Pneumonitis Risk Classification for Lung Cancer Patients

LAURA MACHLAB, Northwestern University
ZHUOYANG ZOU, Northwestern University

1 ABSTRACT

Lung cancer patients receiving radiation therapy are at risk of developing pneumonitis, a condition involving inflammation of the lungs, as a result of healthy lung tissue being exposed to radiation during treatment. There are several risk factors that affect if a patient is at high risk or not for experiencing this as an outcome of their treatment. This study focuses on providing risk classification for lung cancer patients prior to treatment based on their Computed Tomography (CT) scans. To do this, we train and test six different models. Five of which are CNN classifiers, and the sixth is a fine-tuned ResNet-18 model. The highest performing model is a CNN classifier trained on a dataset that has a 50/50 positive/negative ratio, uses the whole lung mask, a learning rate of 0.001, and is trained on 10 epochs. This model has a testing accuracy of 71.9 % and an AUC score of 0.766. However, all six models exhibited overfitting, a result of the limited size of the training dataset used. The results suggest that deep learning models can potentially aid in flagging patients at risk for developing pneumonitis prior to radiation treatment. Future research will focus on reducing overfitting and enhancing model structure for better feature extraction from lung images.

2 INTRODUCTION

Radiation therapy is one of the main treatment options for patients with lung cancer. Radiation therapy aims to kill cancer cells and reduce the size of tumors using high doses of targeted radiation [5]. Though this specifically targets cancerous cells, surrounding healthy tissue is also exposed to the toxicity of radiation. This exposure is damaging and can result in additional health complications. One consequence to healthy lung tissue exposure to radiation is pneumonitis. Radiation pneumonitis is inflammation of the lungs that can lead to permanent scarring and decrease in function of the lungs [6].

Some patients are more likely to develop pneumonitis than others due to a variety of risk factors. While some risk factors are patient-specific features, such as history of smoking or previous lung functioning, the largest risk comes from the amount of radiation the patient is exposed to [7]. If patient risk for pneumonitis can be determined prior to treatment, this information can be used to tailor dosage and treatment plans.

We set out to determine if deep learning image classification can be used to flag lung cancer patients as either high risk or not for developing pneumonitis as a result of the toxicity of radiation therapy. Specifically, our approach uses Convolutional Neural Networks (CNNs) to identify lung features from Computed Tomography (CT) scans that may indicate risk for developing pneumonitis. The patients' CT scans provide information on visual features of their lungs as well as the size and location of the tumor.

3 RELATED WORK

Several studies have focused on utilizing advanced imaging techniques and machine learning methods to predict patient outcomes and complications in the context of lung cancer and radiation therapy. Braghetto et al. conducted a study titled "Radiomics and deep learning methods for the prediction of 2-year overall survival in LUNG1 dataset," where they investigated the performance of radiomics and convolutional neural networks (CNNs) on CT scans for predicting 2-year overall survival in lung cancer patients [1].

In their research, Braghetto et al. employed radiomics, a method that extracts quantitative features from medical images, to characterize tumor heterogeneity and extract prognostic information. They combined radiomics features with deep learning techniques using CNNs to improve the predictive accuracy. By analyzing the LUNG1 dataset, they were able to demonstrate the potential of radiomics and CNNs in predicting patient survival. This study highlights the importance of incorporating imaging data and machine learning algorithms to enhance prognostic predictions in lung cancer patients.

Furthermore, Wang et al. conducted a study titled "A Novel Nomogram and Risk Classification System Predicting Radiation Pneumonitis in Patients With Esophageal Cancer Receiving Radiation Therapy" [2]. Their objective was to develop a predictive model for severe acute radiation pneumonitis (SARP) in patients with esophageal cancer undergoing radiation therapy. They investigated various patient characteristics, including the cystic fibrosis score and planning target volume/total lung volume, to identify indicators of developing SARP.

The study by Wang et al. successfully developed a novel nomogram and risk classification system that incorporated multiple patient characteristics to predict the risk of developing radiation pneumonitis. Their findings reinforced the association between lung-related features and the likelihood of pneumonitis occurrence in patients undergoing radiation therapy for esophageal cancer. This research contributes to the understanding of lung-specific factors that contribute to the development of radiation-induced complications.

