Concordance in Breast Cancer Screening between Traditional and Deep Learning Models

Swati Rajwal, Siddhartha Mantrala, Aditya Prakash
Team SSAnet, Computer Science Department
Emory University

Abstract

Breast cancer is a critical health concern, and its early detection greatly influences the success of treatment and patient survival. Moreover, expertise in interpreting mammograms might not be available in all healthcare settings. To tackle this, our work introduces a novel framework that leverages the clinical and imaging data from the EMory BrEast Imaging Dataset (EMBED) to automate the classification of BIRAD scores. Our approach employs machine learning models for clinical data analysis and deep learning, specifically ResNet50 for feature extraction and transformers, for mammogram classification. Through experimental results, we show our empirical results and also discuss them in detail. Finally, the Cohen Kappa score is also evaluated to show the concordance between machine learning and deep learning models to classify the BIRAD score. We conclude the report by highlighting some of our lessons learnt and future steps in working on this dataset.

1 Introduction & Background

Mammography (X-ray of the breast) plays a crucial role in breast cancer detection, as it can often identify tumors that might be too small for a person or a physician to feel. However, interpreting mammograms requires extensive training and expertise which might not be readily available in all healthcare settings. Since mammography result in imaging data, we can take advantage of computer-aided diagnosis through machine learning (ML) techniques, particularly deep learning models which can potentially aid in the identification of patterns and abnormalities in the breast. Our work focuses on using the EMory BrEast Imaging Dataset (EMBED) [1] to enhance the classification of breast cancer based on the Breast Imaging Reporting and Data System (BIRADS) score.

1.1 Motivation

There have been studies in the past that signify the importance of automating mammogram interpretation [2]. Inspired by the need for a computer-aided system to interpret mammograms, our goal in this project is to design a BIRAD scores classification framework for breast cancer screening using mammograms and clinical data. We believe that such a framework will help in overcoming the lack of available expertise in interpreting mammograms by utilizing deep learning models. Also, such a model can assist healthcare professionals in analyzing more mammograms and identifying potential malignancies in a shorter period of time.

1.2 Dataset

The EMBED dataset consists of clinical parameters (non-imaging dataset) and mammograms (imaging dataset) from approximately 100k anonymized patients. Collected between 2013 to 2020, the size of this dataset is around 2.4 TB. The non-imaging dataset includes features such as age, gender, race, exam date, marital status, breast tissue density, region of interest (ROI) in the breast, Manufacturer, ManufacturerModelName, zip code, pathology severity (the most severe pathology result from a given specimen) and many others. On the other hand, the imaging dataset is much complex and contains multiple mammogram images for each of the multiple examinations for each patient. Multiple examinations for each patient are due to screenings occurring at different stages of time. In short, for a combination of clinical parameters and mammograms, the dataset contains one BIRAD assessment score (called 'asses'). Table 1 shows the BIRAD class label distribution in the dataset. A score 0 means abnormal activity detection and needs more examination to be certain. Score 1 refers to a normal mammogram (no signs of cancer). A score of 2 signifies a non-cancerous tumor (Benign). Finally, score 3 refers to malignancy with higher scores indicating various stages of breast cancer.

Table 1: Class labels distribution in our dataset

BIRAD Score Label	**Data Points**
Abnormal (0)	9169
Negative (1)	45278
Benign (2)	11970
Malignant (3)	6899

1.3 Related Work

Since the dataset has been recently released for academic research[1], there have been only a handful of published works using the EMBED dataset. For instance, we found three relevant studies ([8], [9] and [12]) that have used this dataset for binary classification. Their focus was solely on malignant or benign lesions. Specifically, the study conducted by the authors of [8] attempted to augment EMBED's imaging dataset with the DDSM [2] dataset, but the results reported were not as good as the ones obtained on just EMBED. Specifically, the study conducted by the authors of [8] attempted to augment EMBED's imaging dataset with the DDSM dataset, but did not perform multi-class classification.

