Multiple Multi-Modal Methods of Malignant Mammogram Classification (M6C)

Christopher Vattheuer, Jase Tran
Department of Computer Science, University of Manitoba

***Abstract*—Breast Cancer is a tragic disease which affects approximately 1 out of 8 women. Mammograms are X-Ray scans used to detect early breast cancers that are often susceptible to human error. As such, a strong emphasis has been placed on creating novel techniques to detect early cancers in mammograms scans. Analyzing and detecting breast cancer is a multi-modal task which combines patient health information with their scan data. In this paper, we propose multiple multimodal classification techniques and use them to detect cancers in the RSNA Breast Cancer Dataset. In this work, we perform our own preprocessing and propose our own novel architectures to classify raw mammogram scan data. Our findings indicate that our destructive patching and embedding concatenation techniques lead to strong accuracy scores and significantly faster convergence. Additionally, we compare and validate multiple techniques for aggregating predictions made on multiple scans to determine whether a patient has cancer. Ultimately, we achieved a balanced accuracy of 70.2% on this task.**

Keywords—breast cancer, mammogram, multi-modal, vision transformers, resnet50

INTRODUCTION AND RELATED WORKS

Breast cancer is the most common cancer among women worldwide [1]. The American Cancer Society estimates approximately 13% of women (1 in 8) will be diagnosed with invasive breast cancer, and 3% (1 in 39) will die from the disease in their lifetime in the United States [1]. Early detection is key to improving survival rates. In the same report, the American Cancer Society found that the five-year relative survival rate for early-stage (stage I) breast cancer is >99%, compared to only 29% for late-stage (stage IV) breast cancer [1]. A tool critical for the detection of breast cancers are mammograms, which are X-rays of the breast which can be used to detect early cancers. A key drawback of mammograms is that despite the fact they can be used to detect cancer, they are notorious difficult to analyze, and many cancers often go unnoticed. Part of the difficulty of cancer classification is due to the variability of tissue density and the presence of calcifications. Visual analysis of mammograms is also a time-consuming process, requiring years of specialized training subject to human error.

In recent years, computer vision has emerged as a powerful tool which can aid in early cancer classification in mammograms. The usage of computer vision in such a task allows for analysis of mammograms which are fast, consistent, and objective. Additionally, computer vision can allow for doctors to make better use of their time by having an AI highlight difficult cases or potential cancers allowing doctors to focus their time and attention on important factors. General techniques that have been used for mammogram classification include texture analysis, feature extraction, and machine learning [2]. Multi-view mammogram classification, which uses multiple views of the same breast to improve classification accuracy, is a commonly used approach [3] [4] [5]. This approach is similar to a mammogram analysis technique commonly used by doctors, where scans across each laterality are analyzed side by side often comparing new and old scans. Many approaches build off transfer learning, where models are pre-trained on medical or other image data, which leverages their domain knowledge to improve classification accuracy [6] [7] [8].

Much of the recent boom seen in computer vision can be attributed to Convolutional Neural Networks (CNNs). CNNs work by creating kernels that perform scan images looking for various patterns in patches of pixels. This approach bears similarity to how the human brain processes information, where centers early in the brain specifically look for edges and various textures in vision. Just as the human visual system can recognize an object despite changes in its size, position, or orientation, CNNs are also capable of invariant recognition. CNNs have been successfully implemented across a wide range of tasks, from medical imaging [9] [10] [11] [12] [13] to autonomous vehicles [14] [15] and astronomy [16]. In the context of medical imaging, the advantages of CNNs compared to conventional machine learning classifiers include not requiring hand-crafted feature extraction or segmentation of tumors or organs by human experts [17]. However, CNNs also require very large amounts of data to train their millions of learnable parameters to classify images, and thus are more computationally expensive [17]. Unlike in other applications of computer vision, where training images are often plentiful, large data requirements can be a big challenge in medical imagery since building such large datasets is costly and demands an enormous workload by experts and is often inaccessible due to ethical and legal issues regarding patient privacy [17]. Like other deep neural networks, CNNs have also been found to be vulnerable to

