Causal inference for predicting treatment outcome in breast cancer

DePaul University, Chicago, IL
Jarvis School of Computing and Digital Media

Advisor
Ilyas Ustun, PhD: iustun@depaul.edu

Team Members
Arun Sivakumar: asivakum@depaul.edu

Anisha Patel: apate228@depaul.edu

Ananth Chalumuru: achalumu@depaul.edu

Introduction

Neoadjuvant therapy (NAT) is an early treatment option for breast cancer, which aims to reduce tumor burden and enable patients to undergo breast-conserving surgery instead of mastectomy. While some patients achieve complete recovery, others experience incomplete recovery. Identifying a method to predict which breast cancer patients will achieve pathological complete response (pCR) and which will not after NAT would improve patient stratification and lead to more appropriate treatment regimens. Several studies[10] have investigated using imaging features for predicting neoadjuvant response but these studies do not use images itself to establish causal inference. In our paper, we identify causal relationships between images and pathology labels using a Novel refined CNN approach to learn causation from the raw images themselves.

Dataset:

The study included a group of patients, and prior to treatment, each patient's dynamic contrast-enhanced MR (DCE-MR) images were obtained. There were 64 patients with complete response to neoadjuvant therapy and 224 with no response to neoadjuvant therapy.

Methodology: In order to classify the outcome of neo adjuvant therapy we implemented 4 methods:

A. Classifying from MRI images:

a. Implemented 3D CNN on volumetric MRI images of each patient:

The original DICOM images were used to train the 3D CNN model. A pretrained ResNet10 from MONAI was used with MedicalNet weights and these weights come from training on 23 different medical datasets. The authors of [1] claim that pretraining on medical datasets makes the model converge faster and so we utilize these weights for our experiments.

b. Implemented 2D CNN on the middle slice of annotated images for each patient:

Using the annotation information, the MRI images were cropped in x, and y and z dimensions. The middle slice of the annotations was selected to represent each patient in training. We carry out data augmentation on the images that are separated out for training to avoid overfitting. As we are learning 2D image classification, a pre-trained ResNet50 model with Imagenet weights was loaded for classification. Usage of pre-trained models[2,3,4] for breast cancer classification was explored by the authors of [5] where they reported that ResNet50 model with pre-trained weights gave them the best classification performance. The authors of [5] also explored various hyperparameters and reported hyperparameters that gave them higher classification performance. We use their hyperparameters for training our deep learning model and keep the hyperparameters common for 2D CNN and refined 2D CNN.We discovered that using a 2D CNN model we can learn the pixels to predict if a patient can recover after Neoadjuvant therapy but predict if a patient will not recover after neoadjuvant therapy. This is a limitation of 2D CNN and we addressed this using a Novel Refined CNN algorithm that estimates ground truth and learns these pixels to predict if a patient will not recover after neoadjuvant therapy.We will discuss and provide details on our inspiration for this method in the next section.

c. Implemented refined 2D CNN on the middle slice of each patient using GradCAM(proposed method):

Deep matrix factorization. Deep matrix factorization is the method by which we can quantify a model’s reasoning. This method first introduced by [6] involves carrying out Principal Component Analysis on deep features extracted from the final convolution layer of a model. We can break down the components,and show the percentage of the labels present in each component. Here we refer to training classes as labels and in figure 2 we can see that the model predicted that the picture had a bear using the pixels represented in green. We hypothesize that by using the same method on medical images we can estimate ground truth in the image (ie) we can find the pixels that show us that the patient will not recover after neoadjuvant therapy. We adapt the gradCAM implementation from [9]. For detailed explanation on GradCAM refer to Appendix B

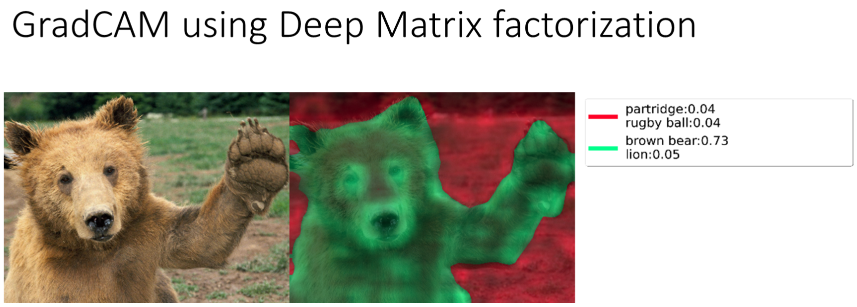


Figure 2 showing that the model predicted the image as bear with 73% of the actual label present in component 2.

