STATS 403 Final Project Report: RSNA Screening Mammography Breast Cancer Detection

Chenglin Zhang cz155@duke.edu

Yijia Xue yx179@duke.edu

Lihui Chen lc349@duke.edu

Contents

1 Problem Introduction

1.1 Background Information, Importance & Motivation

Breast cancer is the most frequently diagnosed cancer in women worldwide with 2.26 million [95% UI, 2.24–2.79 million] new cases in 2020. And in the U.S., breast cancer alone is expected to account for 29% of all new cancers in women [1]. It has the highest incidence rate among cancers in female population according to the statistics from Centers for Disease Control and Prevention (Figure 1). The severity of breast cancer is also warrant mentioning. The

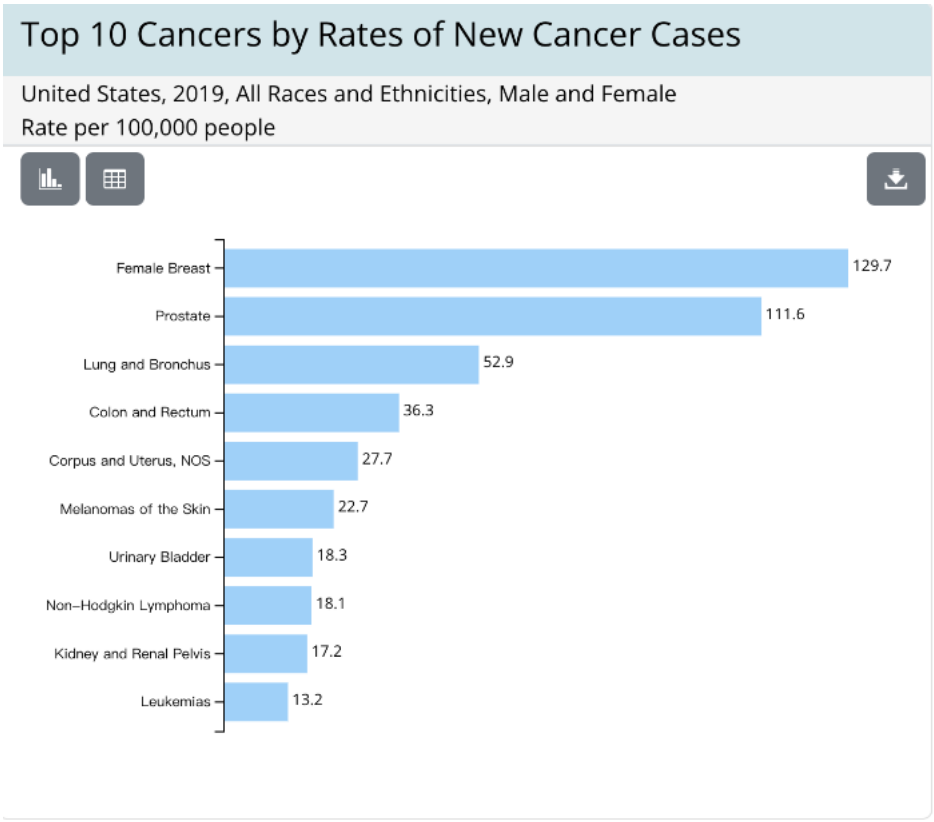

Figure 1: Top 10 Cancers by Rates of New Cancer Cases [2]

death rate of breast cancer is the second highest among all cancers, just behind the Lung and Bronchus cancer according to the Centers fro Disease Control and Prevention [2]. In addition to the high incidence and mortality rate, breast cancer imposes huge burden on patients and their family. According to a survey from Breastcancer.org, a non-profit organization dedicated to promote the knowledge about breast cancer, reveals that, among 1,437 U.S. people who were diagnosed with breast cancer in the past 10 years, 47 % their breast cancer-related out-of-pocket costs were a significant or catastrophic burden, and 37% reduced spending on basic necessities to cover treatment costs [3].

The severity and burden of breast cancer can also be attributed to the late-stage diagnosis because of a lack of early symptom of this disease. Detection at advanced stages of the decease implies the treatment is more difficult and uncertain [4]. Therefore, an early stage detection is proposed in order to prevent the disease from being diagnosed too late. Mammographic screening is a good way of early detection when resource permitted [5]. With the advancement of machine learning/deep learning algorithms, and the popularity of mammography in many countries, we aim to incorporate deep learning techniques to “read” the mammography and returns the diagnosis of the given mammography.

2 Literature Review

2.1 Mammography Breast Cancer Detection

Chaurasia and Culurciello [6] have proposed "Linknet" to exploit the use of encoder in semantic segmentation. Their ideas of bypassing the output of the encoder layer to the decoder layer is similar to the Unet architecture that preserves

the low-dimensional information when using the convolution. The simplicity of the model regarding parameters and GFLOPs compared to other deep networks also enabled them to achieve the state-of-the-art results in CamVid and Cityscapes. Figure 2 exhibits the architecture of their model.

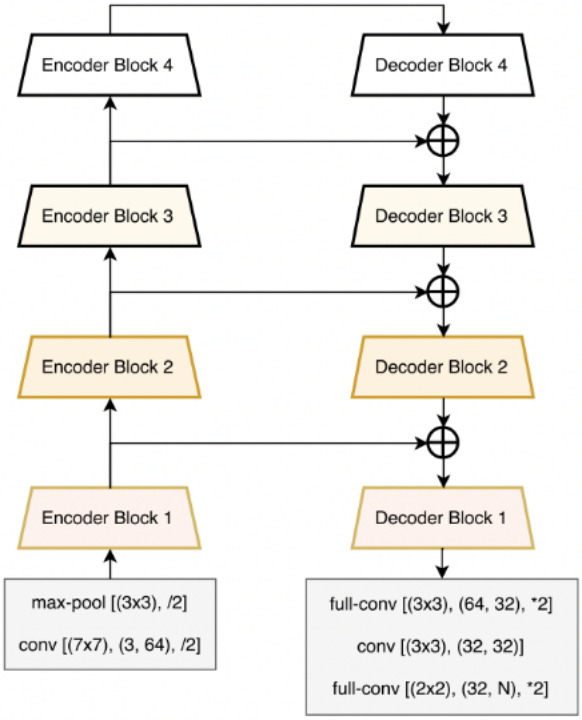

Figure 2: Linknet by Chaurasia and Culurciello

The previous work by Chaurasia and Culurciello has shown the efficacy of deep learning in the tasks such as image segmentation. Moreover, Abunasser et al. [7] focused the task on medical image classification task which is on breast cancer detection, a topic that our study aims to focus, in their study. They utilized convolution-based networks, Xception, to classify a total of eight breast cancer sub-classes. They used techniques such as skip connection and filter concat to fine-tune the network. Their result on breast cancer classification has achieved a promising score of 97.60% Precision, 97.60% Recall and 97.58% F1-Score.

