank layer, a non-linearity layer, and a pooling layer [3]. Most of the stages contain one to three of these stages, followed by a classification layer.

Filter Bank Layer. This layer takes several input feature maps and produces several output feature maps. Each output map is created by convolving a trainable filter with one or more input maps, adding a bias term. This operation allows the filter to detect a specific pattern at every location in the image. If the input shifts slightly, the output also shifts, but the detected pattern remains consistent.

Non-Linearity Layer. After convolution, a non-linear function is applied to each value in the feature maps. Early models used the tanh activation, but more recent models use stronger nonlinearities such as rectified functions. Sometimes, this is followed by local contrast normalization, which encourages nearby features to compete with each other.

Feature Pooling Layer. Pooling reduces the resolution of feature maps by combining values in small neighborhoods. The simplest method is to compute the average over a neighborhood in each map. This operation while sometimes replace by max pooling which is used in modern networks. Pooling makes the model more robust to small shifts in the input. Some newer models remove the pooling layer and use convolutions with larger strides to reduce resolution.

B. Swin Transformer

Computer vision tasks most of the time was dominated by convolutional neural networks (CNNs), especially during the rise of different CNN baseline models, such as Visual Geometry Group (VGG) [7] as being very deep CNN baseline, GoogLeNet [8], which increased the number of convotions on a single layer, YOLO [6], which is used for object detection. Since the start of convolutional neural network in 1989 with the paper of LeCun et al. (1989) [3], increased the efficiency of the CNN baseline, which is currently is going to its saturation point [2], which led researcher over the world to try new architectures for computer vision tasks. This was one of the main reasons to redirect the Transformer architecture, which was the prevalent architecture of the natural language processing task, to the computer vision task. This led to the creation of Swin Transformer with Transformer architecture backbone, which used shift windows and self-attention mechanisms to properly identify not only the necessary batch of pixels, but also see the whole picture of the image.

Swin Transformer model has different structure, which can be seen in Fig 1. It has so called "Swin Transformer Block", which used not the standard multi-head self-attention (MSA) module, but a shifted window self-attention module [4]. Shifted window works as dividing the whole image into different windows (blocks) and apply self-attention mechanism on them, which can be seen in Fig 2 [9], while the basic Vision Transformer is dividing the image into the same windows, where the necessary information could be lost, especially in the edges of the windows.

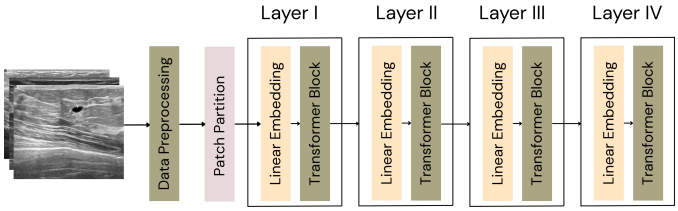


Fig. 1: Swin Transformer architecture.

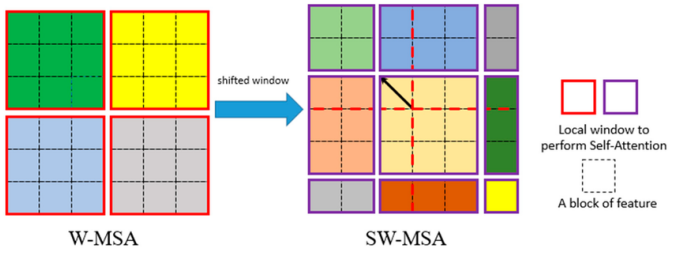


Fig. 2: Shifted window approach.

