## Materials and Methods

### Data

#### Preparation

We retrospectively acquired 119 patients who’d undergone percutaneous ablation therapy (PTA) for metastatic colorectal liver metastasis (CRLM) at our institution. Patient selection criteria included: a minimum follow-up period of 6 months, and a maximum of 2 months between pre-treatment and post-treatment contrast enhanced CT imaging. Follow-up data included three possible outcomes: residual disease, local progression, and no progression. For this work, labels 1 and 2 were combined as progression.

CRLM were manually segmented on pre-treatment contrast-enhanced CT scans, and validated by a radiology trained physician fellow. Segmentations were labeled based on the outcome status of the individual site, ‘GTV\_Recurred’ for progression and ‘GTV\_Not\_Recurred’ for no progression. Each disease site was individually labeled, resulting in a total of 197 sites.

An example of a patient with three sites, two of which locally progressed, Figure 1.

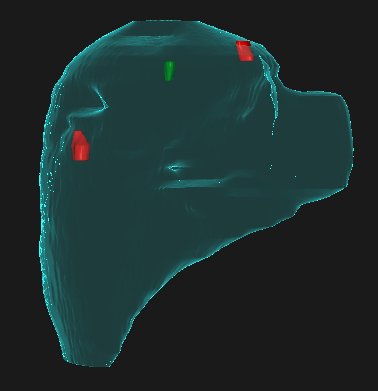


Figure 1: Example of patient with three disease sites. Two of the treated sites developed progression (red) and one site did not progress (green). Note that the expansion of the disease site outside of the liver is an artifact of 3D rendering.

Note that the expansion of the disease site outside of the liver is an artifact of 3D rendering, all sites are kept within the liver using ROI algebra.

#### Image Preparation

Automatic intensity based image registration was performed focused about the liver within the Raystation Treatment Planning system(Bodensteiner, 2018). Liver contours were generated automatically with our in-house liver auto-contouring algorithm(Anderson et al., 2020), manually inspected, and edited as needed.

Following rigid-registration, biomechanical deformable image registration within Raystation (Morfeus(Cazoulat, Owen, Matuszak, Balter, & Brock, 2016)) was applied using the liver contours as boundary conditions. The resampled, post-ablation image, pre-ablation image, and unsampled post-ablation image were then exported for presentation to the model.

Images and segmentations were converted into nifti files using our in-house DicomRTTool (Anderson, Wahid, & Brock, 2021). Using SimpleITK(Beare, Lowekamp, & Yaniv, 2018), all images and masks are resampled to dimensions of 5x1x1mm, with bilinear interpolation and nearest-neighbor resampling, respectively.

#### Presentation to model

Image ‘slabs’ of size 32x64x64 were extracted for each treated site, centered about the segmented disease. Several examples of this, with the pre-treatment image on the left, and the post-ablation on the right, are shown in Figure 2.

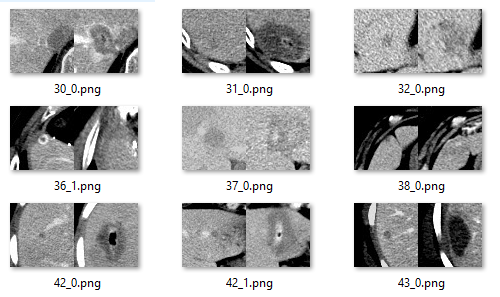


Figure 2: Examples of ‘slabs’ pulled form a variety of patients. Image on the left is pre-treatment and image on the right is post-ablation. Note that the nomenclature is patient ID\_site#, so 42\_0 and 42\_1 are both from the same patient.

#### Distributing the data

Data was randomly shuffled based on patient ID, placing each patient in the validation group until 1/3rd of the progression sites were in the validation set. Next, the remaining patients were moved into validation until 1/3rd of the non-progression sites were in the validation set. All remaining patients were placed in the training set. **Note that by splitting on patient ID we ensure no patients are distributed across both training and validation sets.** The final distribution of data can be seen in Table 1.