2.2 Dataset and Class Imbalance

The dataset is from [8] and it composes of two parts. The first part is the training mammography images, which contain breast images from over 10,000 patients about their CC and MLO on the left and right side of the breast. The other part of the data contains a descriptive table in which there is information about the patient id, whether the particular breast image has cancer, etc.

Our exploratory analysis of the dataset also reveals a huge imbalance problem in the dataset, as the incidence rate of the breast cancer in the image only take up approximately 1%. Therefore, the imbalanced dataset imposes difficulty in training the deep learning models. We would stress this problem in the following data preprocess section.

2.3 Implemented Model and Methodology

In this study, we acquired and preliminarily analyzed the dataset from a Kaggle competition [8], preprocessed the data more a simpler and more accurate training process, established UNet, Vision Transformer, and Efficient Net to classify the mammography, and evaluated our established models using probabilistic F1 score. Our implementation is described in the following chart (Figure 4).

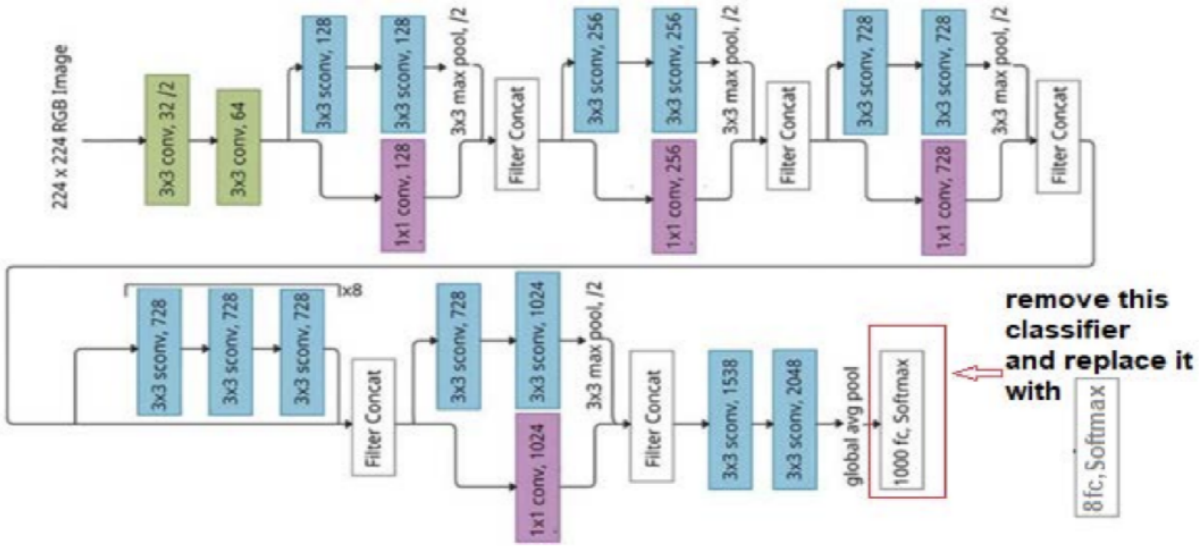

Figure 3: Xception by Abunasser et al.

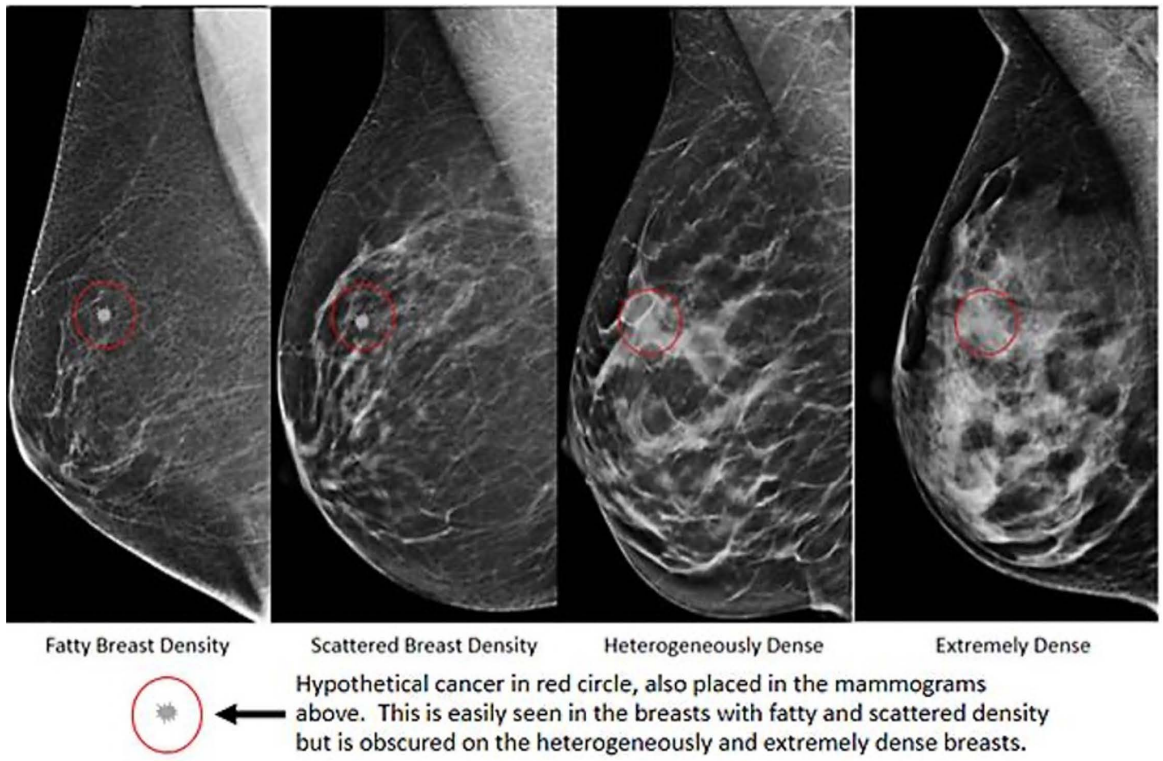

Figure 4: Illustration of how a cancer may be difficult to identify on a mammogram for women with dense breast tissue.[9]

