A Deep Learning Pipeline for Lung Cancer Classification on Imbalanced Data Set

By

Qingyun Qian

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Electrical Engineering

May 31, 2020

Nashville, Tennessee

Approved:
Bennett A. Landman, Ph.D.
Richard A. Peters, Ph.D.

Table of Contents

Page

List of Figures

Figure **Page**

List of Tables

Table **Page**

Chapter I. Introduction

Lung cancer is the leading cause of cancer death and the second most diagnosed cancer in both men and women in the United States and cigarette smoking is the number one cause of lung cancer (Centers for Disease Control and Prevention,2019). The National Lung Screen Trial(NLST) with chest screening of current and former smokers is the largest randomized study of lung cancer screening of high-risk population (Gatsonis et al. ,2011). 68996 CT images from 25183 subjects can be downloaded from NLST study and 626 patients were diagnosed within one year from latest biopsy. Many Computer Aided Diagnosis(CAD) systems are developed in recent years to detect lung cancer at early stage (Donahue et al., 2015; Gao et al.,2019; Hua et al., 2015; Liao et al.,2019; Santeramo et al., 2018; Xu et al.,2019). For Computer Tomography(CT) scans, several types of deep learning architectures are introduced and proven to be best method for medical imaging (Tekade & Rajeswari, 2018). In this project, a deep learning based pipeline is applied on data from the NLST study and approaches to deal with imbalanced data is discussed.

1. Related Works on Lung Cancer Detection

A major goal of CAD is to automatically classify malignant/benign nature of tumors based on image features and traditional CAD scheme usually includes a number image processing tasks followed with a classification and the performance depends heavily on the results of image processing tasks. In recent years, a variety of deep learning based techniques are developed to explore high level features from training images.

Conventional CAD scheme requires image processing and pattern recognition steps to extract quantitative features from nodules (Hua et al., 2015). Farag et al. (2011) applied geometric feature descriptors to extract features form nodule candidates and provided 2% enhancement of in specificity compared with classification based on normalized cross-correlation. Lin et al. (2013) attempted the technique of fractal analysis based on fractional Brownian motion model to extract feature vectors and the approach outperforms previous approaches based on change on CT attenuation value.

A number of current methods for the task of nodule classification are based on convolutional neural networks (CNN). For example, Hua et al. (2015) introduced a deep belief network(DBF) (Hinton, Osindero & Teh, 2006) and a convolutional neural network(CNN) (Krizhevsky, Sutskever, & Hinton, 2012) for nodule classification in CT images and suggested deep learning methods have better performance and promise than conventional CAD scheme. Two-dimensional region of interest(ROI) of pulmonary nodule depicted in a two-dimensional CT slice is served as input data and features are extracted without computing actual morphology and texture features. This method outperforms conventional hand-crafted feature computing frameworks (Hua et al., 2015).

The problem of cancer detection in whole scan is often divided into 2 steps, nodule detection and cancer classification. Liao et al. (2019) proposed a pipeline using 3D deep neural network that detects suspicious nodules and evaluates the whole lung malignancy. The lung is first segmented from other tissues. Then, a 3D region proposal network(RPN) (Ren et al., 2015) using modified U-net (Ronneberger, Fischer, & Brox, 2015) as backbone model is applied to detect all suspicious nodules in the lung. The nodules are scored and their cancer probabilities are combined with the leaky noise-or model to get an overall cancer probability for this scan (Liao et al. ,2019). The model won the first place in the Data Science Bowl 2017 competition with the training and

testing area under the ROC curves(AUC) of 0.90 and 0.87 on the Lung Nodule Analysis 2016 dataset (LUNA) and the training set of Data Science Bowl 2017 (DSB).

2. Related Works on Lung Cancer Detection Based on Serial Scans

In practice, time series of CT scans are often analyzed and methods that incorporate serial imaging data are developed. Recurrent neural networks(RNN) is a state of art deep learning method for video and natural language processing as the network incorporate longitudinal data (Donahue et al., 2015). Xu et al. (2019) evaluated RNN in analyzing time series of images. The model they proposed is used transfer learning of CNN with RNN and the results demonstrated that deep learning can integrate imaging scans at multiple time points to improve clinical outcome predictions (Xue et al., 2019).

Long Short-Term Memory(LSTM) (Hochreiter & Schmidhuber, 1997) is a popular variation of RNNs and a variety of LSTM based methods are proposed for the task of classifying longitudinal CT scans. However, traditional LSTMs are usually applied on data with regular time gaps between observations (Hochreiter & Schmidhuber, 1997) and in the case of medical exams, scans are collected at times of clinic need. As a result, the scans may not be equally spaced in time and number of scans for each subject could vary greatly (Santeramo et al., 2018). Santeramo et al. (2018) modified LSTM architecture to take time interval between consecutive scans into consideration and the time modulated LSTM improved classification performance on both real-world and simulated 2-D chest x-ray data.