In summary, both studies explored the potential of advanced imaging techniques and machine learning methods for predicting patient outcomes and complications in the context of lung cancer and radiation therapy. Braghetto et al. focused on predicting 2-year overall survival using radiomics and CNNs on CT scans, while Wang et al. developed a predictive model for severe acute radiation pneumonitis in patients with esophageal cancer undergoing radiation therapy. These studies collectively underscore the significance of incorporating imaging data, patient characteristics, and machine learning algorithms to improve prognostic predictions and guide

personalized treatment approaches in lung cancer and radiation therapy settings.

4 DATASET

The data we use consists of 1,169 chest CT scans of lung cancer patients prior to treatment. Of the 1,168 CT scans, 812 of them were collected at Cleveland Clinic and 356 were collected at Northwestern Medicine. Each CT has a set of contour labels outlining regions of interest (ROIs) in the CT scan, as well as a label indicating if the patient developed pneumonitis as a result of radiation therapy. There are 16 patients in the Northwestern group and 48 patients in the Cleveland group positive for pneumonitis, meaning that there is a total sample size of 64 for pneumonitis positive patients.

Location of Collection	Number of CT scans	Number of positive Pneumonitis cases
Northwestern Medicine	356	16
Cleveland Clinic	812	48
Total	1168	64

Table 1. CT Scan Data

To balance the dataset, we undersample from the pneumonitis negative cases. We prepare two different distributions. The first distribution is composed of 30 percent positive and 70 percent negative cases, and the second distribution is 50 percent each. To make sure that one hospital is not overrepresented in the negative cases, a proportional amount of negative cases is sampled randomly from each set. This means that in the balanced dataset, the Northwestern CT scans alone have the target distribution – either 30/70 or 50/50 – as do the Cleveland CT scans.

Once the dataset is balanced, we split it into train, development, and test sets. This is done with training data making up 70 percent, development data making up 15 percent, and testing data making up the remaining 15 percent.

We read the CT scans into NumPy arrays using the contours as masks to select CT scan information limited to the decided ROI. The CT scans are saved as DICOM images, and we use DicomRTTool, a python module developed at the University of Texas, to convert the information into NumPy arrays [3]. This method is used to create four datasets. The first comes from the 30/70 positive/negative distribution with both lungs – the whole lung – as the ROI. The following three datasets are composed with the 50/50 positive/negative distribution, one with whole lung as the ROI, one with the ipsilateral lung as the ROI, and one with the contralateral lung as the ROI. Ipsilateral refers to the lung which contains the cancerous tumor, and contralateral refers to the lung without the tumor.

When reading each CT scan using DicomRTTool, 2D NumPy arrays representing slices of the CT are extracted, and every fourth one is added to the dataset. Skipping slices ensures variation between used arrays with the goal of reducing memorization done by the model. Each slice is labeled based on the pneumonitis outcome for the corresponding patient.

5 MODEL

We use two different approaches for pneumonitis risk classification. The first is to train a convolutional neural network (CNN) classifier, and the second is to fine-tune a pre-trained ResNet, ResNet-18, to perform classification. Both models take 2D CT slices as inputs and produce binary classifications as outputs.

5.1 CNN

The CNN has two convolutional layers followed by max pooling, and two fully connected layers. The first convolutional layer extracts 32 feature maps using 3x3 filters, followed by the second convolutional layer that generates 64 feature maps of the same size. Both convolutional layers incorporate padding to preserve the spatial dimensions of the feature maps. Max pooling is then applied, by selecting the maximum value in non-overlapping 2x2 windows, to downsample the feature maps. The flattened feature maps are transformed by the fully connected layers, with ReLU activation introducing non-linearity. The last fully connected layer serves as the output layer, producing predicted probabilities for the two classes. We used the Adam optimizer and cross entropy loss function.

5.2 Resnet18

ResNet-18 is a convolutional neural network (CNN) model that has 18 layers, including convolutional layers, pooling layers, fully connected layers, and an output layer [4]. The architecture follows a modular structure, with a series of residual blocks. Each residual block contains two convolutional layers, batch normalization, and rectified linear unit (ReLU) activation functions. Skip connections, or shortcut connections, establish residual connections between layers, facilitating the flow of information through the network. This design allows ResNet-18 to address the problem of vanishing gradients and enables the training of deeper networks without sacrificing performance. The depth of ResNet models allow them to capture complex patterns, making them suitable for medical image classification. Additionally, transfer learning is leveraged by utilizing a ResNet-18 model pretrained on a large-scale dataset, before fine-tuning it on our specific dataset. This approach harnesses the knowledge and feature representations learned from the pretraining phase, allowing the model to capture relevant features and improving its classification performance.