Refined CNN for medical images. Human in the loop ideologies have improved the deep learning models by introducing domain knowledge into the model. In [8], they perform annotation refinement using the ground truth regions provided by physians to make the deep learning model learn certain regions. Learning these regions makes the model predict cancer using the correct regions.

In our Novel method, We train two models, one model which is trained using all the images and we predict on the same images using this initial model. For the instances that are correctly classified, we train a new second model on these images and for the misclassified instances,we carry out our refinement routine. Our refinement routine involves, two steps

- In step1, we estimate ground truth regions using deep matrix factorization. Deep matrix factorization involves carrying out PCA on the extracted CNN features and we set K =2, as we are carrying out binary classification. From the components received from PCA, we can break down the components based on the percentage of label present in each component and overlay the matrix on top of the image to see which pixels are present in each component.
- Now,in step 2 we give higher weights to the components which have a higher percentage of the correct label and consider these pixels as estimated ground truth. We train the second model on these instances, with the weight mask on top of the image, so that higher weights are given to the loss that comes from the estimated ground truth regions in the cross entropy loss function.

In figure 3, we illustrate our method and show how giving higher weights to the correct regions makes our novel refined CNN model learn these estimated ground truth regions better than baseline CNN.

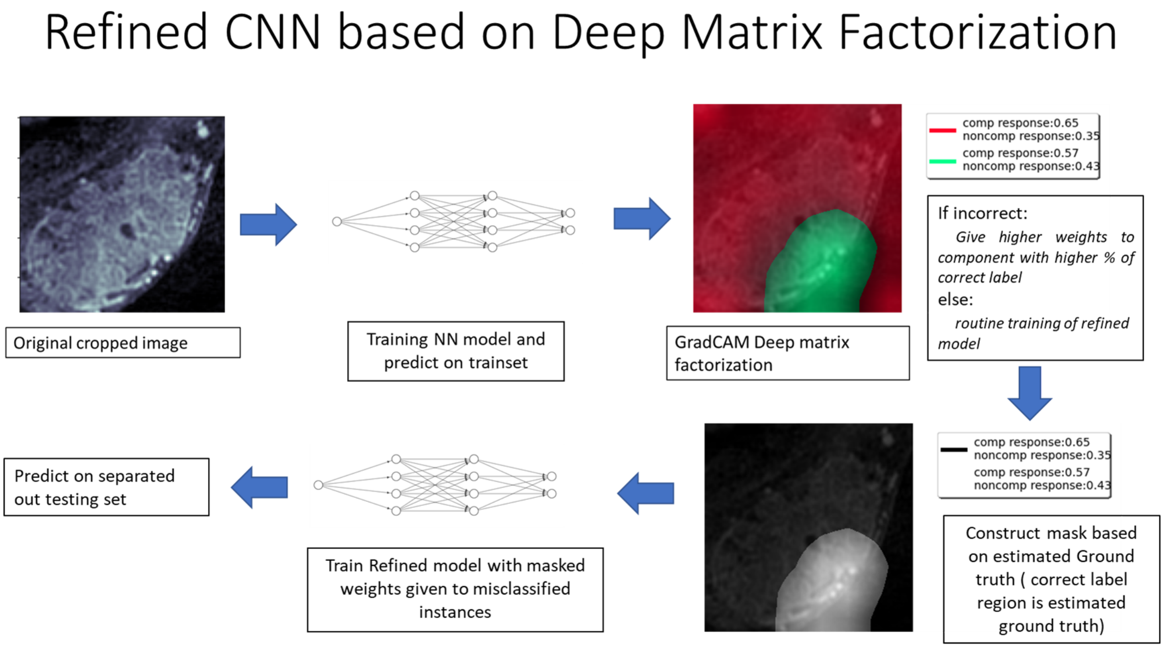

Refined CNN based on Deep Matrix Factorization
Original cropped image
Training NN model and predict on trainset
GradCAM Deep matrix factorization
🟥 comp response:0.65 noncomp response:0.35
🟩 comp response:0.57 noncomp response:0.43
If incorrect:
Give higher weights to component with higher % of correct label
else:
routine training of refined model
comp response:0.65 noncomp response:0.35
comp response:0.57 noncomp response:0.43
Construct mask based on estimated Ground truth (correct label region is estimated ground truth)
Train Refined model with masked weights given to misclassified instances
Predict on separated out testing set