Mathematical Representation of the Input. Let $I \in \mathbb{R}^{H \times W \times C}$, where I is an input, H and W are height and width, and C is a channel dimension ($C = 3$ for RGB, $C = 1$ for grayscale). In real-world we do not operate on the continuous manifold of the image, because it will dramatically increase time and space complexity of the model, because of this, we will apply discretization on the image by using **Patch Partition** P, usually it has a size 4×4. Then our image becomes sequence of vectors, or mathematically,

$$I \to X \in \mathbb{R}^{N \times (P^2 \cdot C)} \tag{1}$$

Where $N = \frac{H}{P} \times \frac{W}{P}$ is the total number of patches (tokens in transformer architecture language). This transformation can be seen as a projection from spatial domain to the sequence domain. We then apply a linear embedding E to project vectors into an arbitrary latent dimension d:

$$Z^{(0)} = XE + E_{pos} \tag{2}$$

where $Z^{(0)} \in \mathbb{R}^{N \times d}$ is the input to the first transformer layer, and E_{pos} is a positional embedding, which we will define later.

Self-Attention Mechanism. Mathematically self-attention mechanism could be understand as **Scaled Dot-Product Attention**. Let Z^{l-1} be the input feature to layer l. We define three linear transformations (learnable weight matrices) $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$.
We compute the Query (Q), Key (K), Value (V) matrices:

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V \tag{3}$$

Then there a mapping Att : $\mathbb{R}^{N \times d} \times \mathbb{R}^{N \times d} \times \mathbb{R}^{N \times d} \to \mathbb{R}^{N \times d}$ defined as:

$$\text{Att}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

Where matrix QK^T is the inner product between every pair of patches, and since matrix Q and matrix K^T are normalized, our product matrix will be proportional to the cosine similarity between two attributes.

Scaling $\frac{1}{\sqrt{d}}$ is natural normalization, because our matrices has mean 0 and standard deviation 1, but their inner sum will have mean 0, but standard deviation to be d, and because of this reasons by dividing $\frac{1}{\sqrt{d}}$ will normalize the resulting matrix.

Softmax is a function which defined as:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^j} \tag{5}$$

Which is in fact a probability distribution, and it acts as a convex combination of value vectors, because of the exponential function.

Window-Based Attention. For this one, Swin transformer partitions the feature map Z into non-overlapping window of size $M \times M$, then the total number of windows defines the same as previous, $\Omega = \frac{H}{M} \times \frac{W}{M}$. Because of this, the self-attention is computed locally for each window, which is also resulted in less time complexity up to being linear relevant to the input image size ($O((HW)^2) \to O((HW) \cdot M^2)$). Additionally, the Swin transformer architecture is using the **Relative Position Bias**, which is necessary to the model properly follow the calculation of the positionsm, and then the self-attention formulas is modified:

$$\text{Att}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}} + B)V \tag{6}$$

Where $B \in \mathbb{R}^{M^2 \times M^2}$ is learnable bias matrix, which calculated as a difference between two patches.

Shifted Window Partition. By using only window based attention, it will create an isolated windows with no communication between other windows. To solve this we define "Shift" operation.

Let l to be a regular layer. For $l+1$, the partitioning grid is shifted by a displacement vector:

$$v_{shift} = (\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor) \tag{7}$$

This shift will create new windows, which will now have different sizes. And now to connect these disconnected windows we perform shifting; however, if we perform naive shift it will require us to pad the whole image, which is computationally inefficient, hence, we introduce the **cyclical** shifting, which will shift top-left window to the bottom-right. Mathematically,

$$Z_{shifted} = roll(Z, \text{shifts} = (-\frac{M}{2}, -\frac{M}{2}) \tag{8}$$

However, it will result on adjacent patches, which originally was disconnected from each other, and in order to prevent attention to each other, we introduce **Mask Matrix** M.

$$\text{MaskedAtt}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}} + B + M)V \tag{9}$$

where $M_{ij} = 0$ if token i and token j are from the same original region, and $M_{ij} = -\infty$ if they are different, which will force softmax to be 0, shown in Fig 3 [4].