On the other hand, the study by the authors of [9] only used image metadata to leverage clinical information for each mammogram. The authors of another paper [12] performed patch classification rather than considering the entire mammogram image. In contrast to the existing literature, our work is the first to perform multi-class BIRAD classification on the EMBED dataset using deep learning, and the first to incorporate clinical data into the pipeline. This allows for a more nuanced and accurate assessment of breast cancer risk compared to previous studies. Moreover, during our literature review, we found more studies ([11], [14]) that cited the EMBED dataset. However, they only do so in the context of fair and diverse datasets and unbiased Artificial Intelligence. These papers do not perform any classification but were useful to get a holistic idea about the potential applications of this dataset.

[1] https://aws.amazon.com/marketplace/pp/prodview-unw4li5rkivs2
[2] Digital Database for Screening Mammography

2 Methodology

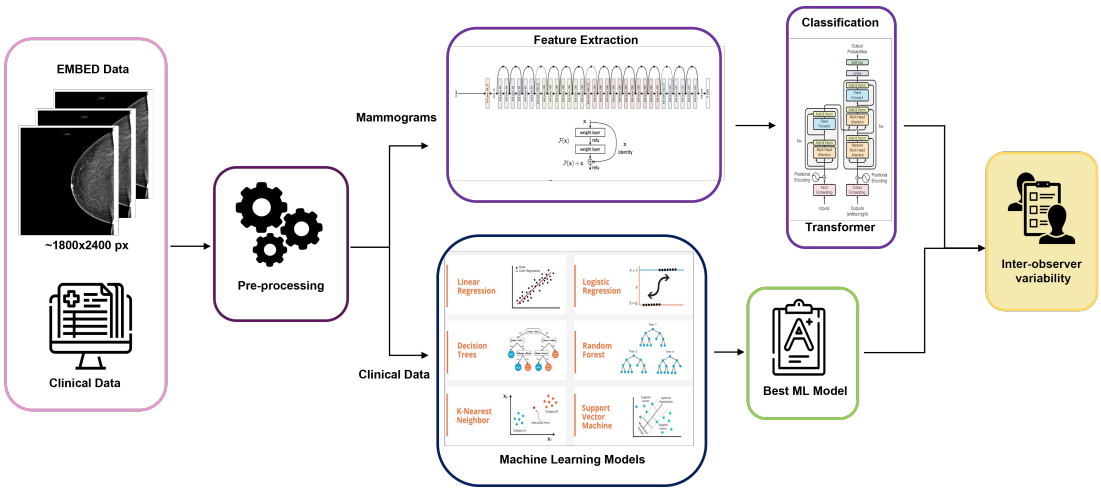


Figure 1: Our Proposed framework

Our proposed pipeline for this project is shown in Figure 1. The pipeline consists of two approaches:

1. Machine learning models trained on clinical data
2. Deep Learning model trained on Imaging dataset

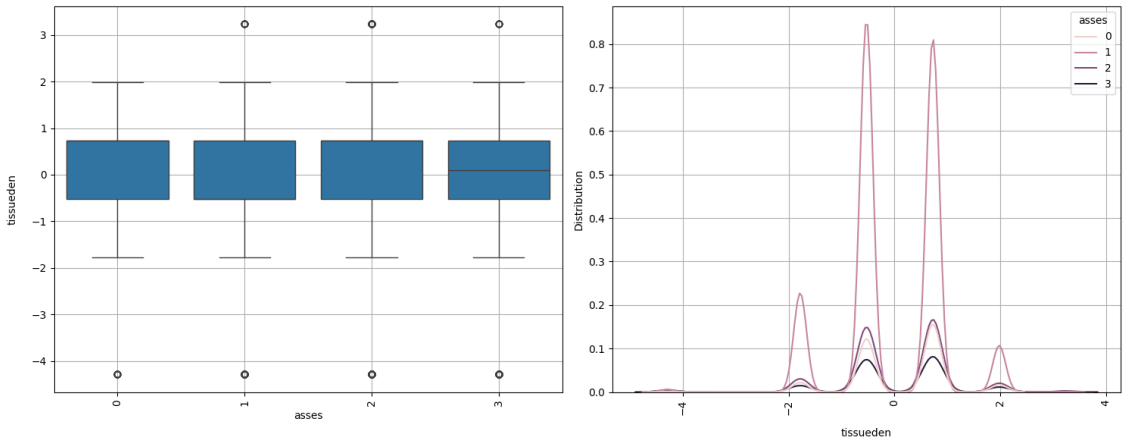
For the clinical data, we train various Machine Learning models and select the best-performing model based on the test set. For the imaging dataset, we train a CNN (ResNet50 in our case) to extract features from the images and a transformer to classify the images.