In our implementation, we imported the pretrained ResNet-18 from PyTorch and then applied a couple of adjustments to the structure. The first modification was to change the number of input channels from 3 to 1, given that the CT scans are grayscale rather than RGB. The second modification was to the output layer to adjust the model for binary classification.

6 RESULTS

6.1 Evaluation Metrics

We evaluate our model using two primary metrics: accuracy and the Area Under the Curve (AUC) score. The AUC score is derived from the Receiver Operating Characteristic (ROC) curve, a graph that illustrates the performance of a classification model across all potential classification thresholds. Specifically, the ROC curve plots two crucial parameters — the True Positive Rate (TPR) and False Positive Rate (FPR). So, the AUC score offers an aggregate measure of model performance across every possible classification threshold. An AUC score of 1 signifies perfect predictive performance, indicating that every prediction made by the model is accurate.
The Area Under the Curve (AUC) for a Receiver Operating Characteristic (ROC) curve is given by the definite integral:

The Area Under the Curve (AUC) for a Receiver Operating Characteristic (ROC) curve can be calculated using Sensitivity (TPR) and 1 - Specificity (FPR):

$$\text{AUC} = \int_0^1 \text{TPR}(FPR)\, dFPR \tag{1}$$

where TPR represents the Sensitivity and FPR represents 1 - Specificity. In a discrete case, the AUC can be approximated by summing the areas of the trapezoids formed by successive points on the ROC curve:

$$\text{AUC} \approx \sum_{i=1}^{n-1} \frac{(FPR_{i+1} - FPR_i)(TPR_{i+1} + TPR_i)}{2} \tag{2}$$

The accuracy for our models can be calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives.

6.2 Overall Results

We perform hyperparameter tuning across the six models we trained or fine-tuned. Our initial CNN had a learning rate of 0.001 and was trained on data with the 30/70 positive/negative ratio and whole lung mask. It converged within 10 epochs and had a testing accuracy of 66.9 percent. We then tried training the same CNN with a decreased learning rate of 0.0001, which produced a testing accuracy of 68.9 percent. This increase in accuracy was not a result of more robust feature learning but rather due to an increase in negative predictions. Both initial CNN models were converging towards classifying all patients as negative, and thus the accuracy was nearing 70 percent, the percentage of all negative cases in the test set. To correct this class imbalance, the following models were then trained on the 50/50 ratio dataset. The CNN trained on the 50/50 whole lung dataset over 10 epochs with a learning rate of 0.001 performed significantly over chance with an accuracy of 71.9 percent. The ResNet-18 model was fine-tuned on that same dataset over 50 epochs, with a learning rate of 0.001, and had an accuracy of 60.8 percent. The two remaining models were trained on the 50/50 dataset, with the same learning rate of 0.001, and over 10 epochs, but one was trained with the ipsilateral lung mask, and the second was trained with the contralateral lung mask. The ipsilateral lung classifier performed with 47.5 percent accuracy on the test set, and the contralateral lung classifier performed with 45.0 percent accuracy on the test set.

Model	Lung ROI	Epoch	Learning rate	Dataset Ratio (neg/pos)	Accuracy	AUC Score
CNN	Whole	10	0.001	0.7	0.6693	0.4913
CNN	Whole	10	0.0001	0.7	0.6893	0.4879
CNN	**Whole**	**10**	**0.001**	**0.5**	**0.7188**	**0.7663**
ResNet18	Whole	50	0.001	0.5	0.6081	0.6096
CNN	Ipsi	10	0.001	0.5	0.4745	0.4532
CNN	Contra	10	0.001	0.5	0.4497	0.4301

Table 2. Model Training Parameters and Subsequent Results

Our highest performing model on both accuracy and AUC metrics was the whole lung CNN trained on the 50/50 dataset. This model had a test accuracy of 71.9 percent and test AUC score of 0.766. While the fine-tuned ResNet had a lower accuracy than the CNNs trained on the 30/70 dataset, it can be seen in the confusion matrix that it has a balanced distribution between positive and negative predictions. Additionally, the fine-tuned ResNet has an AUC score of 0.610, which is the second highest among all models.