Table 1: Distribution of patients and lesions between train and validation set. Note that patients were split up based on patient ID, so a single patient will never be in both training and validation groups.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source Information | | | Status | Image Distribution | |
| Source | Patients | Lesions | ~Progression | Train | Validation |
| Progression |
| MD Anderson  Cancer Center | 119 | 197 | 132 | 94 | 38 |
| 65 | 43 | 22 |

### Architectures

We investigated several alterations on convolutional neural networks for feature extraction, into densely connected networks for outcome. The framework of the feature extraction has three parts: First, convolution blocks are performed in a manner similar to the DenseNet(Huang, Liu, Van Der Maaten, & Weinberger, n.d.) architecture, where previous feature maps are concatenated immediately to later layers. Second, a transition block reduces the number of filters by a fraction ranging from 0.5 to 1, where 1 indicates no reduction and 3D max pooling with a stride of 2. Third, feature extraction ends with a global max pooling layer.

The framework of the densely connected network has two parts: First, a set number of dense connections followed by dropout layer. Second, the model predicts two classes, progression and no-progression, with a soft-max activation.

We investigated several parameters in both the feature extractor, densely connected prediction, and parameters outside of the architecture.

Feature extraction parameters:

* 4, 8, 16, or 32 filters as the number of filters to start the feature extraction
* 4, 8, 16, or 32 as the rate of increase in the number of filters (growth rate)
* 1, 2, 3, 4, or 5 convolution blocks in each dense layers
* 1, 2, 3, or 5 dense convolution blocks followed by transition layers
* 0.5, 0.75, or 1.0 the fraction of filters to decrease by each transition

Densely connected parameters:

* 0, 1, 2, 3, 4, 5, or 7 number of densely connected layer
* 64, 128, 256, or 512 number of connections in each layer
* 0.5 or 0.0 dropout for fraction of connections to drop

Other hyper-parameters:

* Min and max learning rate were created on a model by model basis
* Adam or Stochastic gradient descent optimizers
* Categorical cross entropy, or Cosine loss

### Training the model

We investigated presenting the data in eight different ways, involving combinations of the pre-treatment image, the rigidly registered post-treatment image, the deformably registered post-treatment image, segmentations of the liver, and disease. A breakdown the combinations is shown in Table 2.

Table 2: The methods of presenting data to the model. Note that across all 8 the pre-treatment contrast enhanced CT is presented. Next, either the deformed or rigidly registered post-treatment image is presented. Segmentations of the liver and or disease are also included.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Imaging | | | Contours | |
| Pre-Treatment | Post-Treatment Deformed | Post-Treatment Rigid | Liver | Disease |
| x | x |  |  |  |
| x | x |  | x |  |
| x | x |  |  | x |
| x | x |  | x | x |
| x |  | x |  |  |
| x |  | x | x |  |
| x |  | x |  | x |
| x |  | x | x | x |

To balance the presentation of progression and no-progression, we created two TensorFlow data-generators with interleaving outputs. With a batch size of 32, this ensured each training step had 16 sites with progression and 16 sites without progression. Each generator will reshuffle and repeat, meaning the progression data is sampled more often than the no-progression data. This decision was influenced by recommendations from Tensorflow when working with imbalanced data, <https://www.tensorflow.org/tutorials/structured_data/imbalanced_data>.

#### Cosine Loss

Due to the limited, and biased (more no-progression than progression instances) nature of our dataset, we quickly noticed a problem of model overfitting. Normalization was introduced to the model in the form of L2 normalization and dropout in the dense connections, unfortunately it did not seem to alleviate the overfitting.

Recent work has shown that a large part of this issue can be related to the usage of softmax + crossentropy loss function(Barz & Denzler, n.d.), and propose the usage of a cosine loss function. The cosine loss focuses on increasing the similarity between the L2-norm of the prediction and the ground truth, ‘bounding’ the loss to a unit sphere. In contrast, the cross entropy and softmax contains exponential and log functions, allowing arbitrarily high values to appear.