Figure3 shows an illustration of how a classifier trained using the Refined CNN learns misclassified examples better. In training phase 1, we train a CNN model and predict on the training set and factorize CNN features of misclassified examples. The example mentioned in the diagram was misclassified as a non-complete response. Region/ component with higher percentage of correct label (red region) is considered as estimated ground truth. In the next phase, we give higher weights to estimated ground truth regions and learn the pixels in those regions in the form of higher weights in the loss function with a newly initialized CNN model.

B. Classifying from image features.

There were 529 computer-extracted imaging features representing a variety of imaging characteristics including size, shape, texture, and enhancement of both the tumor and the surrounding tissue. The missing values were replaced by mean. The problem with this approach is the curse of dimensionality. There are 529 features and only 288 instances. In order to overcome this issue, we chose to use Principal Component Analysis. Principal Component Analysis (PCA) is a statistical technique that is used to reduce the dimensionality of large datasets. 3 components explained 99.6% of variability in the image features. The outcome of NeoAdjuvant therapy was classified using the SVM and Random Forest.

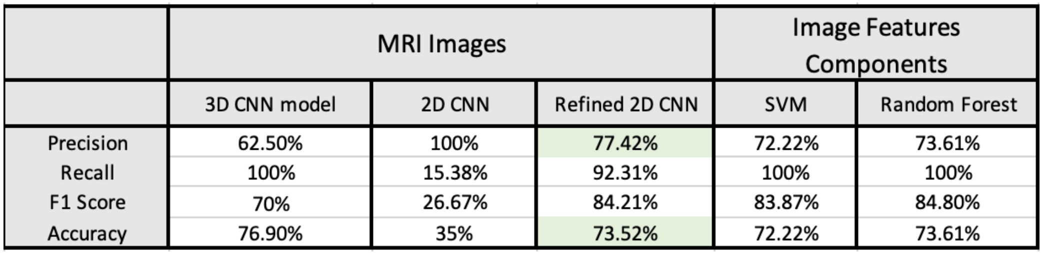
Results:

A: Classifying outcome from MRI images

The 2D ResNet50 CNN model gave the precision of 100%, however, Recall was only 15%, and Accuracy was 35% on the testing data. The refined 2D CNN gave precision and recall of 77.42% and 92.31% respectively.we save the model according to performance on a validation set composed of equal number of samples from each class.Finally, we test the model and report their performance on the original proportion of classes present in the testset. This is done to test how the model performs when the examples are provided as it occurs in the wild, with Out-Of-Distribution proportions in table 1.

B. Classifying outcome from image features.

When classifying the outcome of neoadjuvant therapy using 3 components of image features, Random Forest slightly outperformed SVM. The Random Forest had the following hyperparameters, number of estimators = 10, minimum samples split = 5, minimum samples leaf = 20, max depth =19, and bootstrap = True. For SVM the following were C = 5, kernel was radial with gamma of 1.