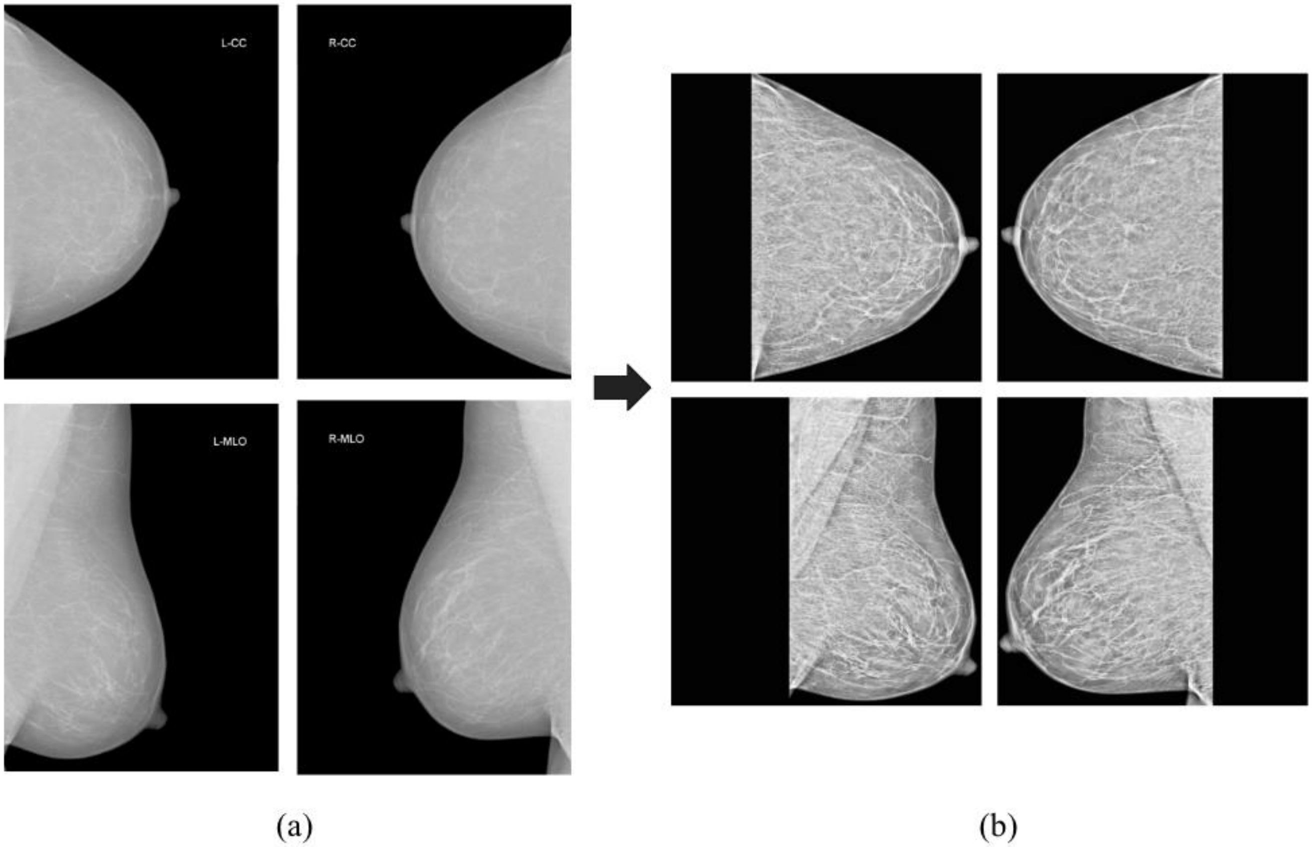

Figure 1: Original images before (a) versus after cropping and enhancements (b). Images are not to scale.

adversarial examples, which are carefully designed inputs that cause the network to change output often without a human being able to notice anything different in the input example. While the impact of adversarial examples in the medical domain is unknown, their very existence indicates that CNNs and artificial networks have fundamental differences to humans in the way they perceive images [17].

An exciting and recent finding in Computer Vision has been vision transformers (ViTs). Unlike traditional convolutional neural networks (CNNs), ViTs build off concepts coming from natural language processing. ViTs make use of self-attention mechanisms to capture long-range dependencies between patches of an image. This approach has been shown to be effective in natural image classification tasks and has recently been applied to medical image analysis, including mammogram classification [18] [19]. One of the strengths of ViTs is their ability to learn representations without the need for hand-crafted features or prior domain knowledge [20]. This makes ViTs highly adaptable to various image analysis tasks. Additionally, ViTs have shown to be effective in capturing subtle image features that may be difficult to detect using traditional methods [21]. However, ViTs are typically much bigger and require tremendous amounts of training data and computational resources to outperform many CNN based techniques. Since ViTs do not have inductive biases, the model's pre-existing knowledge or biases about the problem at hand, which are built into CNNs, they tend to struggle with medium-sized datasets and typically require pretraining [20]. Furthermore, while the base ViT is good at capturing long-range dependencies between patches, it disregards the local feature extraction and thus does not model local information very well. Follow-up solutions to this limitation have focused on incorporating spatial information directly into the transformer architecture and yielded promising results [20]. Despite these challenges, ViTs have demonstrated serious potential across many domains in computer vision such as object detection [22], segmentation [23], image generation [24] and various multimodal tasks [25]. Given a good track record in general computer vision tasks, ViTs seem to be a promising candidate for improving the accuracy of mammogram classification and as such we use it as one of our base models.

While in many Computer Vision tasks, the information contained within an image, such as one from CIFAR10, is typically sufficient to make the association between the image and its corresponding label. Unfortunately, it is not so clear cut with complex tasks such as medical images classification. Tasks such as mammogram cancer detection often require doctors to not only analyze mammograms, but also understand characteristics of their patient such as their age and family history. It may very well be the case that characteristics such as breast tissue density, the presence of calcifications and masses may naturally differ between age groups and thus the way doctors evaluate scans depends on the patient, as some signs which are atypical at younger ages may be typical at older ages, and vice versa. For these reasons, it may be the case that classification techniques that only make use of image data and remain agnostic to other key information may perform worse on more complex tasks. There are many techniques used to integrate information across many domains in classification. Typical multimodal medical image solutions involve fusing different sources of images such as MRI, CT, X-Rays, PET, etc. together to produce a resultant image with excess and complementary information [26]. While X-Rays are the only image modality in our dataset, we consider the patient's information as an additional modality that can provide context to the X-Rays when fused together. In this paper, we investigate the effectiveness of various methods of multi-modal classification on mammogram image and tabular patient data.