2.1 Clinical Dataset & Preprocessing

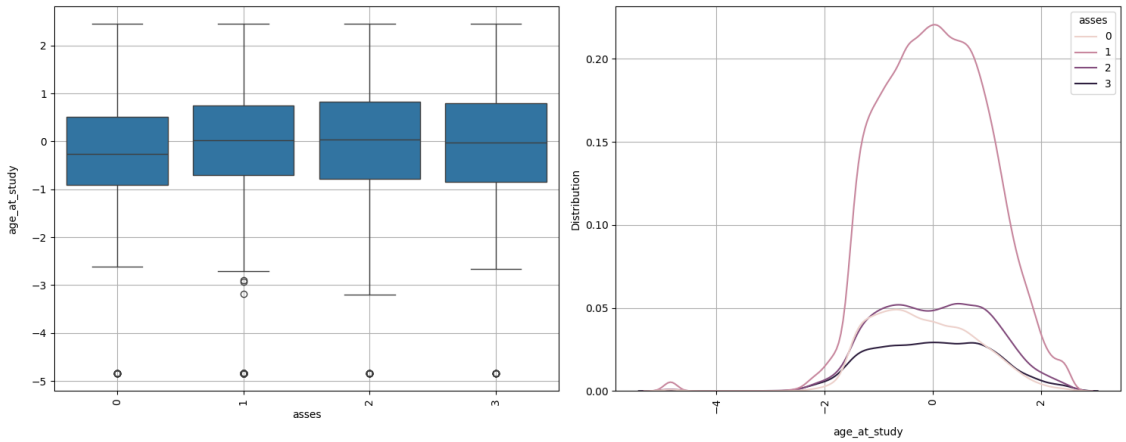
A major challenge we faced was merging information from the clinical and metadata files. The clinical data is indexed by individual findings which can correspond to either breast and also varies by examination. The metadata is indexed by a file wherein each row corresponds to one image. Both of these files can be linked by patient ID ('empi_anon'), exam ID ('acc_anon'), and laterality of the clinical finding and image. We matched the clinical findings of the left breast with files of the left breast, the clinical findings of the right breast with files of the right breast, and the clinical findings of both breasts to all files. We filtered the clinical data to include only those records with a single finding ('numfind'=1). This way, we made sure that each clinical data entry was appropriately matched with its corresponding imaging data.

Table 2: Data size for each set

Dataset	**# Images/Clinical Data**
Train	52782
Validation	13196
Test	7331



(a) Breast density, 'tissueden'



(b) Age at Study

Figure 2: Clinical features (a) Tissue density. (b) Age at study visualization with target.

The dataset was then divided into Train, Validation, and Test sets in the ratio of 70:20:10 as shown in Table 2. Next, we eliminated irrelevant columns, imputed missing values with predefined defaults, and applied transformations to certain columns. Specifically, we encoded categorical variables (including the target variable called 'asses') using LabelEncoder by Sklearn [6] for converting non-numeric categorical data into a numeric categorical data format. Additionally, we created new features by counting the occurrences of unique elements within columns such as the number of dicom images and number of ROIs. Before applying a model, the standard scaling was applied to the feature sets (two features shown in Figure 2). The correlation between the features and target variable is also shown in heatmap as shown in Figure 3 where we put a threshold of 0.9 magnitude or above for feature elimination.