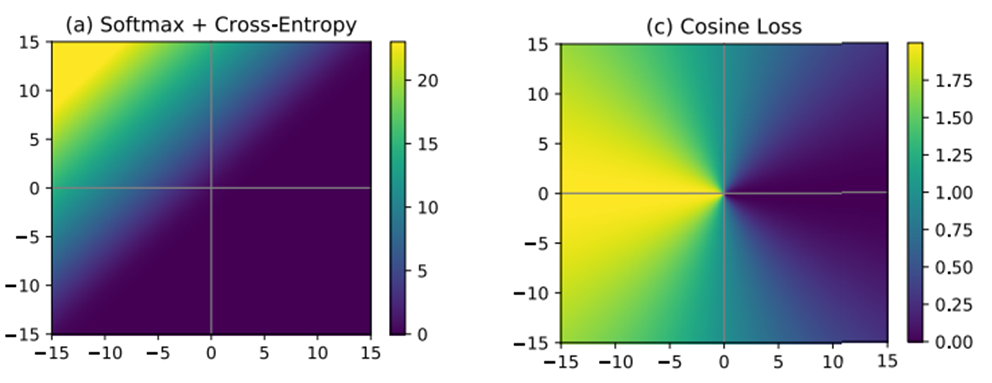


Figure 3: Heat maps of two loss functions in a 2-D feature space. Note that cosine loss is bounded from 0 to 2, while softmax + cross-entropy can take on arbitrarily large values. Figure copied from (Barz & Denzler, n.d.).

After adopting the cosine loss function the model appeared to have less overfitting, and was adopted for the rest of the work.

#### Learning Rates

Unique minimum and maximum learning rates were identified using our inhouse learning rate finder (GitHub link redacted), adapted from code presented in <https://www.pyimagesearch.com/2019/08/05/keras-learning-rate-finder/>. These minimum and maximum learning rates were used in a Cosine annealing learning rate scheduler (GitHub link redacted). Step-size was varied to be 2000, 5000, or 10,000, with a reduction factor of 0.5.

Each architectural style was run three times to ensure poor initialization was not the root cause of poor performance.

### Evaluating Training

Training was monitored using the Tensorboard module(Abadi et al., n.d.). Area under the curve (AUC) was the metric of choice for overall model performance.

Top performing models, as identified by the AUC, were selected for prediction visualization. To provide human interpretable explanations about the model predictions, we use the integrated gradient (IG) method. IG were originally proposed by Sundarajaran et al.(Sundararajan, Taly, & Yan, 2017) to identify which aspects of the image were most important in the decision process. Tensorflow offers a tutorial on implementation of the IG method (<https://www.tensorflow.org/tutorials/interpretability/integrated_gradients>) which we adapted from 2D images to 3D images as seen in our data. Each top performing model from the eight possible input data presentations given in Table 2 was investigated.

## Results

### Quantitative

We present the results between ‘pre-treatment and deformable post-treatment’ or ‘pre-treatment and rigidly registered post-treatment’. The accuracy, % of progression cases correctly predicted, the % of non-progression cases accurately predicted, and the AUC for each data presentation method are presented in Table 3. Note that accuracy, and % correct assume a 50% prediction cutoff.