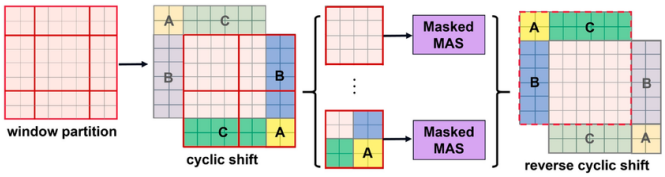
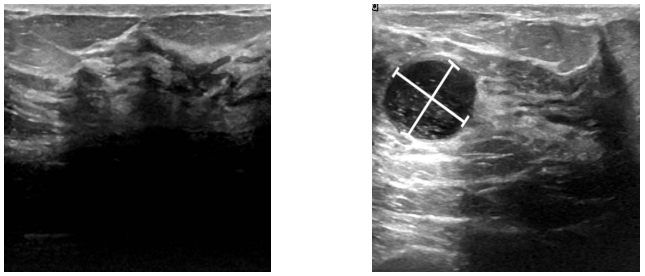


Fig. 3: Efficient batch computation approach.

III. Methodology

A. Dataset Description

The dataset used in this study consists of almost 850 breast ultrasound images of size 500×500, example of which shown in Fig 4, categorized into two classes: lump and no lump. The dataset was originally obtained from Hugging Face [5], but the images of the segmentation masks were removed to avoid noise and ensure cleaner training data. Furthermore, it had class imbalance, which caused a lot of troubles during training, so the dataset was shrunk up to 460 ultrasound images with 30% of no lump images. The cleaned data was manually organized into subfolders for each class, each containing training, validation, and test images (60% train, 20% val, 20% test), shown in Table I.



(a) Example with no lump (b) Example with lump

Fig. 4: Two example images from the Vanilla Dataset.

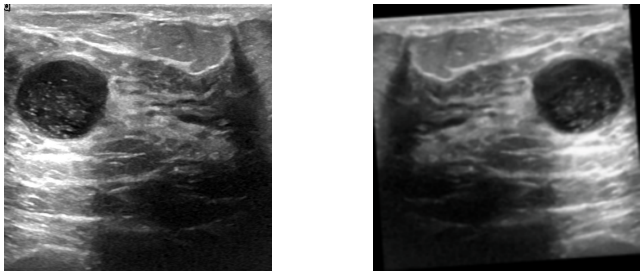
TABLE I: Dataset split by class

Class	Train	Validation	Test
Lump	194	66	66
No lump	80	27	26

B. Data Preprocessing and Augmentation

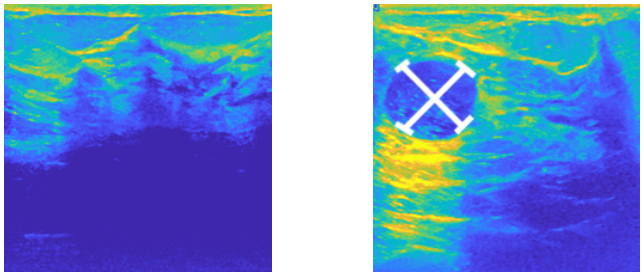
Dataset were resized to 224×224 to efficiently use the CNN based pipeline and Swin Transformer [4] model. Intensity values were normalized (z-scored) using ImageNet mean and standard deviation, which is necessary to efficiently use the pre-trained weights. To create a more robust model that generalizes well to unseen data, several stochastic techniques were applied, such as horizontal flip, affine transformation with scaling, translation, and rotation, multiplicative noise, and Gaussian blur shown in 5. Finally, the preprocessed images were converted to PyTorch tensors to use the dataset for training the models.

Moreover, there was an idea to use pseudocolor image processing to preprocess the dataset to improve the results of the model, so, by using MATLAB software, dataset was preprocessed with PARULA pseudocoloring, shown in Fig 6.



(a) Before Preprocessing. (b) After Preprocessing

Fig. 5: US image before and after Data Preprocessing .



(a) Example with no lump (b) Example with lump

Fig. 6: Two example images from the PARULA Dataset.