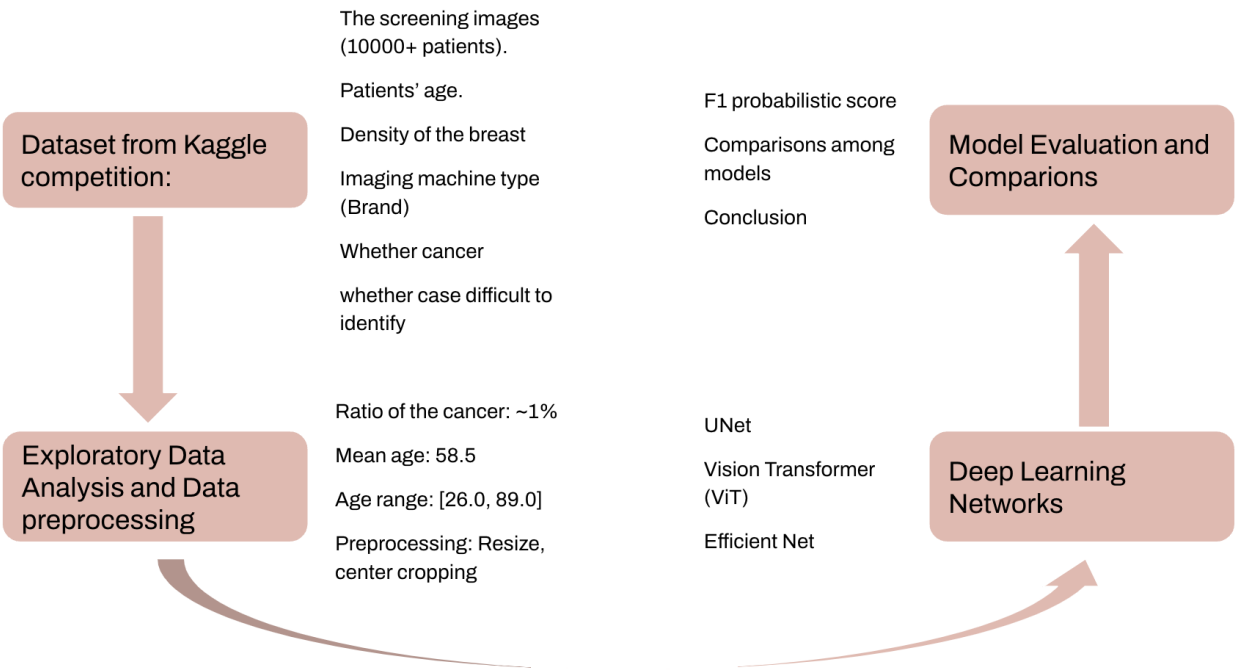

Figure 5: Flowchart of our Proposed Study

2.3.1 U-Net and Data Preprocessing

The U-Net [10] is widely known for medical image segmentation as shown in Figure 2.3.1. We will first talk about the U-Net to go through the many important techniques used in the model such as the overlap-tile strategy, image interpolation (in U-Net, bilinear interpolation is used), skip-connection, etc. Then, we will evaluate the usability of the U-Net model in our project, as well as the potential implications we get in the data (image) preprocessing phase.

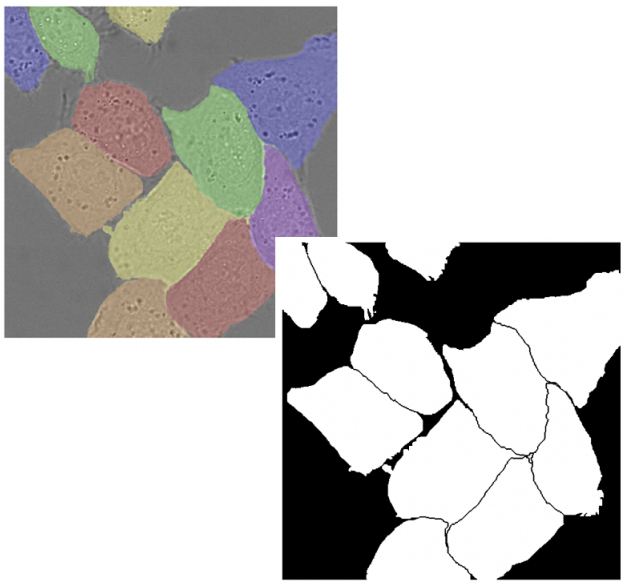

Figure 6: Segmentation [11]

In Figure 7, each blue box represents a multichannel feature map. White boxes represent copied feature maps (skip-connected from the downsampling phase). The arrows denote the different operations.