Gao et al. (2019) took global temporal variation into consideration for the reason that the last scan is typically the most informative in lung cancer detection and proposed Distanced LSTM, a new Temporal Emphasis Model(TEM) to model the global time interval between previous time points to the last scan as a global multiplicative function to input gate and forget gate (Gao et al., 2019). The DLSTM method is trained in a lightweight post-processing manner for the features extracted from Liao et al. (2019) on a subset of NLST data and clinical data. The model is proven to outperform CNN based methods, traditional LSTM and time modulated LSTMs.

3. Introduction to NLST

The National Lung Screening Trial(NLST) is a randomized multicenter study comparing low-dose helical computed tomography (CT) with chest radiography in the screening of older current and former heavy smokers for early detection of lung cancer. Each NLST participant was randomized to undergo a baseline and two annual screenings by using either low-dose CT or chest radiography (Gatsonis, 2011). 68996 CT scans from 25183 subjects can be downloaded from NLST study and 626 patients were diagnosed within one year of latest biopsy.

4. Approaches for Imbalanced Data

In NLST data set, only about 2.5% patients belong to the positive class and the rest are negative, which is a highly skewed data set. The class imbalance problem occurs in a large of domains including diagnostic problems while cancer cases are usually fairly rare compared with normal cases. The skewed data distribution could influence the modelling of rare events (Sun et al.,2009). However, degree of class imbalance is not the only factor influencing the performance. Apart from class distribution, the problem also depends on complexity of concept

represented by the data, the size of the training set and the type of classifier (Japkowicz & Stephen, 2002). Experiments were decided by Japkowicz & Stephen (2002) using three types of machine learning classifiers and they suggest that assuming in a large enough data set, the imbalance problem may not be an obstacle. As the size of NLST is quite large, this project uses subsets of different sizes as training data to explore the problem that to which extent class imbalances are damaging for classification in NLST data with the proposed classification pipeline.

Common solutions for this problem includes data level approaches like resampling data, algorithm-level approaches, cost-sensitive learning and boosting approaches (Sun et al.,2009). The resampling approach is often used in dealing with the problem and it contains oversampling the minority class and eliminating data from majority class. However, the optimal class distribution in each resampled batch might not be 1:1 (Weiss et al., 2003). As the goal of the project is to overcome data imbalance and use as much data as possible, downsizing majority data is not considered. Weighting the loss function would make sure that the modified distribution id biased towards the costly classes (Sun et al.,2009). As the minority class is of more importance, giving it more weight should decrease misclassification for this class.

In this experiment, oversampling the minority class and weighting the loss function are used for this problem.

Chapter II. Method

1. Data

Data from NLST study, consisting of 68996 CT images from 25183 subjects are used in this project, each patient was screened 1 to 3 times in year 1999, 2000 and 2001. CT images from 6335 of these patients have received confirmed biopsy results of cancer diagnosis or not cancer, of which 1060 patients was confirmed with cancer, and 626 of cancer patients were diagnosed within one year of latest biopsy. Theses scans with biopsy results are recommended by the study. Official result of screening exam after comparing with prior screening images was provided for each patient with or without biopsy. One of the following 6 labels results given to each patient (National Cancer Institute, 2014).

1= Negative screen, no significant abnormalities

2 = Negative screen, minor abnormalities not suspicious for lung cancer

3 = Negative screen, significant abnormalities not suspicious for lung cancer

4 = Positive, Change Unspecified, nodule(s) >= 4 mm or enlarging nodule(s), mass(es), other non-specific abnormalities suspicious for lung cancer

5 = Positive, No Significant Change, stable abnormalities potentially related to lung cancer, no significant change since prior screening exam

6 = Positive, other

In previous studies, label 1 for cancer and 0 for not cancer were assigned to only images of patients going through biopsy to assure image quality. In this experiment, in addition to scans from patients with biopsy results, images from 6135 patients without biopsy results but received screening result '1= Negative screen, no significant abnormalities' were also assigned label 0. Labels given by biopsy diagnosis are referred as 'hard labels' and other labels given only with screening results were 'soft labels'.

With only 10.6% of patients with hard labels and 5.2% of all data used has label 1, the NLST data is quite large but imbalanced. 7 different subsets of the data with increasing size of training sample and degree of class imbalance were used to investigate the effect negative samples have.

Subset 1(626 patients):

Images of all 626 positive patients were used and 626 patients with biopsy result of not cancer (later referred as negative data part A) were randomly selected so that number of positive and negative samples were equal. The data were then randomly divided into 5 folds. For each trial, 1 of the 5 folds were used as test set, 1 as validation set and rest 3 folds as training sets. Models with best performance on the validation set were used on the test set so that test data were not involved in any process of training. After 5 trials, average test AUC is recorded.

Subset 2(2383 patients):

For the rest of data were also used in this project but as training data only. Apart from 626 negative patients selected in subset 1, other 4642 patients with biopsy diagnosis of not cancer (later referred as negative data part B) were randomly divided into four groups of approximately 1/4 of data so that 1/4, 1/2 and all of them can be added to training data for subset 2,3, and 4. For subset 2, one of the four groups, approximately 1/4 of these 4642

patients were added but these new data were used only in training. Data in subset 1 were divided into 5 folds the same way. For each trial, apart from 3 folds of subset 1, the 1/4 of negative data part B were added to the training set. Test and validation sets were same as subset 1 to keep same test data to compare.