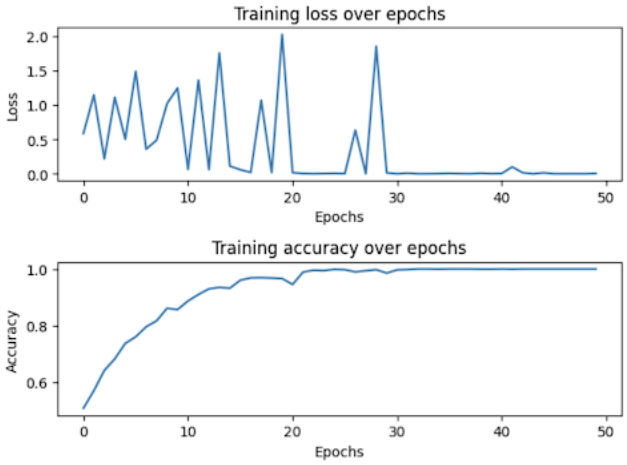
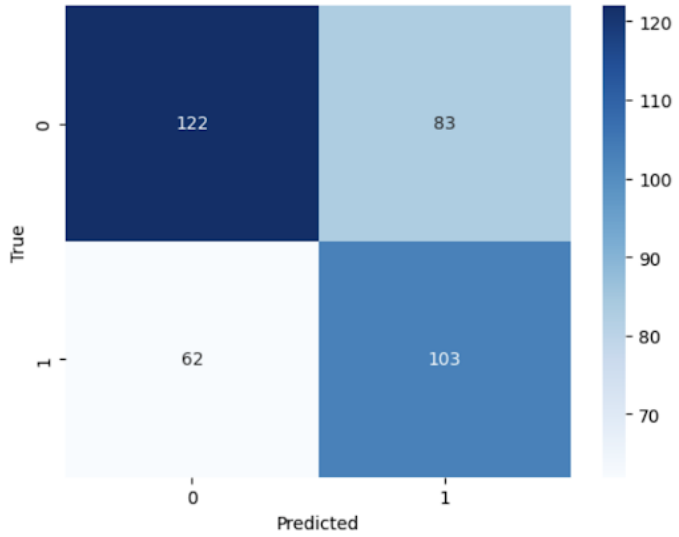


Fig. 1. Learning rate of Resnet18 during training

7 DISCUSSION

While we successfully trained a model that can predict patients' risk for developing pneumonitis as a result of radiation treatment with accuracy above chance, we see the limitations in our results. The largest limitation comes from the size of the dataset used to train the six models. A manifestation of this can be seen in the steep training loss and accuracy curves. Within 10 epochs of training the CNN classifier, there is convergence in the loss and accuracy. This is




```
Accuracy: 0.6081081081081081
ROC AUC: 0.6096821877309682
```

Fig. 2. confusion matrix of ResNet18 Model on testing data

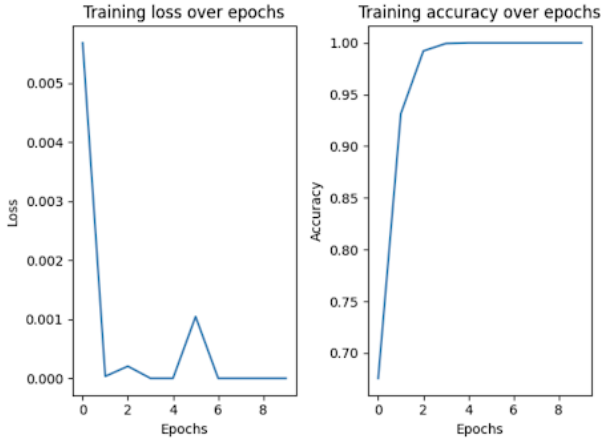
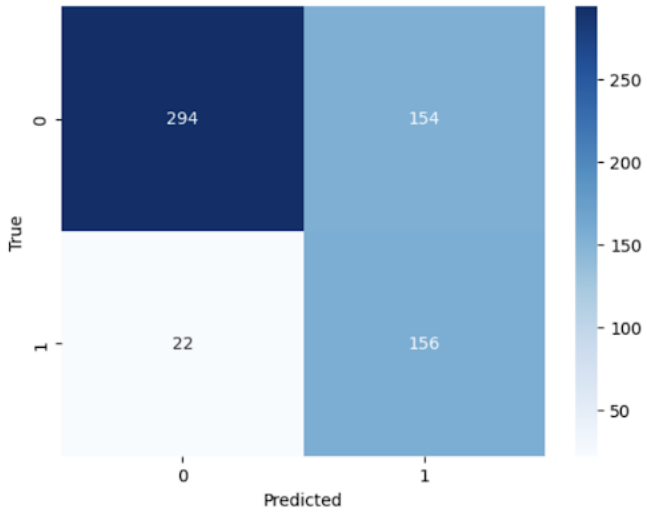