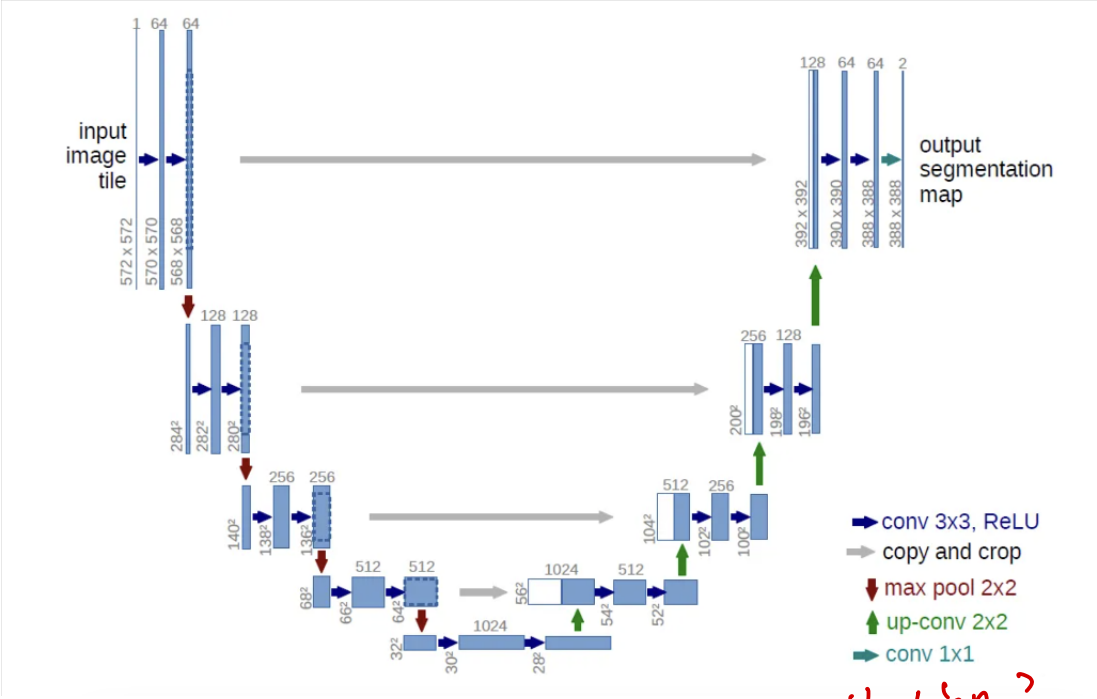

Figure 7: Basic Structure of the U-Net

citation ?

The U-Net mainly contains three important operations, which are

- Downsampling (Encoding)
- Upsampling (Decoding)
- Skip-Connection

In the discussion below, we will mainly focus on upsampling (image interpolation) and skip-connection, as well as other notable techniques used in the architecture.

Overlap-tile Strategy The overlap-tile strategy is used for the seamless segmentation of arbitrarily large images. Because medical images are of very high resolutions, feeding the entire image into the model would consume a lot of GPU memory. By using the overlap-tile strategy, one image input is divided into several tiles (in Figure 8, one tile is the yellow box), and we want to predict its segmentation. Missing input data (padding) is extrapolated by mirroring to make the division boundary continuous in the segmentation result. For example, in Figure 2.3.1, if the cutting line goes through one of the cells, then the final concatenation of the segmentation mask will become discrete.

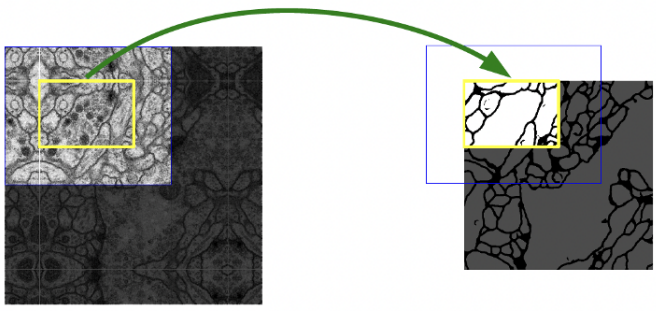

Figure 8: Overlap-tile Strategy

citation ?

Skip-connection The skip-connection is a very important technique used in the U-Net. As we can see in Figure 9, in the bilinear interpolation, the boundary of the object will get blurred as we do upsample, which will influence the

performance of the generated masks. In order to make up for the information lost in the downsampling during the encoding stage, between the encoder and the decoder of the network, the U-Net uses the Concat layer to fuse the feature maps. Thus, more high-resolution information from the downsampling phase can be used during upsampling. So, the detailed information (semantic boundaries) in the original image can be recovered better. Thus, the segmentation performance (accuracy) will be improved.

Image Interpolation The image interpolation technique is used in the deconvolution process of the U-Net. In the original architecture, bilinear interpolation is used to do upsampling of the encoded features. Some image interpolation methods are shown in Figure 9.

Figure 9: Different image interpolation examples

citation ?

For example, in Figure 10, we want to do upsampling from a 2×2 image to a 4×4 image. We first need to identify the source pixel of each pixel in the upsampled image in the original image. The equation is given by equation 1.

$$X_{src} = (X_{dst} + 0.5) \times \frac{Width_{src}}{Width_{dst}} - 0.5 Y_{src} = (Y_{dst} + 0.5) \times \frac{Height_{src}}{Height_{dst}} - 0.5 \quad (1)$$

treat equations as part of paragraphs

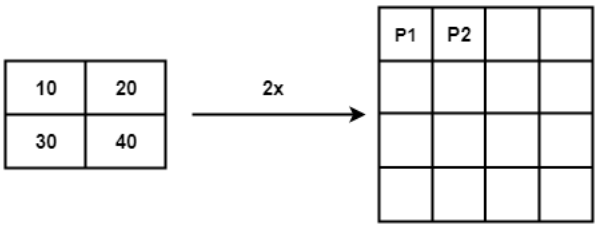

Figure 10: Example of upsampling

The computed coordinate of the source pixels may not necessarily be integers. It may fall among several pixels. The bilinear Interpolation uses the four closest pixels around the computed source pixel to derive the actual value of the destination pixel. The actual formula is given by the equation below (according to the parameters shown in Figure 11).

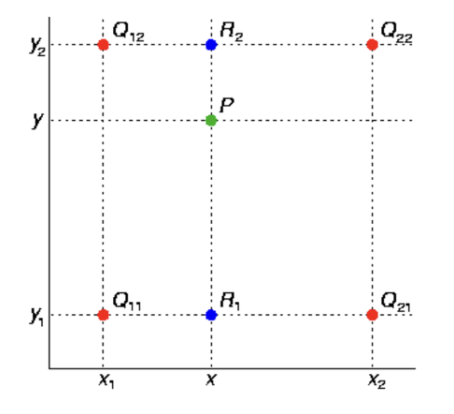

Figure 11: The four nearest pixels in the source image

$$
\begin{aligned}
f(R_1) &\approx \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21}) \\
f(R_2) &\approx \frac{x_2 - x}{x_2 - x_1} f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} f(Q_{22}) \\
f(P) &\approx \frac{y_2 - y}{y_2 - y_1} f(R_1) + \frac{y - y_1}{y_2 - y_1} f(R_2) \\
f(x, y) &\approx \frac{f(Q_{11})}{(x_2 - x_1)(y_2 - y_1)} (x_2 - x)(y_2 - y) + \\
&\quad \frac{f(Q_{21})}{(x_2 - x_1)(y_2 - y_1)} (x - x_1)(y_2 - y) + \\
&\quad \frac{f(Q_{12})}{(x_2 - x_1)(y_2 - y_1)} (x_2 - x)(y - y_1) + \\
&\quad \frac{f(Q_{22})}{(x_2 - x_1)(y_2 - y_1)} (x - x_1)(y - y_1) \\
&= \frac{1}{(x_2 - x_1)(y_2 - y_1)} (f(Q_{11}(x_2 - x)(y_2 - y) + f(Q_{21}(x - x_1)(y_2 - y) + \\
&\qquad f(Q_{12}(x_2 - x)(y - y_1) + f(Q_{22}(x - x_1)(y - y_1))
\end{aligned}
$$