C. Model Configuration and Training Setup

CNN baseline architecture is based on Keras module, and it consists three convolutional blocks with 32, 64, and 128 filters with 3×3 kernel and ReLu activation function, each with 2×2 max-pooling layer to increase the efficiency of the baseline. As a classifier head, it contains flatten, dropout with 0.2, and the output layer with 3 neurons and softmax activation function. Training setup for the CNN baseline is shown in Table II

TABLE II: Training Setup CNN

Parameter	**Value**
Loss Function	Cross-Entropy
Optimizer	AdamW
Learning Rate	1×10^{-3}
Batch Size	32
Epochs	20
Hardware	RTX 2060
Checkpoint Strategy	Save best on min. validation loss

Training setup for the Swin Transformer model is shown in Table III

TABLE III: Training Setup Swin Transformer

Parameter	**Value**
Loss Function	Cross-Entropy Loss
Optimizer	AdamW
Learning Rate	3×10^{-5}
Weight Decay	0.2
Batch Size	16
Epochs	15
Hardware	NVIDIA T4 (Google Colab)
Mixed Precision	Enabled (AMP)
Checkpoint Strategy	Save best on min. validation loss

IV. Experiments and Results

A. CNN Baseline

In the beginning, we tried to use the algorithm optimizers Adam, AdamW, and SGD, but Adam had worse results than AdamW, and SGD was too slow for repeated experiments, although it was more accurate than AdamW. Then we rescaled the images of the dataset to 224×224 for the convenience of the model learning. Also, we initially used two convolution layers of kernels (filters) for processing the dataset, but we changed it to three layers, so model will show better results and also still avoid the overfitting problem. In addition, we stopped at 20 epochs for the learning.

The model produces the test accuracy and f1-score in the range between 0.7 and 0.8. As we can see in the confusion matrix (Table IV) for the CNN based model, it identifies each case in a decent way, so it is considered as a great results.

In the process of experiments, we decided to pass the dataset through PARULA to make it more visible, but after training the CNN model, the results became worse, which indicates that PARULA pseudo-coloring does not help to identify the lump. The validation accuracy fell by $10\% - 15\%$ with the new dataset, especially the results with malignant and normal data showed wrong determination in about 70% cases, so the attempt failed. Shown in the confusion matrix, Table V

Actual / Predicted	**Clear**	**Lump**
Clear	21	5
Lump	14	52

TABLE IV: Confusion Matrix for CNN baseline with Vanilla Dataset

Actual / Predicted	Clear	Lump
Clear	13	13
Lump	6	60

TABLE V: Confusion Matrix for CNN baseline with PARULA Dataset

B. Swin Transformer

The Transformer is one of the most convenient and straightforward model architectures to use, due to this, the experiment started with 384×384 resolution; however, because of lack of computational resources, it was more accurate to use 224×224 resolution. Then, after unsuccessful trials with different optimizers, it was obvious that AdamW optimizer was the best. After that, we tried different learning rates and weight decays, which were ideally founded after more than 250 model trainings, which started with 0.4 f1-score and 65% precision, and after preprocessing was improved up to 82% precision and 0.72 f1-score (Table VI), which distinguish the ability of the model to optimize and get the higher score in metrics, and it still worse than CNN baseline. Next, we tried to use the second dataset with PARULA pseudocoloring, which in theory should give better results, since transformer based model try to look the entire picture, and after several training the model start to differentiate the no lump and lump cases better (Table VII). Finally, unfortunately for us, the transformer based model is not good to work on such small datasets, where we cannot train the model for several hours, and because of that it had a little change in its class weights, which indicates that it was not a good solution for such problems.

Actual / Predicted	Clear	Lump
Clear	5	21
Lump	10	56

TABLE VI: Confusion matrix for Swin Transformer with Vanilla Dataset.

Actual / Predicted	Clear	Lump
Clear	10	16
Lump	13	53

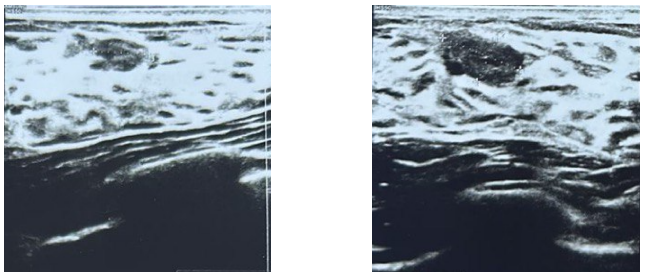
TABLE VII: Confusion matrix for Swin Transformer with PARULA Dataset.