Table 3: Accuracy, % of progression sites accurately predicted, % of non-progression sites accurately predicted, and the area under the curve (AUC) values for the eight presentation methods of data. Note that accuracy and % correct assumes a prediction cut-off of 50%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Description** | **Accuracy** | **% Progression Correct** | **% Non-Progression Correct** | **AUC** |
| Primary + Secondary Deform | 67 | 18 | 95 | 0.66 |
| Primary + Secondary Deform + Liver | 70 | 23 | 97 | 0.68 |
| Primary + Secondary Deform + GTV | 77 | 55 | 89 | 0.81 |
| Primary + Secondary Deform + GTV + Liver | 70 | 36 | 89 | 0.78 |
| Primary + Secondary Rigid | 63 | 0 | 100 | 0.66 |
| Primary + Secondary Rigid + Liver | 67 | 32 | 87 | 0.66 |
| Primary + Secondary Rigid + GTV | 72 | 32 | 95 | 0.72 |
| Primary + Secondary Rigid + GTV + Liver | 67 | 23 | 92 | 0.75 |

For deformable registered images, the presentation of pre-treatment and post-treatment had an AUC value of 0.66. Pre-treatment, post-treatment, and liver segmentations had an AUC of 0.68. Pre-treatment, post-treatment, and disease segmentations had and AUC of 0.81. Pre-treatment, post-treatment, liver and disease segmentations had and AUC of 0.78. The received operating characteristic (ROC) lines for the individual data presentation methods of pre-treatment and deformed post-treatment images are shown in Figure 4.