2.2 Imaging Dataset & Preprocessing

For the imaging dataset, we had mammograms which are DICOM (Digital Imaging and Communications in Medicine) images. The DICOM images were first converted to PNG using Python libraries such as pydicom[3]. Then, in order to optimize the dataset for deep learning models, a preprocessing step was applied to standardize the dimensions of the mammogram images from ∼2000x1800 to 224×224 pixels. Following resizing, we applied

[3] https://pypi.org/project/pydicom/

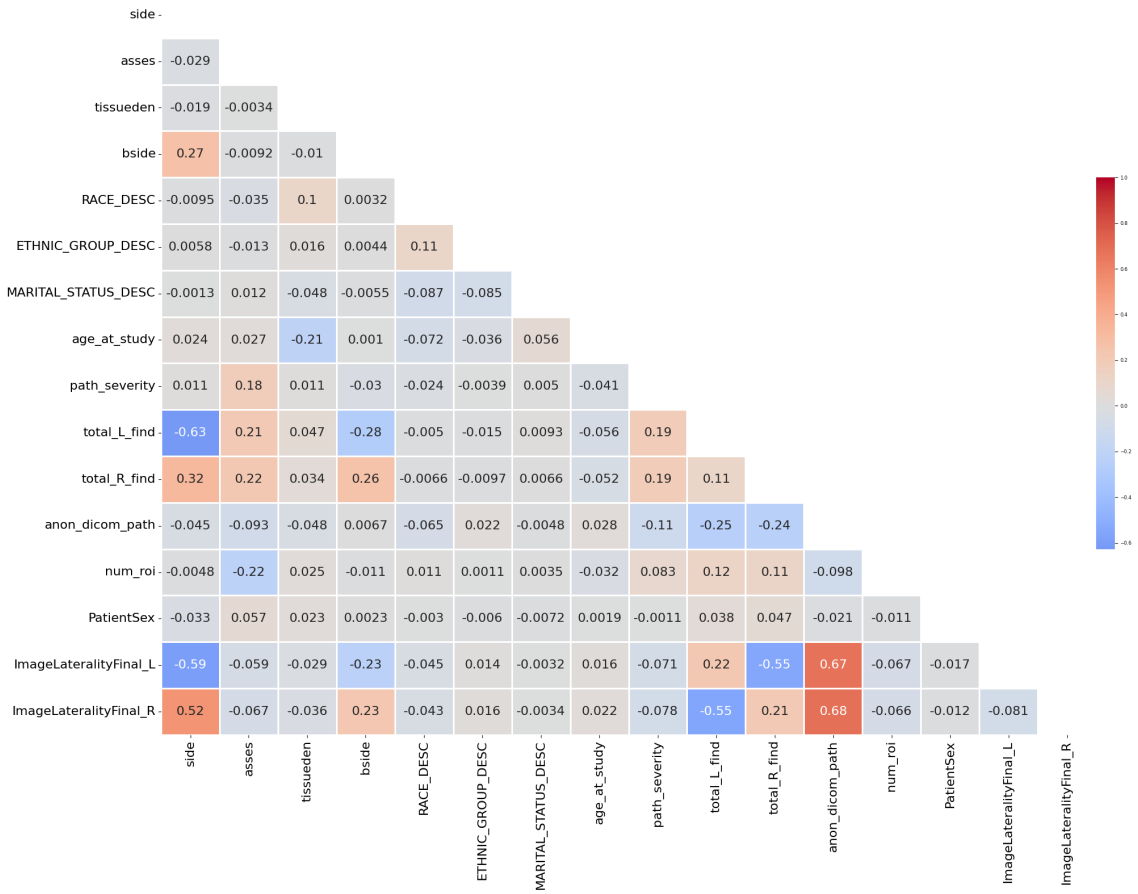


Figure 3: Features Correlation Heatmap after preprocessing (Threshold 0.9)