OUR METHODS

The Dataset and Data Preprocessing

Mammogram data, as with much of medical data is often difficult to access. The most common dataset used, the NYU Breast Cancer Screening Dataset [27] is private and requires special permission to access. Luckily, a public mammogram dataset was recently released on Kaggle as part of a $50,000 mammogram classification competition. The dataset is called the RSNA Screening Mammography Breast Cancer Detection Challenge [28] and contains 54,706 DICOM (.dcm) scans across 11,913 patients as well as metadata about each patient and scan. Of these scans, 1158 (approximately 2%) were identified as containing cancer, and these were distributed across 486 patients. The mammogram scans were captured using one of six different views (CC, MLO, ML, LM, AT, LMO), across ten different machines from two different sites. In addition to the images, the dataset also includes tabular data with patient information and image metadata, such as the patient's age, whether a biopsy was performed, whether it was invasive, the BIRADS score, if there is an implant, density level, and whether the image is a difficult negative case. Despite only containing 54k scans, each is very detailed, and the total dataset takes up just over 314GB. The dimensions of the scans are non-square and varied depending on the id of the machine which performed the scan. At full size, each scan is around 3000 pixels in each dimension. Additionally, depending on the machine that performed the scan, different idiosyncrasies appear. This includes tags placed on the scan and different image characteristics. Because of the incompatible scan formats, very large scan sizes and lack of scan contrast, quite a bit of preprocessing is necessary before classification can begin. It has previously been shown that attempting to classify mammograms with limited preprocessing yields very poor results [3]

In mammogram scans, much of the content of the scan tends to be a black background, whereas the informative portion containing the breast X-Ray tends to only use up a small fraction of the scan. A concern was that if we were to naively scale down scans to 224x224 pixels, then this would relegate the important breast information to a very small and compressed area, leading to poor performance. For this reason, instead of simply scaling down images, we decided to use a cropping technique from [29] that would instead crop the scan around the breast. That way, when scans were later scaled down to 224x224, much more of the original detail in the scan could be preserved. The cropping technique we used makes use of edge detection and centering, followed by our own padding technique to make the images square.

We now had 224x224 images of breast scans for each patient, however each image had very low contrast making it difficult to discern detail. A common technique across many mammogram classification tasks is to increase the contrast of the scans to bring out this important detail [3].To do so, we applied the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique. Using CLAHE requires providing a strength and kernel size, and finding parameters that felt appropriate was entirely done by hand by, through visual analysis of the resulting image. Using CLAHE also produces some image grain, which we reduced using a combination of bilateral filtering, and fast Non-Local Means denoising. The goal was to enhance the features that are important for classification and reduce graininess in the images.

A key weakness of our preprocessing is that we are by no means experts in mammogram analysis, and techniques such as CLAHE, bilateral filtering and fast non-local means denoising all can be tuned. This means that our iterative selection of these parameters may be flawed leading to reduced accuracy later in our pipeline.

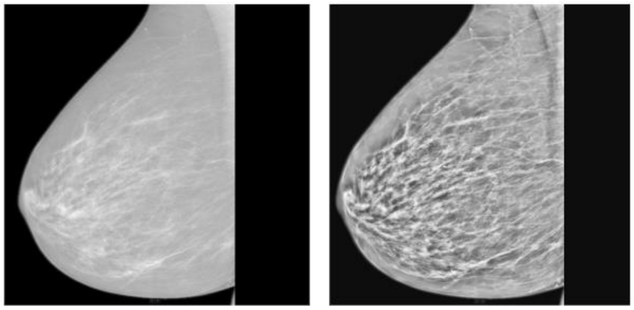

Figure 2: Cropped image before (left) versus after (right) our enhancements.

For preprocessing the tabular data, we performed feature selection, one-hot encoding, and data normalization. As the positive class (cancer) is rare in our dataset, we rebalanced our data by oversampling the minority class using replacement. To avoid overfitting on our heavily oversampled cancer class, we optimistically applied various image augmentation techniques, such as random cropping, flipping, brightness changes, blurring, Gaussian noise, and grid dropout.

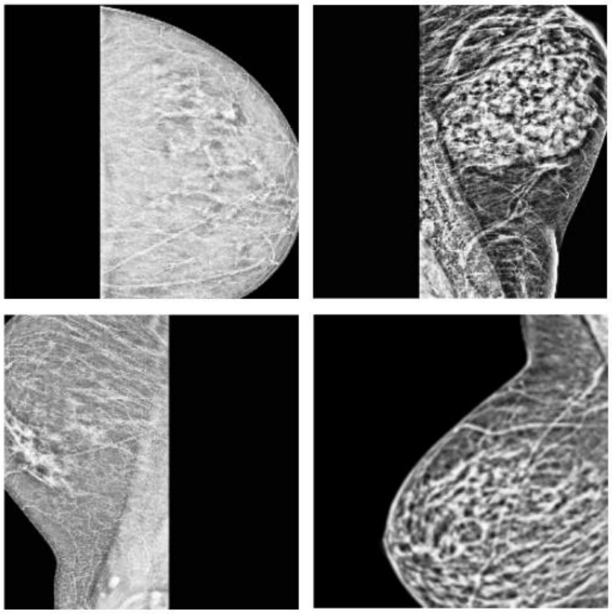

Figure 3: Augmented Training Samples

Our Architecture

When determining what architecture best fit this problem, an immediate first idea was to perform multi-view classification. However, each patient in the RSNA dataset contained a variable number of scans. Some patients had only 1 scan while others had up to seven each performed at various views. This meant that there was unfortunately no guarantee that each scan could be processed in pairs group up by laterality and views. Additionally, this meant that predicting whether a breast contained cancer required aggregating information across a variable number of scans.