C. Medical images from Kazakhstan

Lastly, but most importantly, we need to discuss the possibility of trained models to properly classify the ultrasound image from Kazakhstan. Because of this, we obtained two ultrasound images of left and right breasts with small lumps. Shown in Fig 7, and the following description in Fig 8.

After more than 300 times of trainings of different models, we decided to pick the CNN baseline, since Swin Transformer model was not good enough to differentiate between lump and no lump ultrasound images. So, the only the model which we can trust is CNN model.

Without applying any transformations to the images we obtain the matching results with the given ultrasound images; both of the images were classified as having lumps on it, which is decent results for the CNN model.



(a) US of left breast (b) US of right breast

Fig. 7: Two Ultrasound images from Kazakhstan.

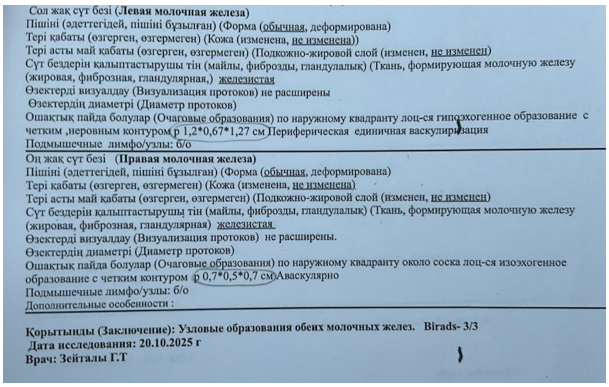

Сол жақ сүт безі (**Левая молочная железа**)
Пішіні (әдеттегідей, пішіні бұзылған) (Форма (обычная, деформирована)
Тері қабаты (өзгерген, өзгермеген) (Кожа (изменена, не изменена))
Тері асты май қабаты (өзгерген, өзгермеген) (Подкожно-жировой слой (изменен, не изменен)
Сүт бездерін қалыптастырушы тін (майлы, фиброзды, гландулалық) (Ткань, формирующая молочную железу (жировая, фиброзная, гландулярная,) железистая
Өзектерді визуалдау (Визуализация протоков) не расширены
Өзектердің диаметрі (Диаметр протоков)
Ошақтық пайда болулар (Очаговые образования) по наружному квадранту лоц-ся гипоэхогенное образование с четким ,неровным контуром р 1,2*0,67*1,27 см Периферическая единичная васкуляризация
Подмышечные лимфо/узлы: б/о
Оң жақ сүт безі (**Правая молочная железа**)
Пішіні (әдеттегідей, пішіні бұзылған) (Форма (обычная, деформирована)
Тері қабаты (өзгерген, өзгермеген) (Кожа (изменена, не изменена)
Тері асты май қабаты (өзгерген, өзгермеген) (Подкожно-жировой слой (изменен, не изменен)
Сүт бездерін қалыптастырушы тін (майлы, фиброзды, гландулалық) (Ткань, формирующая молочную железу (жировая, фиброзная, гландулярная) железистая
Өзектерді визуалдау (Визуализация протоков) не расширены.
Өзектердің диаметрі (Диаметр протоков)
Ошақтық пайда болулар (Очаговые образования) по наружному квадранту около соска лоц-ся изоэхогенное образование с четким контуром р 0,7*0,5*0,7 см Аваскулярно
Подмышечные лимфо/узлы: б/о
Дополнительные особенности :
Қорытынды (Заключение): Узловые образования обеих молочных желез. Birads- 3/3
Дата исследования: 20.10.2025 г
Врач: Зейталы Г.Т