Loss Function Firstly, in the U-Net model of the original paper [10], the Pixel-wise Softmax Function is defined by

$$p_k(x) = \exp a_k(x) / \left(\sum_{K} k' = 1 \exp(a_{k'}(x)) \right) \tag{2}$$

where $a_k(x)$ represents the activation value of pixel x in the k-th channel of the feature map. K represents the total number of classes. The cross-entropy then penalizes at each position the deviation of $p_{\ell(x)}(x)$ from 1 using

$$E = \sum_{x \in \Omega} w(x) \log(p_{\ell(x)}(x)) \tag{3}$$

where $\ell : \Omega \leftarrow \{1, \cdots, K\}$ is the true label of each pixel and $w : \Omega \leftarrow \mathbb{R}$ is a weight map that we introduced to give some pixels more importance in training. For example, in Figure , we want the boundary among the cells to be accurately predicted as the background instead of the cells. However, it is often the case that the predicted mask will merge the cells and thus mistakenly classify the pixels among the cells. Thus, adding a weight matrix and giving more penalties for the pixels among the cells will increase the classification performance.

The weight map is given by

$$w(x) = w_c(x) + w_0 \exp\left(-\frac{(d_1(x) + d_2(x))^2}{2\sigma^2}\right) \tag{4}$$

where $w_c : \Omega \to \mathbb{R}$ is the weight map to balance the class frequencies, $d_1 : \Omega \to \mathbb{R}$ denotes the distance to the border of the nearest cell and $d_2 : \Omega \to \mathbb{R}$ the distance to the border of the second nearest cell. In the original paper, the parameters are set to be $w_0 = 10$ and $\sigma \approx 5$ pixels.

The implications of the U-Net in our project However, as stated before, in our project, we only have classification data, i.e. we only have the 0-1 label for our training set instead of masks indicating the exact place (pixels) of the cancer tissue. Thus, the U-Net may not work in this case. We should still use convolutional networks for classification tasks, where the output of an image is a single class label (or a probability derived from the Softmax function). But the U-Net is so important that we cannot neglect it in deep learning for medical image data processing. Also, it can still provide valuable insights into our project, especially in data preprocessing, in which we can use the image interpolation methods (affine transformation, projective transformation) to create more positive samples (because the data is highly imbalanced) and data augmentation methods to increase the model robustness (affine transformation and projective transformation are used in the data preprocessing of the award-winning models [12], which we will talk about below in this report).

improper for sci. writing

Data Preprocessing Different techniques in the image data preprocessing phase are involved in this project: distorting, flipping (horizontal), cropping, resizing, brightness, contrast, drop out. As stated in the dataset introduction part, the image dataset has left and right breasts so that their direction is the opposite in the mammogram. Also, we cropped and resized the data to exclude the blank part of the mammogram and kept the breast mammogram only. The data were resized from a very high-resolution image (∼5MB for each mammogram) into 1024×1024, 512×512 (unit: pixel) images to reduce the data size, thus boosting the training and preventing the model from overfitting. The contrast and brightness of each image were also adjusted so that the tissues and potential cancer tissues may be easier to be identified. Examples of the preprocessed data are shown in Figure 13. The data are retrieved from [13].

More details of the image data preprocessing can be found at the **Further Study** section of this report. Next, we will move on to discuss the other models that we implement in this project.

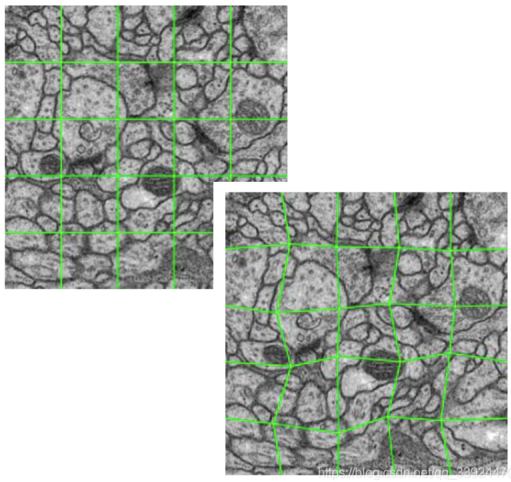

Figure 12: Example of distortion

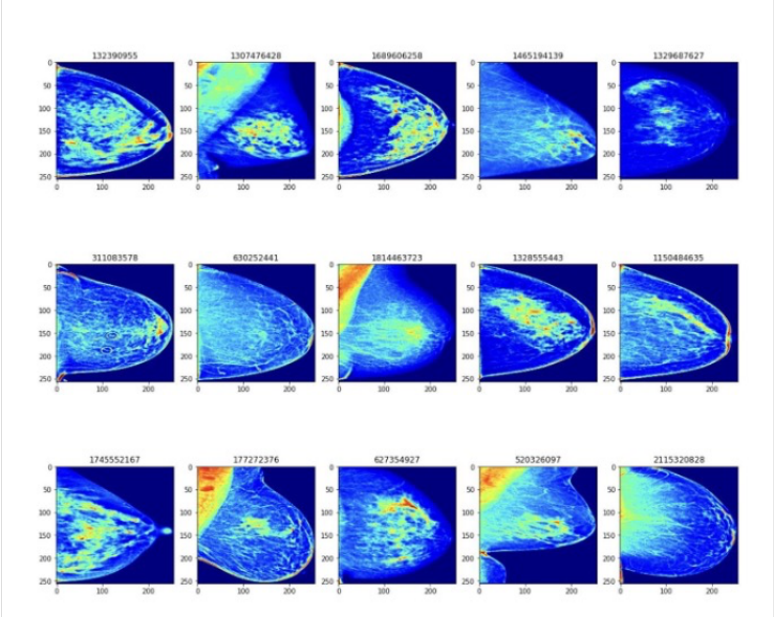

Figure 13: Preprocessed data

2.3.2 Efficient Net

Efficient Net[14] mainly introduces a new scaling method that balances network depth, width, and resolution to achieve better performance. The compound scaling method proposed in this paper is based on the observation that simply increasing the width, depth, or resolution of a ConvNet can lead to diminishing returns in terms of accuracy and efficiency. Instead, the authors propose a more principled approach that balances all three dimensions using a single compound coefficient ϕ.