Given the variable number of scans per patient, we break our classification strategy into two key stages. The first stage analyzes individual scans to determine the likelihood of cancer in each. The second stage aggregates probability information across multiple scans to produce a prediction as to whether a breast has cancer. Across both stages, we try multiple techniques for each to see what performs best.

Stage 1 Techniques

Method 1: Naïve Classification

In this technique, we make no use of the additional patient metadata, and use either a ResNet50 or a ViT to classify patient images. A key benefit of this technique is its simplicity, as it allows for simple training of models on a binary classification task. A key drawback is that this technique makes no use of any patient information while processing images which may otherwise influence how a professional would classify images.

Method 2: Destructive Patching

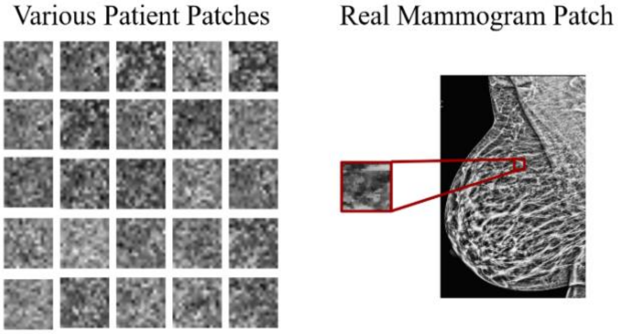

Figure 6a: Produced Destructive Patches vs Real Mammogram Patch

In this technique, a neural network first analyzes patient information to produce a 16x16 ‘Patient Patch’. This generated 16x16 image is then placed in the top right corner of the scan. The idea behind this technique was that ViTs break down images into 16x16 patches whose attention between each other influence prediction outcomes. Thus, producing a visual patch which embeds patient information forces understanding of a patient into the visual processing of a scan. This technique should

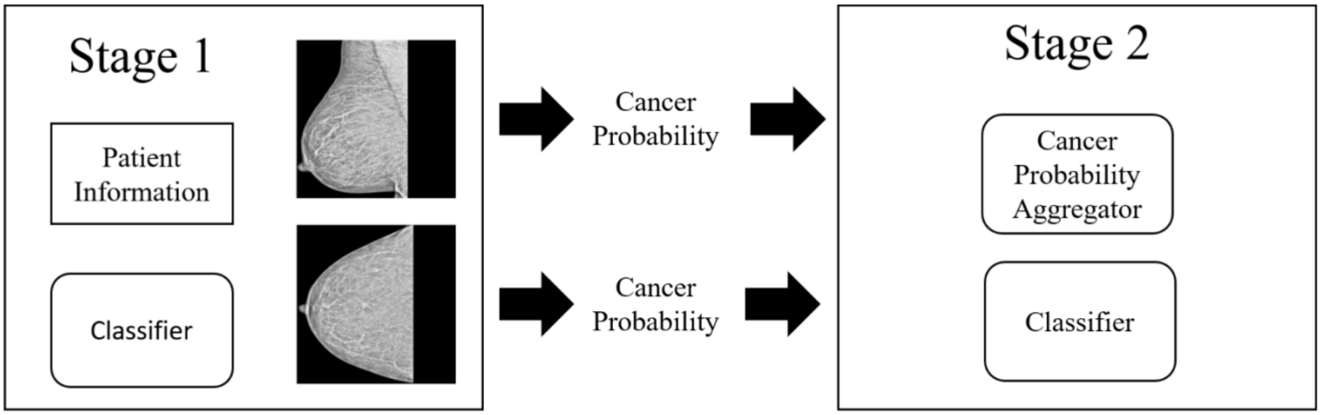

Figure 4: Overview of Cancer Detection Stages

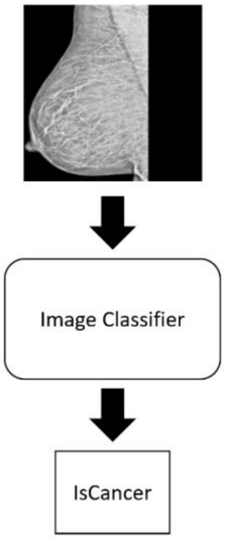

Figure 5: Summary of Naïve Classification

allow for a network to get a more in-depth understanding of a scan allowing for greater accuracy than Naïve classification. In practice we found that this technique produced patches that were similar to real patches in mammogram scans.

A key weakness of this technique is that it alters the source image in a destructive way. Should it turn out that the patch covers key information in a scan such as a telltale sign of breast cancer then incorporating patient information in this way may reduce quality of predictions. This technique is tried for both a ResNet50 and a ViT.