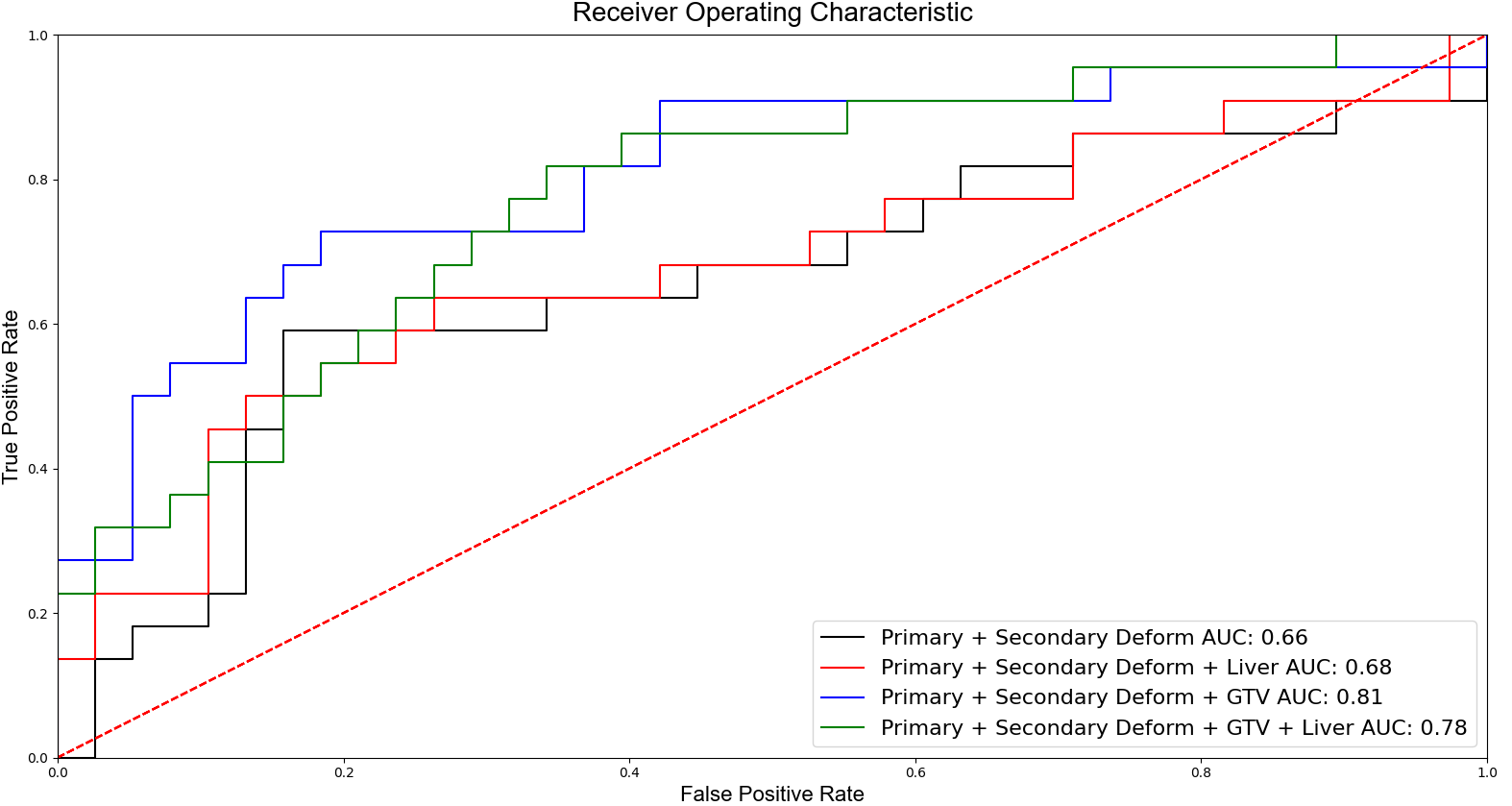


Figure 4: Receiver operating characteristic lines for each presentation of pre-treatment and deformable registered post-treatment data to the model. Area under the curve (AUC) recorded in legend.

For rigidly registered images, the presentation of pre-treatment and post-treatment had an AUC value of 0.66. Pre-treatment, post-treatment, and liver segmentations had an AUC of 0.66. Pre-treatment, post-treatment, and disease segmentations had and AUC of 0.72. Pre-treatment, post-treatment, liver and disease segmentations had and AUC of 0.75. The received operating characteristic (ROC) lines for the individual data presentation methods of pre-treatment and rigidly registered post-treatment images are shown in Figure 5.