The method involves scaling the depth, width, and resolution of the network according to the following equations:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi} \tag{5}$$

where d, w, and r represent the depth, width, and resolution of the network respectively. The constants α, β, and γ are determined by a small grid search and subject to the constraint that $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$.

The authors show that this compound scaling method outperforms other single-dimension scaling methods on image classification tasks such as ImageNet. They also demonstrate that this approach can be used to efficiently scale up existing ConvNets without sacrificing accuracy.

Overall, this paper provides a novel methodology for scaling ConvNets that achieves better accuracy and efficiency by balancing all dimensions of network width, depth, and resolution in a principled way using a single compound coefficient. Since this the model is commonly implemented in various competitions in Kaggle, we also adopt it as the baseline model.

2.3.3 Vision-Transformer-Based Model

Architecture of Vision Transformer

Vision transformers[15] are based on the transformer model originally used in natural language processing (NLP), which takes one-dimensional sequences of word tokens as input. However, since images are two-dimensional, vision transformers partition them into smaller two-dimensional patches, which are then treated as word tokens in the transformer model. The input image, with height H, width W, and C channels, is divided into $N = \frac{HW}{P^2}$ patches of size $P \times P$ to align with the input structure used in NLP [16]. Prior to feeding the patches to the transformer encoder, they undergo flattening, sequence embedding, learnable embedding, and patch embedding in the order shown below:

- Flattening each patch into a vector X_{np}, where $n = 1, \ldots, N$, with a length of $P^2 \times C$ was performed.
- The flattened patches were then mapped to D dimensions using a trainable linear projection E, producing a series of embedded image patches.

- Adding to the input sequence, the sequence of embedded image patches was prefixed with a learnable class embedding X_{class} that corresponded to the classification outcome Y.
- Positional information was incorporated into the input by adding one-dimensional positional embeddings E_{pos}, which were also learned during training, to the patch embeddings.

The embedding vectors that are obtained from the operations mentioned above are expressed by z_0:

$$z_o = \left[X_{\text{class}}\,; X_p^1 E; \ldots; X_p^N E\right] + E_{pos} \tag{6}$$

The embedding vectors z_o are input to the transformer-encoder network, which is a stack of L identical layers, to conduct the classification task.At the L-th layer of the encoder output, the classification head is fed with the value of X_{class}. In the pretraining stage, a multilayer perceptron (MLP) with a single hidden layer implementing the GELU nonlinearity is utilized as the classification head. In the fine-tuning stage, a single linear layer is utilized as the classification head.

The vision transformer utilizes the encoder components of the original NLP transformer architecture. The input consists of a sequence of embedded image patches of size 16 x 16, along with positional data and a learnable class embedding. A patch size of 16 x 16 was chosen to strike a balance between performance and computational cost. The learnable class embedding is fed to a classification head connected to the output of the encoder, which produces a classification output based on its state. Figure 14 illustrates the model architecture based on the vision transformer. In this work, the original vision transformer model pre-trained on the ImageNet dataset was modified by replacing its last layer with a flattening layer followed by batch normalization and an output dense layer.

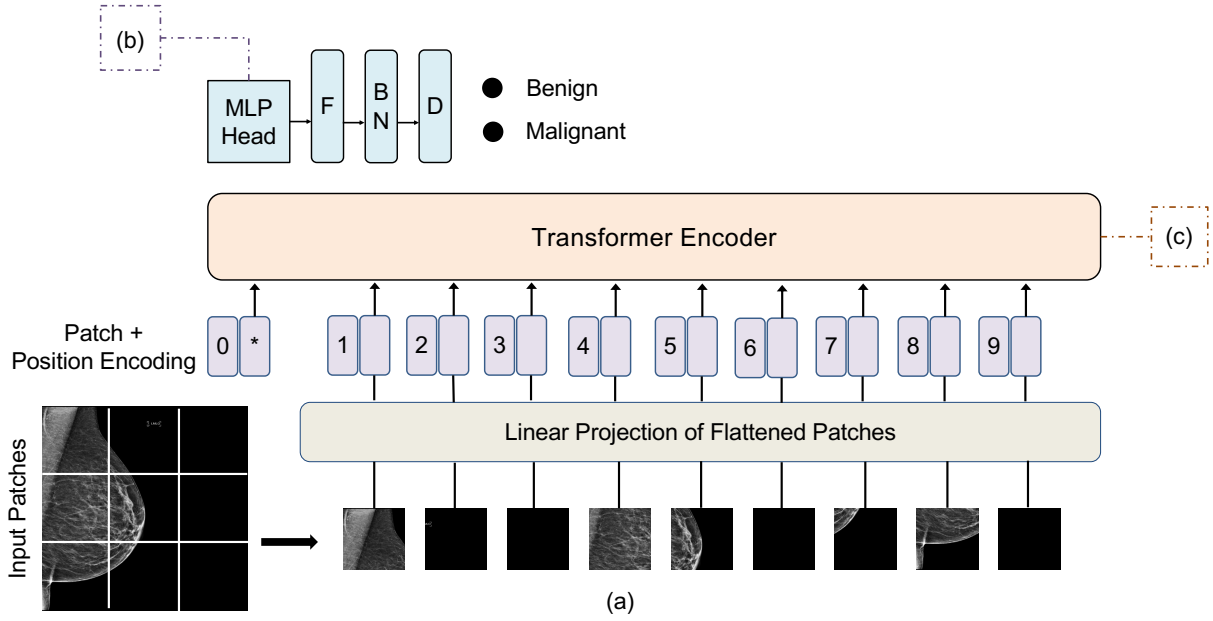

Figure 14: (a) Architecture of Vision Transformer applied to mammography breast cancer detection