Method 3: Early Concatenation

This technique is specific to ViTs and requires a patient embedding to first be produced using a neural network. This embedding is then concatenated to early patch embeddings produced to by the ViT. A key concern with Destructive Patching was that due to the positional embeddings of patches, the model may struggle to understand the relation of the top right corner Patient Patch with further away patches in the image. To combat this, it was hypothesized that uniformly concatenating the patient embedding to each patch may allow for a more comprehensive understanding of images. A key weakness of this technique is that it increases the size of ViTs, greatly slowing down processing time and removing any possibility of using pretrained ImageNet ViT models available on Pytorch.

Method 4: Bidirectional Cross Attention

Making use of findings from a recent work, UNINEXT [30] which standardizes multimodal classification techniques across all perception tasks, we propose a method inspired by their technique. In this method, we first use a neural network to produce a patient embedding but combine it with the patch embeddings produced by a ViT through Bidirectional Cross Attention Module. This technique was proposed in the paper, although it was done between image embeddings and text embeddings produced by large language models. We hypothesized that this technique may fail early on in training due to uninformative patient embeddings being produced.

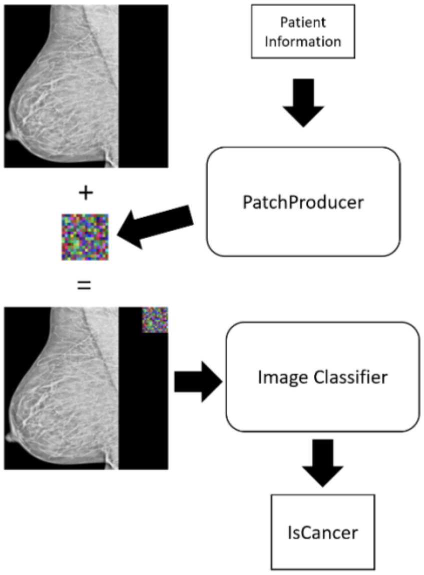

Figure 6: Summary of Destructive Patching

Stage 2 Techniques

Once probabilities are produced for each scan, we explore multiple methods for aggregating per scan cancer probabilities. The outcomes of each of these techniques are evaluated by Random Forest, Multi-Layer Perceptron, SVM, KNN and Ada-Boost classifiers to explore their effectiveness.

Method 1: Average Probability Feature

In this technique, the average probability of cancer across each scan is gathered and passed as a single float value to a classifier in addition to the other tabular patient information. This technique generalizes the results seen across all scans at the cost of heavily simplifying it and potentially hiding very high or very low probability values on individual scans.

Method 2: Max Probability Feature

This method is highly similar to Method 1, however instead of returning the average value in addition to patient features it returns the maximum probability seen on any scan.

Method 3: Min Probability Feature

Like method 1 and 2, this one simplifies all cancer probabilities into the smallest across each scan for a breast.

Method 4: Average, Min, Max Probability Features

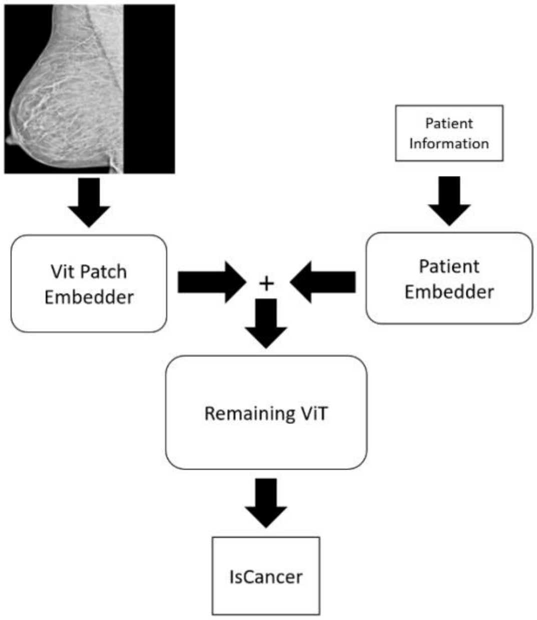

Figure 7: Summary of Early Concatenation

To maintain more relevant information about the scans done on a breast, this technique returns not only the average, minimum or maximum, but rather all three in combination with the other patient features. A hypothesized weakness of the first three methods is that they may oversimplify cancer probabilities, and this method helps give a classifier more context.

Method 5: Padded Probability Features

Whereas Method 1, 2 and 3 each simplify probability information to make classification easier, this method attempts to preserve all prediction information done by the Stage 1 classifier by passing all prediction information, padded out for consistent length to a classifier. This technique preserves the most information about patient scans but may be ineffective due to many features with a relatively small amount of data to train on.

Experiments and Results

Stage 1 Results

Balanced Accuracy and Macro F1 scores of methods.

Method	ViT		ResNet50	
Naïve Classification	62.1%	0.361	65.1%	0.362
Destructive Patching	65.3%	0.465	67.5%	0.463
Early Concatenation	64.3%	0.480	-	-
Bidirectional Cross Attention	50%	-	-	-