	MRI Images			Image Features Components	
	3D CNN model	2D CNN	Refined 2D CNN	SVM	Random Forest
Precision	62.50%	100%	77.42%	72.22%	73.61%
Recall	100%	15.38%	92.31%	100%	100%
F1 Score	70%	26.67%	84.21%	83.87%	84.80%
Accuracy	76.90%	35%	73.52%	72.22%	73.61%

Table 1 shows us that Refined 2D CNN has the highest precision across all models; Although SVM and Random Forest classifier had a recall of 100% they predicted that no patients will show response to neoadjuvant therapy implying that SVM and Random Forest are overfitting to the majority class and cannot predict the minority class well, more details attached in appendix A. Compared to 2D CNN's our method refined 2D CNN's has higher accuracy as we learn to differentiate both classes better.

Discussion:

It is important to determine which patients will benefit from NeoAdjuvant therapy. Our technique of refined 2D CNN shows that it is possible to classify which patients give a complete response to NAT. The dataset was quite small for this kind of problems, more instances will help in making the model even more robust.

Summary statement:

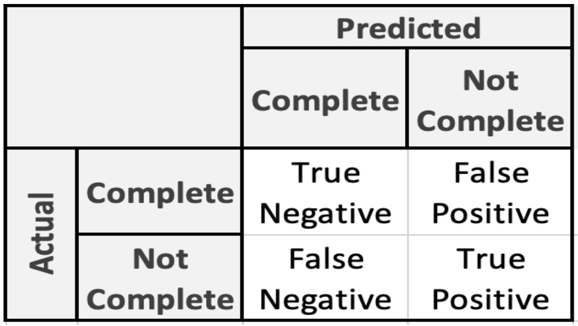
- Our work is the first to estimate ground truth labels from the images using Deep matrix factorization and so, we believe that novel methodologies deserve to be considered as finalists
- From the high classification accuracy and precision of our Refined 2D CNN model, we confirm that we identified causal features from the images and utilized these causal features to improve prediction performance of our method.

References:

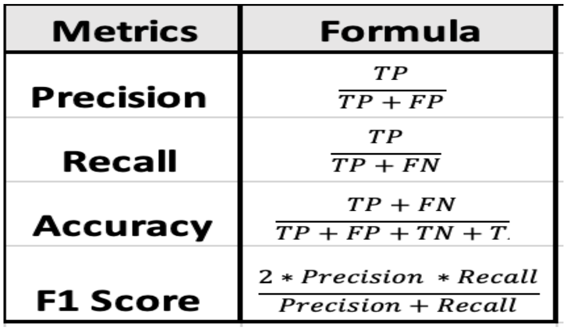
1. S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer Learning for 3D Medical Image Analysis," in IEEE Transactions on Medical Imaging, vol. 38, no. 6, pp. 1538-1548, June 2019, doi: 10.1109/TMI.2019.2901476.
2. S. A. Hossain et al., "Breast Cancer Diagnosis Using Deep Convolutional Neural Networks and Support Vector Machines," in IEEE Access, vol. 7, pp. 17976-17985, 2019, doi: 10.1109/ACCESS.2019.2891163.
3. N. F. F. Ibrahim et al., "Deep Learning for Breast Cancer Diagnosis Using Mammography Images: A Review," in IEEE Access, vol. 9, pp. 23728-23744, 2021, doi: 10.1109/ACCESS.2021.3050492.
4. M. A. Salah et al., "Breast Cancer Classification from Histopathological Images with Inception Convolutional Neural Networks," in IEEE Access, vol. 8, pp. 155191-155202, 2020, doi: 10.1109/ACCESS.2020.3017387.
5. Bae, S., Yoo, C., Choi, Y., Kim, S., & Lee, K. (2019). Multiclass classification of breast cancer histology images using convolutional neural networks. IEEE Access, 7, 174420-174431. doi: 10.1109/ACCESS.2019.2956152
6. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618-626). https://doi.org/10.1109/ICCV.2017.74
7. Patel, K., & Varshney, K. R. (2021). Deep matrix factorizations for ground truth estimation from CNN features. arXiv preprint arXiv:2105.10734.
8. S. Shakya, M. Vasquez, Y. Wang, R. Tchoua, J. Furst, and D. Raicu, "Human-in-the-loop deep learning retinal image classification with customized loss function," in IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, Oct. 2021, pp. 3901-3905, doi: 10.1109/ICIP42928.2021.9506785.
9. Gil, J. (2017). PyTorch implementation of Grad-CAM. GitHub repository. https://github.com/jacobgil/pytorch-grad-cam.
10. A. Bahl, E. Ojeda-Fournier, and S. S. McDonald, "Multivariate Machine Learning Models for Prediction of Pathologic Response to Neoadjuvant Therapy in Breast Cancer using MRI features: A Study Using an Independent Validation Set," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 1, pp. 297-305, Jan. 2019, doi: 10.1109/JBHI.2018.2832193.