2.3.4 Transfer Learning

Transfer learning was used to train vision transformer models on the mammography dataset using pre-trained examples from the vast ImageNet natural image dataset. The goal was to categorize breast mammograms into two classes—those from benign and malignant tissues—using the vision transformer's expertise from the substantial natural image collection. To do this, we removed the pre-trained prediction head and substituted a D $\times$ K feedforward layer, where K = 2 represents the total number of classes in the downstream direction. Our goal in this application of transfer learning was to improve our ability to learn the target function $f_t(\cdot)$ in the target domain D_t by drawing on our prior understanding of the source domain D_s and the learning task T_s. There are m training examples $\left\{\left(x^1, y^1\right), \ldots, \left(x^i, y^i\right), \ldots, \left(x^m, y^m\right)\right\}$

in the ImageNet dataset, with x^i denoting the ith input and y^i the ith label. By then minimizing the objective function in equation 7, where $\langle y^{ij} \mid x^{ij}, W_0, W_1, b\rangle$ is the Softmax output probability function, and b is the bias, W_1 was generated using the weights of the ImageNet pre-trained vision transformer model W_0.

$$J\left(\langle W_1, b \mid W_0\rangle\right) = \frac{-1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{m} y^{ij}\log\left(P\left\langle y^{ij} \mid x^{ij}, W_0, W_1, b\right\rangle\right) \tag{7}$$

3 Results

3.1 Experimental setup

In this task, what we predict is the likelihood of each breast of each patient that getting cancered.We compared the performance of the Vision Transformer pretrained on ImageNet[17] with the Efficient Net on the same dataset: RSNA screening mammography breast cancer detection dataset. The training hyperparameters and model implenmentation details can see Table 1

Hyperparameter	Value
Learning rate	2×10^{-4}
Weight decay	1×10^{-5}
Batch size	64
Epoch number	20
Loss	Binary cross-entropy with logits loss
Optimizer	Adam
Training and validating ratio	4:1

Table 1: Training hyperparameters of the implemented model.

3.2 Evaluation Metric

In this task, we adopt the probabilistic F1 score [18] for performance evaluation. Probabilistic F1 score is an extension of the traditional F1 score that takes into account the uncertainty of a model's predictions. It is a metric used to evaluate classification models in natural language processing. The Probabilistic F1 score considers both precision and recall, as well as the confidence level of the model's predictions. This allows for a more thorough evaluation of the model's performance, especially in cases where there are razor-thin margins and low-resource test sets. The mathematical formulation of probabilistic F1 score is shown below.

need to define the terms

$$pTP_{C_j} = M\left(\mathbf{x_i}, C_j\right) * I_{C_j = y_i} \tag{8}$$

$$pFP_{C_j} = M\left(\mathbf{x_i}, C_j\right) * I_{C_j \neq y_i} \tag{9}$$

$$\begin{aligned} p\,\text{Precision} &= \frac{pTP}{pTP + pFP} \\ p\,\text{Recall} &= \frac{pTP}{pTP + pFN} \end{aligned} \tag{10}$$

Here $\mathbf{x_i}$ denotes the ith sample, C_j denotes the jth possible class, y_i denotes the ith label. pTP, pFP, pPrecision, pRecall represents the probabilistic true positive,probabilistic false positive,probabilistic precision, probabilistic recall.

$$pF_1 = \frac{2pPrecision * pRecall}{pPrecision + pRecall} \tag{11}$$

3.3 Experimental Results

The experiment results are shown in Table2.Five-fold cross-validation was used to compare the model performances. From the table, we can see that the vision-transformer-based transfer-learning approach provided the highest quantitative and statistical measures for predicting the likelihood of breast mammograms as being from benign or malignant tissues. This proves the effectiveness and quality of the vision-transformer-based transfer-learning approach for detecting breast cancer from mammograms. The possible reason is that vision transformer has the ability to capture global information from the early layers and the deep self-attention mechanism that enables features in each patch to be carefully analyzed for decision-making.

Model	pF1 Score	pPrecision	pRecall	pAUROC
EfficientNetV2[19]	0.45	0.45	0.46	0.44
Vit-base	0.52	0.53	0.52	0.52

Table 2: Testing Result on RSNA Screening Mammography Breast Cancer Detection Dataset

4 Future Work & Conclusion

4.1 Limitations

Lack of in-depth Feature Analysis
For each patient, there are four mammography images(Left/Right with two kinds of image-forming conditions), so the features may have some inner correlations. However, in the project, we omit this information and do not come up with a method to extract more useful information from the inner correlation between these four images.

4.2 Further Study

This project was originally a Kaggle competition ending on 28th Feb. 2023 [8]. The award-winning model was made public according to Kaggle's rules. In this section, we analyzed some notable data preprocessing techniques and Deep Learning architectures in the award-winning model.

Award-winning Data Preprocessing Methods The award-winning model employed the following data preprocessing methods:

- Original image arrays were converted into 2048 x 2048 x 1
- Images were then cropped to exclude blank space (a simple rule-based cropping line as shown in Figure 15)
- YOLO model was trained to generate breast bounding-box (compared to simple rule-based breast extraction, YOLO cropped images usually have a smaller region, which seemed to prevent our models from overfitting)
- Affine transformation, Vertical/Horizontal flip, brightness/contrast, blur, CLAHE, distortion, dropout

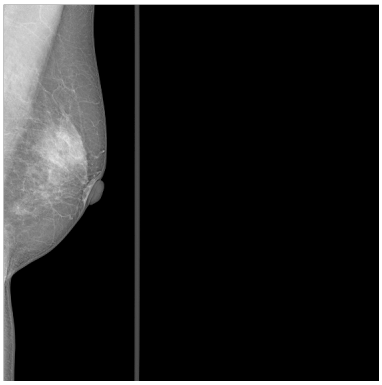

Figure 15: Breast Bounding Box

Award-winning Architecture: the Multi-View Fusion Model In recent years, the multi-view fusion model is really a hot topic. For example, in Mei et al.[20]'s work, they proposed the pyramid image fusion method to train the model. The pyramid structure allows for the extraction of features at different scales, which are then synthesized using a multimodal strategy to improve the robustness and accuracy of the method. The idea of cropping the image to different scales to catch both the global features and the local features may also be suitable for this project.