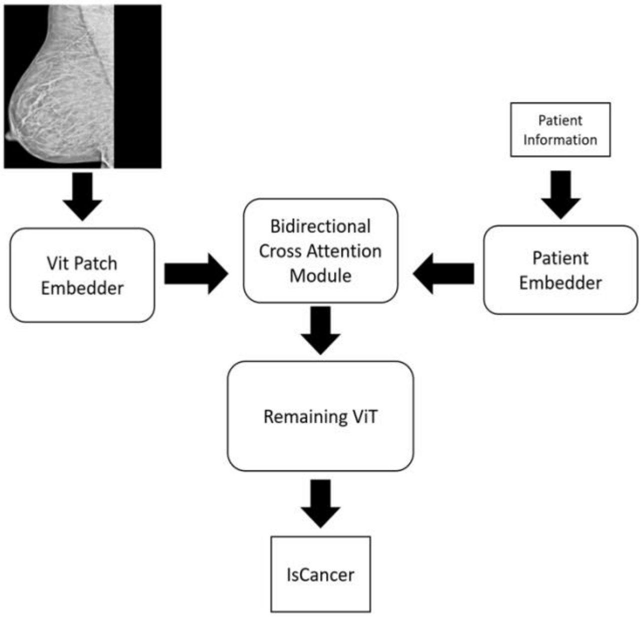

Figure 8: Summary of Bidirectional Cross Attention

When analyzing these results, we were first surprised to find that the ResNet50 outperformed the ViT across all our techniques. We believe this difference exists due to the fact that much of the benefit of using ViTs come from long training sessions and extensive pretraining. Due to our limited access to medical data and computational resources, we were unfortunately unable to make use of a medical pretraining stage for ViTs.

We were quite pleased to find that the multi-modal techniques allowed for greater classification accuracy on single images, with Destructive Patching being our most successful technique, giving +3.1% accuracy on ViTs and +2.4% accuracy on the ResNet50 when compared to Naïve Classification. What was surprising about this result was the degree to which the ResNet50 was able to make use of the Patient Patch. The Patient Patch was specifically designed with ViTs in mind due to their 16x16 patches and attention mechanism, but this technique seemed to also translate to the ResNet50, greatly boosting its classification abilities. This accuracy improvement demonstrates that producing an image-based embedding of non-image data can aid in multi-modal image classification.

When examining the results of Early Concatenation, we found that it boosted accuracy at the cost of greatly slowing down epoch times. While not as performant as Destructive Patching, this technique showed promise, increasing balanced accuracy from 62.1% to 64.3%. Unfortunately, this method was designed exclusively for ViTs and as a result could not be tested on the ResNet50.

Finally, we were disappointed by the limited results seen in the Bidirectional Cross Attention technique. Across many different training parameters that we grid searched on, the model would fail to catch on and the initial training accuracy would refuse to increase. It is quite possible that if we had trained for much longer and grid searched more exhaustively that this model could have caught on and trained well. Despite the limited results demonstrated by this technique, we believe that attention-based methods for tabular and image data classifications hold great potential.

An unexpected but very useful property emerged during multi-modal training, in that compared to Naïve Classification, Methods 2 and 3 converged several times faster. These multi-modal techniques would often achieve accuracy levels in the first epoch that would take naïve classifier many more epochs to reach across both the ResNet50s and ViTs. The combination of both increased accuracy and much faster convergence highlights the benefit of these multi-modal techniques.

Stage 2 Results

Across each of the Stage 1 Methods, we then apply the Stage 2 techniques whose balanced accuracy and f1 scores are recorded below.

Naïve Classification ViT

Aggregate Type	Best Classifier	Balanced Accuracy	Macro F1
Average Probability Feature	Random Forest	66.5%	0.380
Max Probability Feature	Random Forest	66.3%	0.402
Min Probability Feature	Random Forest	66.7%	0.413
Average, Min, Max Probability Feature	Random Forest	63.7%	0.435
Padded Probability Features	SVM	50.2%	0.504

Naïve Classification ResNet50

Aggregate Type	Best Classifier	Balanced Accuracy	Macro F1
Average Probability Feature	MLP	67.7%	0.407
Max Probability Feature	Random Forest	67.8%	0.386
Min Probability Feature	Random Forest	66.4%	0.468
Average, Min, Max Probability Feature	Random Forest	58.9%	0.449
Padded Probability Features	Random Forest	50.0%	0.493

Destructive Patching VIT

Aggregate Type	Best Classifier	Balanced Accuracy	Macro F1
Average Probability Feature	Random Forest	63.0%	0.435
Max Probability Feature	Random Forest	65.2%	0.470
Min Probability Feature	Random Forest	64.1%	0.459
Average, Min, Max Probability Feature	Random Forest	61.9%	0.466
Padded Probability Features	SVM	50.5%	0.506

Destructive Patching ResNet50

Aggregate Type	Best Classifier	Balanced Accuracy	Macro F1
Average Probability Feature	Random Forest	69.8%	0.458
Max Probability Feature	Random Forest	67.1%	0.387
Min Probability Feature	Random Forest	70.2%	0.461
Average, Min, Max Probability Feature	Random Forest	67.5%	0.421
Padded Probability Features	KNN	50.8%	0.511

Early Concatenation ViT

Aggregate Type	Best Classifier	Balanced Accuracy	Macro F1
Average Probability Feature	Random Forest	63.49%	0.4814
Max Probability Feature	Random Forest	65.7%	0.4675
Min Probability Feature	Random Forest	64.8%	0.466
Average, Min, Max Probability Feature	Random Forest	57.4%	0.52529
Padded Probability Features	KNN	50%	0.4939