Fig. 8: Description of the US images

V. Discussion and Conclusion

Under the same preprocessing, the CNN baseline reached 0.70–0.80 test accuracy and f1-score, even with class imbalance (no lump $\approx 30\%$), which was a great results for such a simple deep learning architecture in current realities. The Swin Transformer, with windowed self-attention and ImageNet pretraining, showed consistently higher validation loss and stronger precision trends at 224×224, indicating higher rate of overfitting and inability to train on a small scaled datasets. Brief error analysis suggests threshold tuning and post-hoc calibration are as impactful as backbone choice. Practically, we recommend lightweight imbalance remedies (class-weighted or focal loss, batch-aware sampling, modest over-sampling of *no lump*), metrics beyond accuracy (ROC–AUC, macro PR–AUC, calibration error), and modest scaling to higher resolution with test-time augmentation. In summary, CNN provide a clearer path to improved robustness on breast US; with targeted imbalance handling and calibration, further reductions in false positives are attainable without prohibitive compute. Looking ahead, we plan to extend the pipeline with multimodal vision–language models that fuse images with clinical text (e.g., BI-RADS, notes, metadata) to enable retrieval-augmented triage, explanation-grounded reports, and interactive QA for clinician-facing decision support.

Vision–Language Models as an Additional Modality. The main problem with our dataset is that we don't have any medical information about the patients, just the images. Because of this, we decided to look at a different approach. To address our limited dataset size, we experimented with a pretrained medical vision–language model, **LLaVA-Med-v1.5-Mistral-7B**. Unlike CNN or Swin, which require supervised training, this model operates fully *zero-shot*:

given an ultrasound image (and optional clinical history), it produces a diagnostic label together with a short explanation and recommendation. Despite never being trained on our dataset, LLaVA-Med consistently distinguished *clear* vs. *lump* cases and generated coherent, radiologist-style reasoning. This confirms that medical VLMs can inject strong prior knowledge and interpretability even when only small image datasets are available.

REFERENCES

[1] H. Chen and A. L. Martel, "Enhancing breast cancer detection on screening mammogram using self-supervised learning and a hybrid deep model of Swin Transformer and CNN," *arXiv preprint arXiv:2504.19888*, 2025. doi: https://arxiv.org/abs/2504.19888

[2] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021. doi: https://openreview.net/forum?id=YicbFdNTTy

[3] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989. doi: https://doi.org/10.1162/neco.1989.1.4.541

[4] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *arXiv preprint arXiv:2103.14030*, 2021. doi: https://doi.org/10.48550/arxiv.2103.14030

[5] S. Raisharma, "Breast Ultrasound Images Dataset," *Hugging Face*. Accessed: Nov. 9, 2025. [Online]. Available: https://huggingface.co/datasets/ShivamRaisharma/breastcancer

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788. doi: https://doi.org/10.1109/CVPR.2016.91

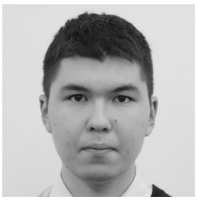
[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015. doi: https://arxiv.org/abs/1409.1556

[8] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9. doi: https://doi.org/10.1109/CVPR.2015.7298594

[9] Z. Tang, B. Wu, W. Wu, and D. Ma, "Fault detection via 2.5D transformer U-Net with seismic data pre-processing," *Remote Sens.*, vol. 15, no. 4, Art. no. 1039, Feb. 2023. doi: https://doi.org/10.3390/rs15041039

[10] O. Tanimola, O. Shobayo, O. Popoola, and O. Okoyeigbo, "Breast cancer classification using fine-tuned Swin Transformer model on mammographic images," *Analytics*, vol. 3, no. 4, pp. 461–475, 2024. doi: https://doi.org/10.3390/analytics3040026

Madiyar Kairolla Contribution to project: Constructed the pipeline with Swin Transformer, and trained the model over 250 times. Moreover, plotted confusion matrix for the Swin Transformer model. In addition, explained mathematical details of the Swin transformer pipeline.



Dosbolat Adekenov Contribution to project: Contribution to project: constructed and trained the CNN model. In addition, plotted confusion matrix for CNN model.

Aibar Akimbayev Contribution to project: studied background information and preprocessed the PARULA dataset.

Aldiyar Sagat Contribution to project: Explored pretrained medical VLMs (Florence-2, BioViL, LLaVA-Med) for zero-shot ultrasound interpretation. Developed preprocessing and augmentation pipelines and carried out VLM inference experiments. Wrote the Abstract, Discussion, and Conclusion, and formatted all References in IEEE style.