In the competition, the creators of the award-winning model encoded the two views of the mammograms globally to get the saliency map and multiply it with the locally encoded features of the cropped ROI of the mammograms. The model then concatenates the global and local features together to get the classification results. Two visualizations of the architecture are shown in Figure 16 and Figure 17.

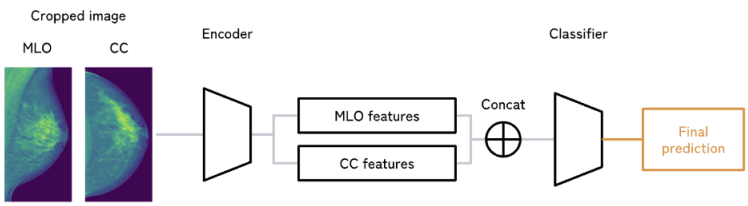

Figure 16: Award-winning Architecture 1

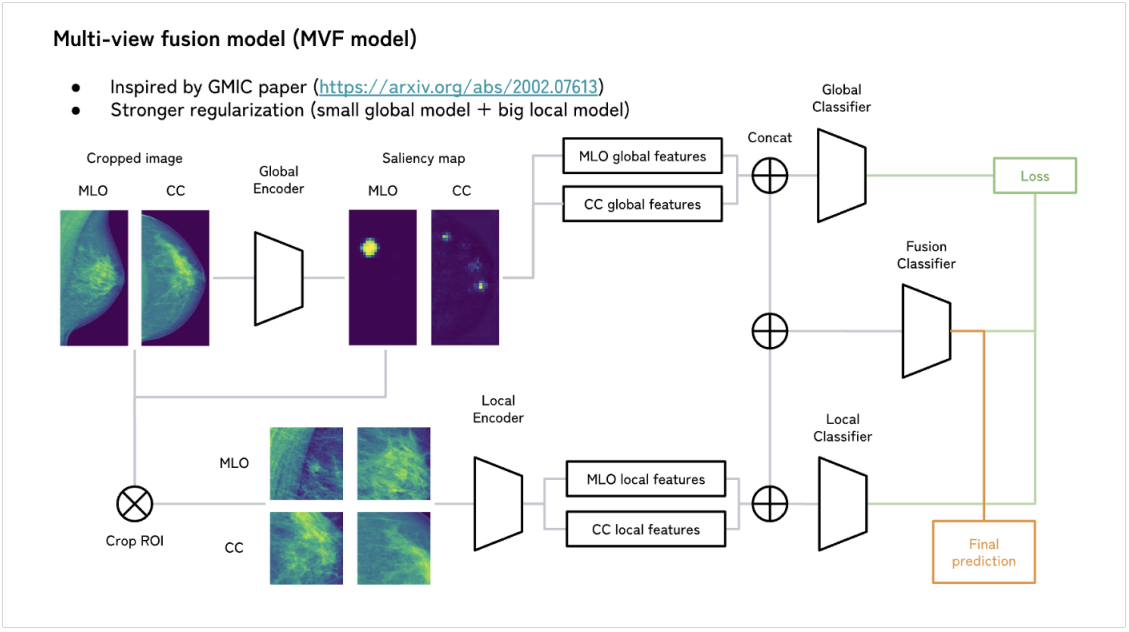

Figure 17: Award-winning Architecture 2

5 Author contributions

Chenglin Zhang: Implementing the Transfer Learning model. Report Writing: includes the section related to U-Net, Data Preprocessing, Further Study of Data Preprocessing, and Model analysis.
Lihui Chen: Implementing the Unet model. Report Writing: includes the section related to Introduction, Literature Review, Exploratory Data Analysis, Implemented Model and Methodology.
Yijia Xue: Implementing the Vision Transformer and Transfer Learning model. Report Writing: includes the section related to Efficient Net/ Vision Transformer/Transfer Learning/Results/Model Limitations

6 Acknowledgement

We thank Prof. Dongmian Zou for the generous and insightful guidance in this session's Deep Learning course, that it will empower our future learning greatly.

References

[1] Sergiusz Łukasiewicz, Marcin Czeczelewski, Alicja Forma, Jacek Baj, Robert Sitarz, and Andrzej Stanisławek. Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. *Cancers*, 13(17):4287, 2021.

[2] U.S. Cancer Statistics Working Group. Rate of new cancers in the united states, 2022.

[3] Breastcancer.org. The financial effects of breast cancer., 2022.

[4] Marina Milosevic, Dragan Jankovic, Aleksandar Milenkovic, and Dragan Stojanov. Early diagnosis and detection of breast cancer. *Technology and Health Care*, 26(4):729–759, 2018.

[5] Benjamin O Anderson. Global summit consensus conference on international breast health care: guidelines for countries with limited resources. *The breast journal*, 9:S40–1, 2003.

[6] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pages 1–4. IEEE, 2017.

[7] Basem S Abunasser, Mohammed Rasheed J AL-Hiealy, Ihab S Zaqout, and Samy S Abu-Naser. Breast cancer detection and classification using deep learning xception algorithm. *International Journal of Advanced Computer Science and Applications*, 13(7), 2022.

[8] Kaggle. Rsna screening mammography breast cancer detection, 2023.

[9] Society of Nuclear Medicine and Molecular Imaging. Illustration of how a cancer may be difficult to identify on a mammogram for women with dense breast tissue, 2019.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[12] RABOTNIKUMA on Kaggle. 6th place solution: Multi-view multi-lateral multi-stage approach, 2023.

[13] PAUL BACHER on Kaggle. Rsna bcd 1024x512 preprocessed, 2023.

[14] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Gelan Ayana and Se-woon Choe. Buvitnet: Breast ultrasound detection via vision transformers. *Diagnostics*, 12(11):2654, 2022.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] Reda Yacouby and Dustin Axman. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems*, pages 79–91, 2020.

[19] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[20] Shuang Mei, Hua Yang, and Zhouping Yin. An unsupervised-learning-based approach for automated defect inspection on textured surfaces. *IEEE Transactions on Instrumentation and Measurement*, 67(6):1266–1277, 2018.