Appendix A:

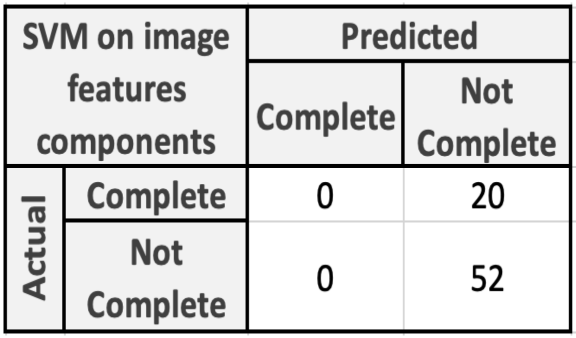
In this section, we show confusion matrices of binary classification of our methods.we report performance of 3D CNN model, with testing batch size of 10, since these models require large memory spaces for 3D convolutions.Testing set for the 3D CNN model comprises of 5 instances from each binary class. Note that we split the dataset into two independent splits for training and validation, we use the validation set for testing and do not separate out further as the dataset size is small.



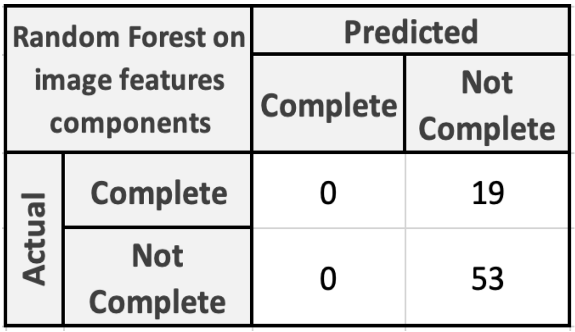
		Predicted	
		Complete	Not Complete
Actual	Complete	True Negative	False Positive
	Not Complete	False Negative	True Positive



Metrics	Formula
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
Accuracy	$\frac{TP + FN}{TP + FP + TN + T.}$
F1 Score	$\frac{2 * Precision \ * Recall}{Precision + Recall}$

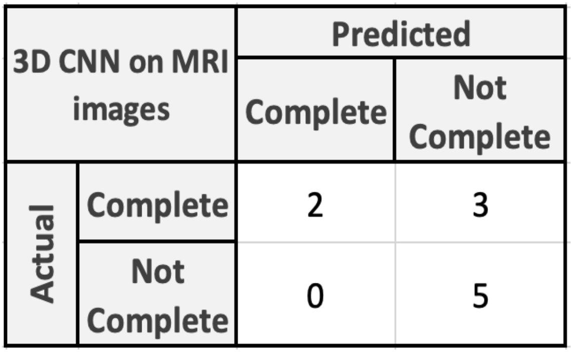


SVM on image features components		Predicted	
		Complete	Not Complete
Actual	Complete	0	20
	Not Complete	0	52

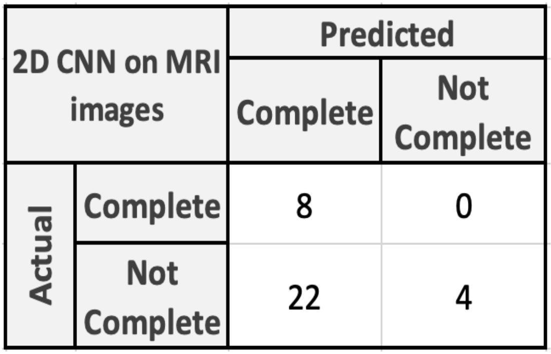


Random Forest on image features components		Predicted	
		Complete	Not Complete
Actual	Complete	0	19
	Not Complete	0	53