Fig. 3. Learning rate of 50/50 whole lung CNN model during training

an indication that the model is memorizing the training data rather than learning the correct features to be looking for in each lung CT scan. This is also the cause for the gap between training and testing accuracy.

A second limitation comes from the masks we were able to use to build datasets. Initially, we had planned two additional models, the first using only information from the ipsilateral lung, subtracting the gross tumor volume (GTV). The second additional model was intended to contain both lungs, subtracting the GTV. Because our data was sourced from two different hospitals, the labeling conventions are different. The Cleveland data does not have standardized




```
Accuracy: 0.7188498402555911
ROC AUC: 0.7663272471910112
```

Fig. 4. confusion matrix of 50/50 whole lung CNN Model on testing data

labels across CT scans, so we are unable to select "ipsilateral - GTV" as a mask. So, our results only come from models trained on data containing the GTV in the training data

8 FUTURE WORK

To address the current limitations, future work relies heavily on increasing the size of the dataset used to train the models. Because there are many more pneumonitis negative CT scans than positive, increasing the dataset requires increasing the number of negative samples. To do this, current positive CT scans can be oversampled using data augmentation.

There are a few other next steps that we have identified. The first is standardizing the contour drawing and naming conventions among all CT scans used to train the additional models as described in the previous section. Secondly, after fine-tuning ResNet-18 again after expanding the training dataset, we think it would be beneficial to find other relevant pre-trained models suited for CT scan classification to compare performance. Finally, we would like to try using the radiomics method used in Braghetto et al. and compare its performance to the simple CNNs and fine-tuned ResNet-18.

9 CONCLUSION

In this paper, we presented our study on predicting the risk of pneumonitis in patients prior to radiation treatment using machine learning techniques. Our best model achieved an accuracy of 71% in classifying patients at risk of developing pneumonitis. We identified several areas for future improvement, including increasing the size of the dataset, specifically the number of positive cases, to enhance the model's training, standardizing contour labels to train models on

datasets using different masks, and fine-tuning alternate pre-trained models specific to our problem space.

REFERENCES

[1] Braghetto, Andrea et al. "Radiomics and deep learning methods for the prediction of 2-year overall survival in LUNG1 dataset." https://www.nature.com/articles/s41598-022-18085-z

[2] Wang, Lu et al. "A Novel Nomogram and Risk Classification System Predicting Radiation Pneumonitis in Patients With Esophageal Cancer Receiving Radiation Therapy." https://www.redjournal.org/article/S0360-3016(19)33659-4/fulltext

[3] Anderson, B. M.; Wahid, K. A.; Brock, K. K. "Simple Python Module for Conversions Between DICOM Images and Radiation Therapy Structures, Masks, and Prediction Arrays." Practical Radiation Oncology, Volume 11, Issue 3, 2021, Pages 226-229, ISSN 1879-8500, https://doi.org/10.1016/j.prro.2021.02.003. https://www.sciencedirect.com/science/article/pii/S1879850021000485

[4] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." Microsoft Research. https://arxiv.org/pdf/1512.03385.pdf

[5] "Radiation Therapy to Treat Cancer" National Cancer Institute https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy

[6] "Radiation Pneumonitis" Canadian Cancer Society https://cancer.ca/en/treatments/side-effects/radiation-pneumonitis

[7] "What is Radiation Pneumonitis and How Is It Treated?" Healthline https://www.healthline.com/health/radiation-pneumonitis