CLAHE (Contrast Limited Adaptive Histogram Equalization) [7] to enhance the contrast of the images, which is particularly important for medical images like mammograms. It is useful in enhancing the visibility of important features such as mass patterns and has been used in existing research work on mammogram images [10]. Figure 4 shows one such mammogram where the contrast levels are flat and using CLAHE we enhance the contrast in the image. We apply this to all the images, CLAHE being adaptive, does not affect the contrast levels of images with good dynamic range. Finally, as the images are in grayscale, we merge them along the channel dimension to create three-channel inputs. This step was necessary to make the images compatible with the ImageNet pre-trained ResNet50 model.

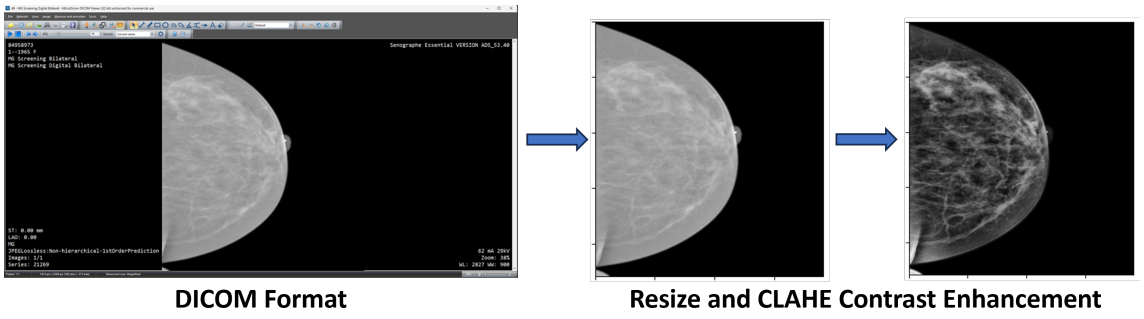


Figure 4: Conversion of the original DICOM image[4]to PNG with enhancement

[4] https://www.microdicom.com/

3 Experimental Results

3.1 Machine Learning Modeling Choices

For modeling the clinical dataset, we utilized Naïve Bayes as a baseline model due to its simplicity and quick training time. We also used decision trees, logistic regression, and random forest since they can provide insights into the decision-making process (which is valuable in clinical settings). Lastly, we used ensemble methods like Random Forest or Gradient Boosting to leverage the strengths of individual models to improve overall performance. For hyperparameter optimization, we used a parameter search space and GridSearchCV on the pre-processed dataset as described in detail in Section 2.1. Specifically, we used 5-fold cross-validation for evaluating each candidate model's performance over the param search space. The optimal parameters finally selected for further training are shown in Table 3. After optimal parameter selection, the models were trained across the training set and evaluated across various metrics on the test set. The results for all the models are shown in Section 3.3.

Table 3: Hyperparameters of Machine Learning Models

Model	**Hyperparameters**
Naive Bayes	`alpha=0.01`
Decision Tree	`criterion='gini', max_depth=10`
Logistic Regression	`C=10, penalty='l2', max_iter=3000`
Random Forest	`max_depth=None, n_estimators=50`
KNN	`n_neighbors=7`
SVM	`probability=True, C=1, kernel='rbf'`
Gradient Boosting	`n_estimators=100, learning_rate=0.1`