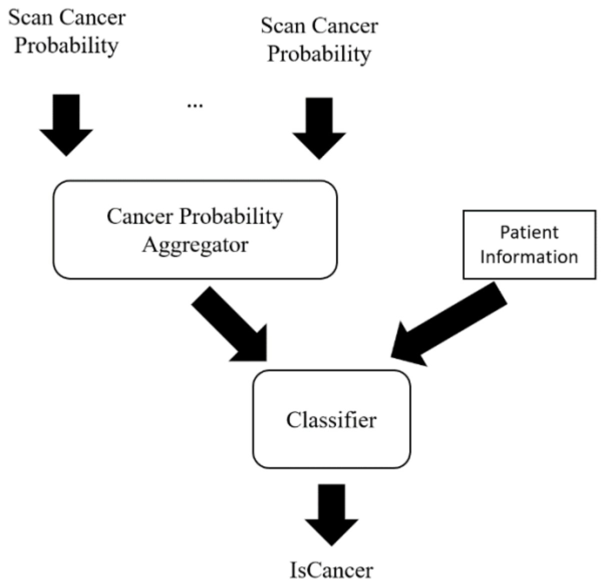

Figure 9: Summary of Stage 2 Process

Overall, the findings in stage two yielded a few very interesting trends. The classifier which was nearly consistently the best to use and the fastest to train was the Random Forest Classifier. Across all Stage 1 techniques, the best aggregate type to use was consistently either the Max or Min Probability Feature. While there were concerns that these methods would oversimplify breast scans, they performed significantly better than methods which returned more information such as Average, Min, Max Probability and Padded Probability Features.

Unfortunately, padding out probability features consistently resulted in poor accuracy, even when an RNN was applied to the task using a variable number of scans per patient. These results suggest that simple metrics which generalize the results of several scans may yield stronger results than examining each scan, at least for a small dataset.

Another interesting trend was that the ViT based approaches appeared to make less use of the second stage methods than the ResNet based approaches. It was theorized that the ViT would be able to make better usage of the patient patch and through its attention modules than the ResNet50. It would then make sense that the second stage would have limited addition gain when aggregating prediction scores. On the contrary, the best performing ResNet approach benefited both from the Stage 1 Destructive Patient Patch and the Stage 2 Minimum Probability Aggregation.

In general, our findings indicate that adding a second stage to aggregate probabilities across multiple scans of the same breast is highly beneficial for cancer detection. Overall, ViT based approaches were able to increase their accuracy to 66.7%, and Resnet50 based approaches achieved the highest accuracy across all methods at 70.2% accuracy.

The best performing technique, ResNet50 with Destructive Patching and Minimum Probability Aggregation had the following characteristics of its predictions on the test set:

Characteristic	Value
True Positives	40
False Positives	693
True Negatives	1634
False Negatives	17
True Positive Rate (Sensitivity)	0.701
True Negative Rate (Specificity)	0.702
Precision	0.05457

To baseline the efficacy of our technique we compare it to information about the predictive accuracy of doctors in mammogram classification. Research has shown that after screening when patients are called back in for additional tests, they have cancer fewer than 10% of the time [31]. Using our technique to flag patients who may have cancer would result in 5.45% of those called in to have cancer, which is moderately under the baseline for human performance. However, it is worth noting that when doctors classify mammograms, they may have significant more patient information and context than provided to our algorithm. Where our algorithm falls short is in our false negative rate of 0.2978, where the human baseline for false negative rate is approximately 0.125 or 1/8 [32].

Limitations And Future Work

There are multiple key limitations of our work which create room for future work. The first limitation comes from a lack of time and computational resources. Large and powerful models like ViTs benefit from having long training times on very large amounts of data. We unfortunately lack the resources required to train these models for extended periods of time and as such had to train for fewer epochs and perform less exhaustive grid searches than preferable.

Additionally, much of the research done in Mammogram classification uses large medical datasets such as the NYU Breast Cancer Screening Dataset. This private dataset comes with over one million examples which were preprocessed by specialists. When compared to our method which made use of a dataset 1/20th the size and had handmade preprocessing techniques, it was possible that some of our preprocessing and data augmentation was in fact hurting performance. Trying our multimodal methods on a large, standardized dataset would yield more informative results which could be compared to state-of-the-art techniques.

Conclusion

In this paper, we explore our own mammogram preprocessing technique, as well as several techniques for combining image and tabular patient information for the detection of cancer. By experimenting with several methods for the combination of image and tabular data, we ultimately achieve a strong result of 70.2% balanced accuracy on the Kaggle RSNA Breast Cancer Competition Dataset and in addition provide some theoretical and practical justification for our novel techniques.

References

[1] A. N. Giaquinto, H. Sung, K. D. Miller, J. L. Kramer, L. A. Newman, A. Minihan, A. Jemal and R. L. Siegel, "Breast Cancer Statistics, 2022," *CA: A Cancer Journal for Clinicians,* vol. 72, no. 6, pp. 524-541, 2022.

[2] P. Oza, P. Sharma, S. Patel and A. Bruno, "A Bottom-Up Review of Image Analysis Methods for Suspicious Region Detection in Mammograms.," *Journal of imaging,* vol. 7, no. 9, p. 190, 2021.

[3] H. Nasir Khan, A. R. Shahid, B. Raza, A. H. Dar and H. Alquhayz, "Multi-View Feature Fusion Based Four Views Model for Mammogram Classification Using Convolutional Neural Network," *IEEE Access,* vol. 7, pp. 165724-165733, 2019.