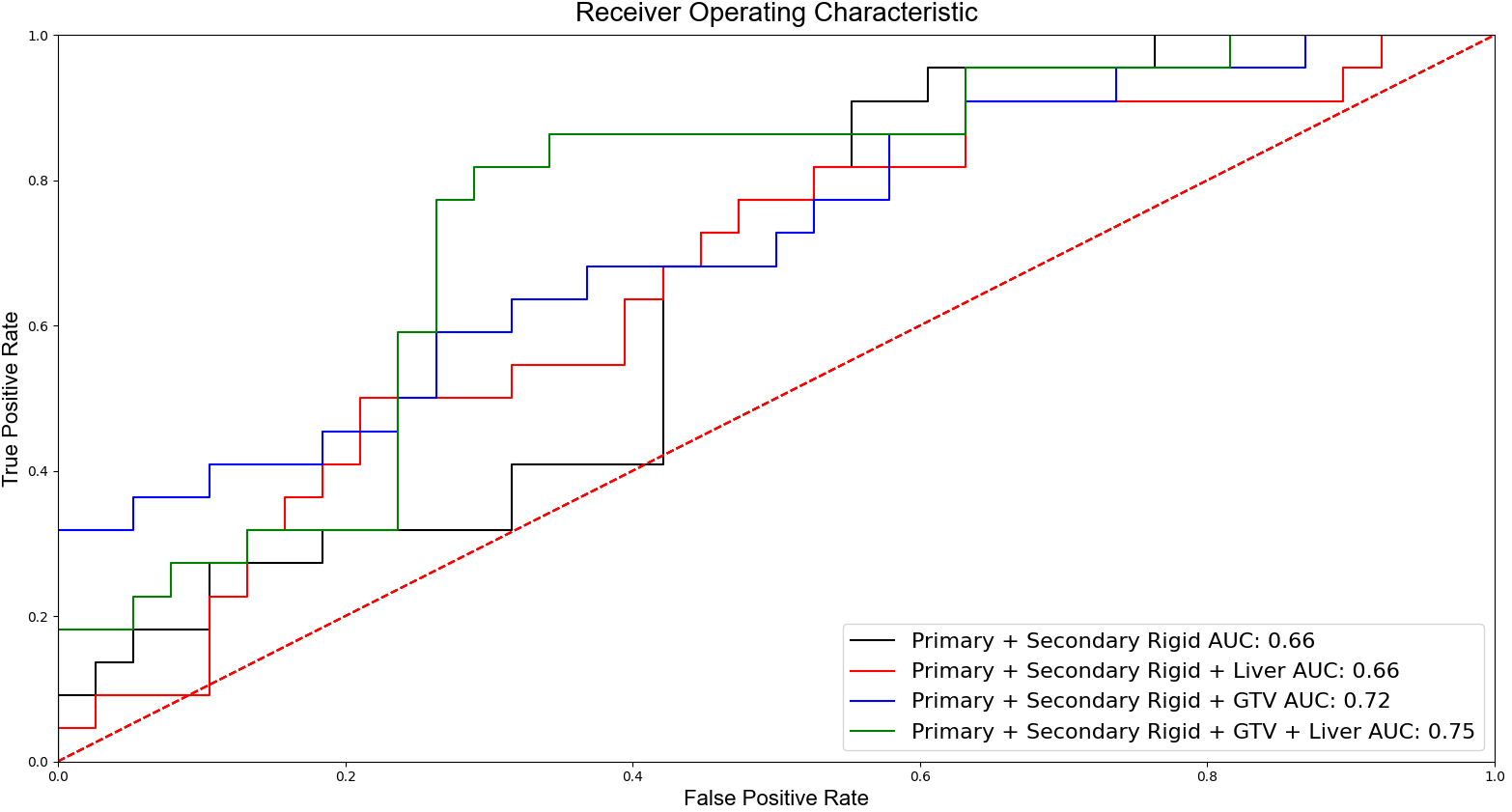


Figure 5: Receiver operating characteristic lines for each presentation of pre-treatment and rigidly registered post-treatment data to the model. Area under the curve (AUC) recorded in legend.

### Qualitative

Qualitative visualization of the IG varied across models and sites. We used visualizations of the IG to identify if the model was focusing on aspects of the image relevant to the PTA procedure. For example, in Figure 6, the model was presented the pre-treatment image, rigidly registered post-treatment image, and the contours of the liver and disease. Visualizing the IG, the model appears to be most focusing on regions overlapping between the ribs and lung, an artifact of the rigid registration, rather than near the disease site.



Figure 6: Overlapping heat map from Integrated Gradients on a progression case. Images (from left to right) are Pre-treatment CT, rigidly registered post-treatment CT, pre-treatment liver, disease, and post-treatment liver. Note that disease is blank, as the disease is not present in these slices. However, the IG show importance was heavily weighted to an area far from the disease.

## Discussion

Across all eight models there appears to be a similar relationship to the AUC as different data is presented. We suspect the pre-treatment + rigid secondary performed as well as the deformation model because large deformation was not present in most cases, and the disease and ablation sites were aligned. Similarly, the presentation of the liver in both cases provided negligible improvement, which we believe is also due to the fact that the image is already centered about the disease in the liver. The slightly poorer performance of the pre-treatment + deformed secondary + disease and liver was curious; investigating into the IG, we saw that the model appears to sometimes focus on the boundary edge of the liver, Figure 7, leading to confusing results.



Figure 7: Overlay of the integrated gradients onto images presented to the deformable model. From the left, the pre-treatment CT, post-treatment CT, liver contour, and disease. Note that the IG are focusing on the boundary of the liver contour in the top right corner and not the disease.

## Conclusion

While the AUC values from our top performing model are decent from a prediction standpoint at 0.81, the visualization of the IG shows that the model is not always ‘looking’ in what the user might consider to be a useful place. The model predictions prediction should be used as added information, not the determining factor in evaluating ablation performance. We believe that the prediction output of the model along with the IG can help provide meaningful information to the interventional radiologist.