3.2 Deep Learning Modeling Choices

For the image dataset, we use deep learning models to perform the BIRAD classification task. The deep learning approach comprises two steps: Feature extraction followed by classification. The EMBED dataset consists of multiple patients with each patient having multiple examinations over a period of time. Each examination comprises multiple mammograms, which are images of breast X-rays captured from different views and positions. Labels are provided at the examination level, but not for each image in the examinations. The dataset does not indicate which image was used for deciding the BIRADS score. To tackle this issue, we randomly select a single image from each examination and its corresponding label and train a ResNet50 as a feature extractor for the task of classifying those images. For this task, we train ImageNet pre-trained ResNet50 for 25 epochs, with Adam Optimizer at a learning rate of 0.001, with $\beta 1 = 0.9$ and $\beta 2 = 0.999$. After training the ResNet50 for feature extraction, we freeze this model and use it as a feature extractor for training the transformer model network proposed by [15]. For the classification task, we train the transformer on the task of classifying each examination that a patient goes through. As each examination consists of a varying number of mammograms, we use the transformer here as it can take input sizes of varying lengths to perform classification tasks. Given an examination, we use ResNet50 to extract features from the mammograms and pass this set of features of dimension (no of images in the examination, 2048), to train the

transformer network. With the advantage of the attention mechanism of the transformer, we use the information from all the images to classify the BIRADS score for the examination. We trained a 4 multihead-attention transformer for 25 epochs (results in Section 3.3), with Adam Optimizer at a learning rate of 0.001, with $\beta 1 = 0.9$ and $\beta 2 = 0.999$. This way we have an end-to-end deep learning framework for classifying each examination to a BIRADS score.

3.3 Emperical Results & Comparison

We are using a combination of accuracy, ROC AUC, precision, recall, and F1/F2 scores (both micro and macro) to provide a comprehensive understanding of each model's performance. Using this set of metrics allowed us to assess not only the overall accuracy of each model but also how well it performs in terms of false positives and negatives, its ability to handle class imbalances, and its capacity to correctly identify the BIRAD score label. Table 4 reports the experimental results across various metrics for each candidate model.

Table 4: Model Performances across various Evaluation Metrics

Model	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1 Micro**	**F1 Macro**	**F2 Micro**	**F2 Macro**
Naïve Bayes	0.660	0.886	0.722	0.349	0.660	0.356	0.660	0.343
Decision Tree	0.768	0.903	0.684	0.611	0.768	0.606	0.768	**0.599**
Logistic Regression	0.764	0.910	0.728	0.589	0.764	0.593	0.764	0.580
Random Forest	0.750	0.907	0.581	0.586	0.750	0.583	0.750	0.585
KNN	0.735	0.878	0.564	0.570	0.735	0.565	0.735	0.568
SVM	**0.773**	**0.914**	**0.770**	**0.622**	**0.773**	**0.607**	**0.773**	**0.598**
Gradient Boosting	0.764	**0.918**	0.680	0.597	0.764	**0.611**	0.764	0.598
ResNet50[5]	0.653	0.8723	0.457	0.331	0.653	0.330	0.653	0.323
Transfromer	0.615	0.554	0.154	0.25	0.615	0.19	0.615	0.222

We observe that SVM performs better across multiple metrics and the ROC AUC curve is shown in Figure 5. It also scores well in F1 Micro (0.773), indicating good balance between precision and recall for individual classes.

Concordance between ML and Transformer: In order to see inter-observer variability, we calculated Cohen's kappa score between the gold standard and ML model which turns out to be 0.47858. Cohen's kappa score between the gold standard and transformer model is 0, and Cohen's kappa score between the ML model and transformer is also 0. This is was happening as the transformer was assigning all the data points as class 3.

[5]ResNet50 used only for feature extraction purpose.