Table2 above shows us that SVM and Random forest that learns from Image features overfit to the dominant class and classify all instances as Not complete

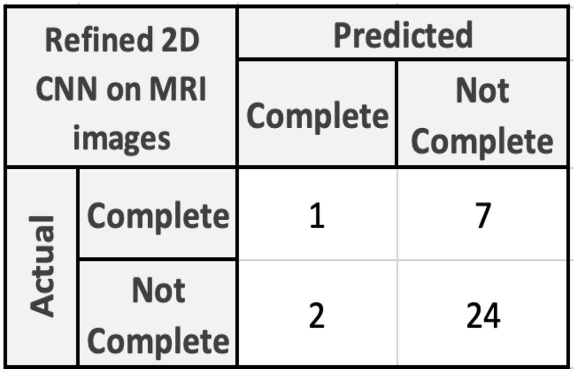


3D CNN on MRI images		Predicted	
		Complete	Not Complete
Actual	Complete	2	3
	Not Complete	0	5



2D CNN on MRI images		Predicted	
		Complete	Not Complete
Actual	Complete	8	0
	Not Complete	22	4

Table3 above shows us that 3D and 2D CNN, methods learn from raw Image themselves, are less prone to overfitting compared to SVM and Random Forest, particularly 2D CNN is able to predict Complete response better



Refined 2D CNN on MRI images		Predicted	
		Complete	Not Complete
Actual	Complete	1	7
	Not Complete	2	24

Table4 above shows us that our method Refined 2D CNN, learns from raw Image themselves and is able to predict Not complete response well, showing that it can predict patients having no response to Neoadjuvant therapy from patients having complete response.

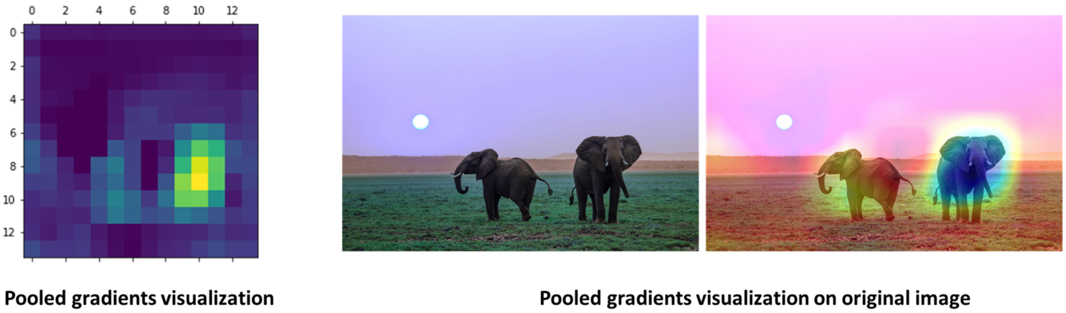
Appendix B

GradCAM.Usage of explainable AI techniques [7] could provide us information on which pixels influence the model to make a prediction. To show an easy example of where a model looks at to make a prediction, we carried out the following steps:

- Loaded a pre-trained model VGG19 trained on Imagenet
- Model was trained on an Imagenet dataset with elephants as the 386th class / 1000 classes by the authors.
- Model predicted that the Image we gave was an elephant and it belongs to class 386.
- We get the activation gradients (backpropagation - traceback to the original image) of the 386th class from the prediction and plot it as visualization.
- Gradients are influencers and we show which pixels influenced the model to give this prediction in Figure 1.

After carrying out the aforementioned steps, we can plot the Gradients on top of the image. We can see that the model looked at the body and the ears of the elephant to predict that the image contains an elephant.

GradCAM visualizations of the prediction class



Pooled gradients visualization

Pooled gradients visualization on original image

Figure1 shows that the pooled gradients plotted on top of the image shows that pixels of the elephant in the right played a vital role in the model predicting that it belongs to african elephant class. Using this method we can visualize and establish causal inference of where the model sees to make the prediction. We further use Deep Matrix factorization with GradCAM to quantify the model’s reasoning of its prediction which is explained below.