[4] L. Xia, J. An, C. Ma, H. Hou, Y. Hou, L. Cui, X. Jiang, W. Li and Z. Gao, "Neural network model based on global and local features for multi-view mammogram classification," *Neurocomputing,* vol. 536, pp. 21-29, 2023.

[5] K. J. Geras, S. Wolfson, Y. Shen, N. Wu, S. G. Kim, E. Kim, L. Heacock, U. Parikh, L. Moy and K. Cho, "High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks," 2018.

[6] S. Vesal, N. Ravikumar, A. Davari, S. Ellmann, A. Maier, A. Campilho, F. Karray and B. ter Haar Romeny, "Classification of Breast Cancer Histology Images Using Transfer Learning," in *Springer eBooks*, Springer International Publishing, 2018, p. 812–819.

[7] S. Boumaraf, X. Liu, Z. Zheng, X. Ma and C. Ferkous, "A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images," *Biomedical Signal Processing and Control,* vol. 63, p. 102192, 2021.

[8] S. Khan, N. Islam, Z. Jan, I. U. Din and J. J. P. C. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognition Letters,* vol. 125, pp. 1-6, 2019.

[9] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe and S. Mougiakakou, "Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network," *IEEE Transactions on Medical Imaging,* vol. 35, no. 5, pp. 1207-1216, 2016.

[10] S. Hussain, S. M. Anwar and M. Majid, "Segmentation of glioma tumors in brain using deep convolutional neural network," *Neurocomputing,* vol. 282, pp. 248-261, 2018.

[11] J. Ma, F. Wu, J. Zhu, D. Xu and D. Kong, "A pre-trained convolutional neural network based method for thyroid nodule diagnosis," *Ultrasonics,* vol. 73, no. 0041-624X, pp. 221-230, 2017.

[12] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," *Procedia Computer Science,* vol. 90, no. 1877-0509, pp. 200-205, 2016.

[13] W. Sun, T.-L. (. Tseng, J. Zhang and W. Qian, "Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data," *Computerized Medical Imaging and Graphics,* vol. 57, pp. 4-9, 2017.

[14] H. Gao, B. Cheng, J. Wang, K. Li, J. Zhao and D. Li, "Object Classification Using CNN-Based Fusion of Vision and LIDAR in Autonomous Vehicle Environment," *IEEE Transactions on Industrial Informatics,* vol. 14, no. 9, pp. 4224-4231, 2018.

[15] S. Yang, W. Wang, C. Liu and W. Deng, "Scene Understanding in Deep Learning-Based End-to-End Controllers for Autonomous Vehicles," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 49, no. 1, pp. 53-63, 2019.

[16] C. J. Burke, P. D. Aleo, Y.-C. Chen, X. Liu, J. R. Peterson, G. H. Sembroski and J. Y.-Y. Lin, "Deblending and classifying astronomical sources with Mask R-CNN deep learning," *Monthly Notices of the Royal Astronomical Society,* vol. 490, no. 3, pp. 3952-3965, 2019.

[17] R. Yamashita, M. Nishio, R. K. G. Do and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging,* vol. 9, no. 4, pp. 611-629, 2018.

[18] X. Chen, K. Zhang, N. Abdoli, P. W. Gilley, X. Wang, H. Liu, B. Zheng and Y. Qiu, "Transformers Improve Breast Cancer Diagnosis from Unregistered Multi-View Mammograms," *Diagnostics,* vol. 12, no. 7, p. 1549, 2022.

[19] Y. Su, Q. Liu, W. Xie and P. Hu, "YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms," *Computer Methods and Programs in Biomedicine,* vol. 221, p. 106903, 2022.

[20] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang and T. Dacheng, "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 1, pp. 87-110, 2023.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-End Object Detection with Transformers," 2020.

[23] H. Wang, Y. Zhu, H. Adam, A. Yuille and L.-C. Chen, "MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers," 2021.

[24] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen and I. Sutskever, "Zero-Shot Text-to-Image Generation," 2021.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," 2021.

[26] M. A. Azam, K. B. Khan, S. Salahuddin, E. Rehman, S. A. Khan, M. A. Khan, S. Kadry and A. H. Gandomi, "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics," *Computers in Biology and Medicine,* vol. 144, p. 105253, 2022.

[27] N. Wu, J. Phang, J. Park, Y. Shen, S. G. Kim, L. Heacock, L. Moy, K. Cho and K. J. Geras, "The NYU Breast Cancer Screening Dataset v1.0," 16 Sep 2019. [Online]. Available: https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf.

[28] Radiological Society of North America, "RSNA Screening Mammography Breast Cancer Detection," [Online]. Available: https://www.kaggle.com/competitions/rsna-breast-cancer-detection/data.

[29] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. Cho and K. J. Geras, "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Medical Image Analysis,* vol. 68, p. 101908, 2021.

[30] B. Yan, Y. Jiang, J. Wu, D. Wang, P. Luo, Z. Yuan and H. Lu, "Universal Instance Perception as Object Discovery and Retrieval," 2023.

[31] American Cancer Society, "Getting Called Back After a Mammogram," [Online]. Available: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/getting-called-back-after-a-mammogram.html. [Accessed 15 04 2023].

[32] American Cancer Society, "Limitations of Mammograms," [Online]. Available: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html. [Accessed 15 04 2023].