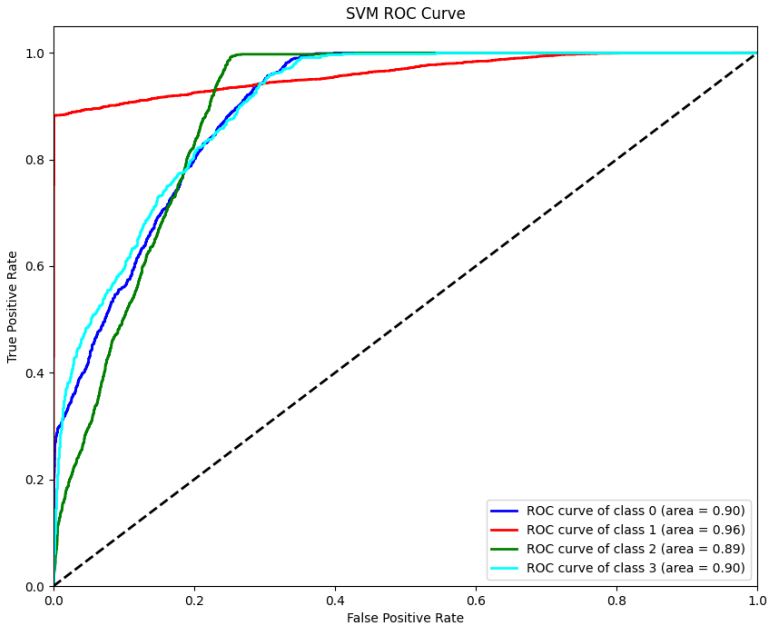


Figure 5: Trained SVM Performance on test set

4 Discussion

In this project, we performed multi-class BIRADS score classification on the EMBED dataset using both the clinical as well as imaging dataset. The novelty of our work is that we are the first one to incorporate clinical data into the pipeline with mammograms. This is crucial since the BIRAD scoring involves not only the identification of features in mammograms but also the interpretation of these features in a clinical context. On top of that, we implemented the idea of using the transformer model to mimic a radiologist in assigning the BIRADS score to the patients. This allowed for a more nuanced and accurate assessment of breast cancer risk compared to previous studies. Through our experimental results, we observe that SVM outperformed the baseline as well as the other models across all the evaluation metrics (AUC 0.914) for the clinical dataset. However, we found that there is still some room for improvement, especially for the class 2, which has the lowest AUC score (Figure 5).

For processing the mammograms, we made a random selection of mammograms per examination. This decision was made after a thorough discussion within the team. Using only these last images in each examination could have introduced a bias in our analysis if the last images were systematically different from earlier ones (for instance, they might focus on areas of concern identified in earlier images). That is why we decided to use random selection to ensure a more representative sample of the entire examination since our goal was to understand general patterns across a wide range of mammogram examinations.

Moreover, it is important to highlight that time-series analysis was not included in the current study but is identified as an important aspect for future work, as it could provide additional insights into the progression of findings across sequential mammograms. Currently, the transformers are utilized in our case because each examination has a different number of images. As a radiologist uses multiple images to diagnose each examination and assign the BIRADS score, we use the attention mechanism of the transformer to make a decision on varying numbers of mammogram images for the classification of the BIRAD

score. Another reason why we decided to utilize transformers is the architecture's ability to handle sequential and context-dependent data which is present in multiple images of a single examination. This is something which cannot be combined by a CNN to assign the BIRAD score.

Also, we found that both modalities (imaging and non-imaging) were useful. We focused on both modalities however, the computation times varied a lot across both modalities. For instance, the imaging dataset training of one epoch took anywhere between 3-6 hours depending on the sample set. However, the maximum it took for modeling clinical datasets on the standard machine learning was 2-3 hours for hyperparameter tuning.

Some of the other lessons learned during this project were handling large datasets (2.4 TB) and the significance of organizing, cleaning, preprocessing, and many challenges faced during these stages. We also had to set up the H100 server which was a critical step to install all the necessary packages and libraries to run computationally demanding jobs such as model training.

Future Work

This framework can used as an inspiration for developing similar frameworks that perform BIRAD classification using different modalities. Exhaustive hyper-parameter search and tuning is something that could be a starting point when using this framework. Different off-the-shelf CNNs like Inceptionv2, EfficientNet, MobileNet,. etc, can be used for feature extraction, likewise for the transformer network. More sophisticated approaches can be used to tackle the non-uniform distribution of the classes in the dataset.

Code and Dataset Availability

Python version 3 was used as the programming language and code related to this project can be located in this GitHub Repository: https://github.com/cosinesimilarity1/Team-SSAnet. Please note that image processing such as preprocessing, training, testing, and evaluation was done on the H100 server at Emory University which took more than 6 hours for just one epoch. For other models (standard machine learning models), Google Colab was used.

Acknowledgements

We would like to thank Prof. Joyce Ho for teaching the coursework CS 534. This work has been inspired by the concepts learned during the lectures. We also extend our gratitude towards Sergio Gramacho and Edgar Leon for providing H100-related technical support. We are also grateful to the Judges and fellow classmates for providing helpful feedback on our madness talk.

References

[1] Jeong, Jiwoong J., et al. "The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images." Radiology: Artificial Intelligence 5.1 (2023): e220047.

[2] Pesapane, F., Trentin, C., Ferrari, F., Signorelli, G., Tantrige, P., Montesano, M., .& Cassano, E. (2023). Deep learning performance for detection and classification of microcalcifications on mammography. European Radiology Experimental, 7(1), 69.

[3] Wanders, Alexander JT, et al. "Interval cancer detection using a neural network and breast density in women with negative screening mammograms." Radiology 303.2 (2022): 269-275.

[4] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[5] C.Szegedy et al. "Rethinking the inception architecture for computer vision". arXiv preprint arXiv:1512.00567, 2015.

[6] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[7] Reza, A. M. (2004). Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. Journal of VLSI signal processing systems for signal, image and video technology, 38, 35-44.

[8] Hwang, I., et al. (2023). Impact of multi-source data augmentation on performance of convolutional neural networks for abnormality classification in mammography. Frontiers in Radiology. https://doi.org/10.3389/fradi.2023.1181190

[9] V. Nalla et al., "Influence of Convolutional Neural Network Depth on the Efficacy of Automated Breast Cancer Screening Systems," 2023 IEEE Symposium on Computers and Communications (ISCC), Gammarth, Tunisia, 2023, pp. 1-4, doi: https://doi.org/10.1109/ISCC58397.2023.10217947

[10] J. Dabass, S. Arora, R. Vig and M. Hanmandlu, "Mammogram Image Enhancement Using Entropy and CLAHE Based Intuitionistic Fuzzy Method," 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2019, pp. 24-29, doi: 10.1109/SPIN.2019.8711696.

[11] Chen, R. J., Chen, T., Lipková, J., Wang, J. J., Williamson, D. F. K., Lu, M., Sahai, S., & Mahmood, F. (2021). Algorithm fairness in AI for medicine and healthcare. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2110.00603

[12] Zhang, L., Brown-Mulry, B., Nalla, V., Hwang, I., Gichoya, J. W., Gastounioti, A., Banerjee, I., Seyyed-Kalantari, L., Woo, M., & Trivedi, H. (2023). Multivariate analysis on performance gaps of artificial intelligence models in screening mammography. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.04422

[13] Rajpal, S., Lakhyani, N., Singh, A. K., Kohli, R., & Kumar, N. (2021). Using handpicked features in conjunction with ResNet-50 for improved detection of COVID-19 from chest X-ray images. Chaos, Solitons & Fractals, 145, 110749.

[14] Jeong, J., Vey, B. L., Bhimireddy, A. R., Kim, T. J., Santos, T., Correa, R., Dutt, R., Mošunjac, M., Oprea, G., Smith, G. H., Woo, M., McAdams, C. R., Newell, M. S., Banerjee, I., Gichoya, J. W., & Trivedi, H. (2023). The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.4 Million Screening and Diagnostic Mammographic Images. Radiology, 5(1). https://doi.org/10.1148/ryai.220047

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.