Subset 3(3505 patients):

Similar to subset 2 except 2 groups, half of negative data part B were added to training set apart from 3 folds of data in subset 1.

Subset 4(5895 patients):

Similar to subset 2 and 3 but all negative data part B were added to training.

Subset 5(7485 patients):

For subsets 5,6 and 7, 6145 patients with only negative screening results and no abnormalities (negative data part C) were used. Similar to part B, these negative data were also separated into 4 groups and added to training set. In subset 5, all of negative data part B and 1/4 of negative data part C were added to 3 folds of subset 1 as training set for each fold.

Subset 6(9014 patients):

Similar to subset 5, 3 folds of subset 1, all of negative data part B and half of part C were used as training set for each trial.

Subset 7(12029 patients):

All negative data in part B and part C were added in training.

2. Data Quality Assurance

All data downloaded from the websites are in Digital Imaging and Communication in Medicine(DICOM) format and a quality assurance (QA) program is run to remove problematic images with missing slices like figure 2 before converting them to NIFTI format. The first step of QA is to read DICOM header to check if instance number matches number of DICOMs. If number of DICOMs is smaller than number of instance number read in the header, missing slices exist. The next step is to check if slice distance, as missing slice may cause difference in slice distance even if instance number and DICOM number matches. Moreover, images with less than 20 slices were also removed (Gao,2019). Images passed QA were then converted into NIFTI format.

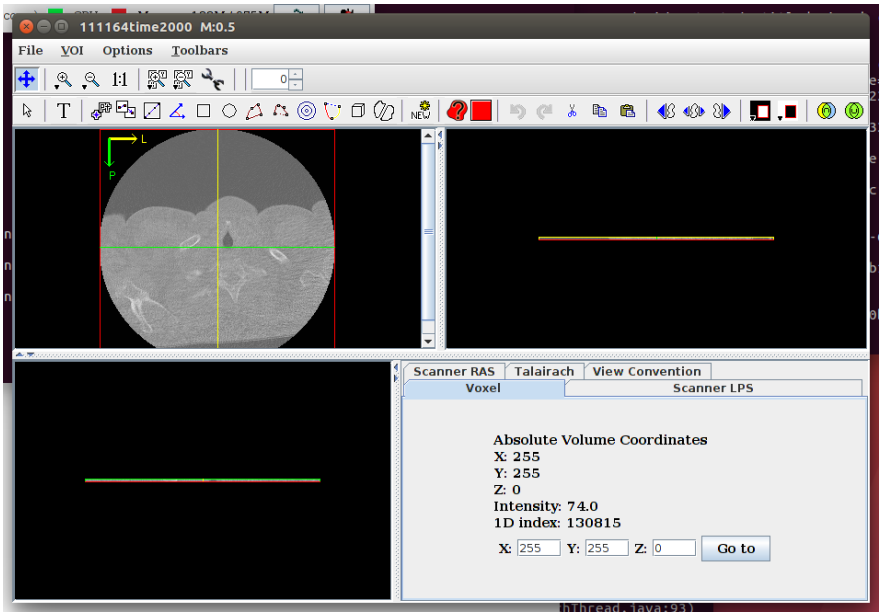

Figure 1 CT image in NLST data set with missing Slices

In the NLST data set, instead of missing slices, some images have problems of duplicate slices. These duplicate slices are removed when converting the scan to NIFTI and after visually checking, these scans with duplicate slices were kept.

The QA process was run on all 68996 DICOMs downloaded including those not used for following experiments and 1073 images were removed because of missing slice or uneven slice distance. For patients with biopsy results, images failing QA were visually checked and Table 1 shows count for images failing QA.

Table 1 Count for images failing QA for patients with biopsy results

	Cancer	Non-cancer
Total # Images	2285	13836
Missing Slices	16	29
Less than 20 Slices	12	22
Slice Distance Problem	33	77
Total # with Problems	61	128

3. Preprocessing

After converting images, the preprocessing steps proposed by Liao et al.(2019) were used to segment lung out of other tissues. Figure 3 shows the steps for preprocessing.

The images in NLST data set are already in Hounsfield Unit (HU) and the first step is to extract a mask slice by slice. For each 2D image, a Gaussian filter is applied then binarized at -600 threshold. Then small and eccentric 2D connected components were removed and then 3D connected components at the center position were kept.

Convex hull of the mask was then computed to capture nodules attached to the outer wall of the lung. The masks were then separated into left and right lungs by eroding iteratively until broken into two components with similar volumes. Finally, images were clipped within [-1200,600] converted from HU to UINT8. Components outside the masks were filled with intensity 170 (Liao et al.,2019). Figure 2 and 3 shows the tri-planar view of a scan before and after pre-processing.

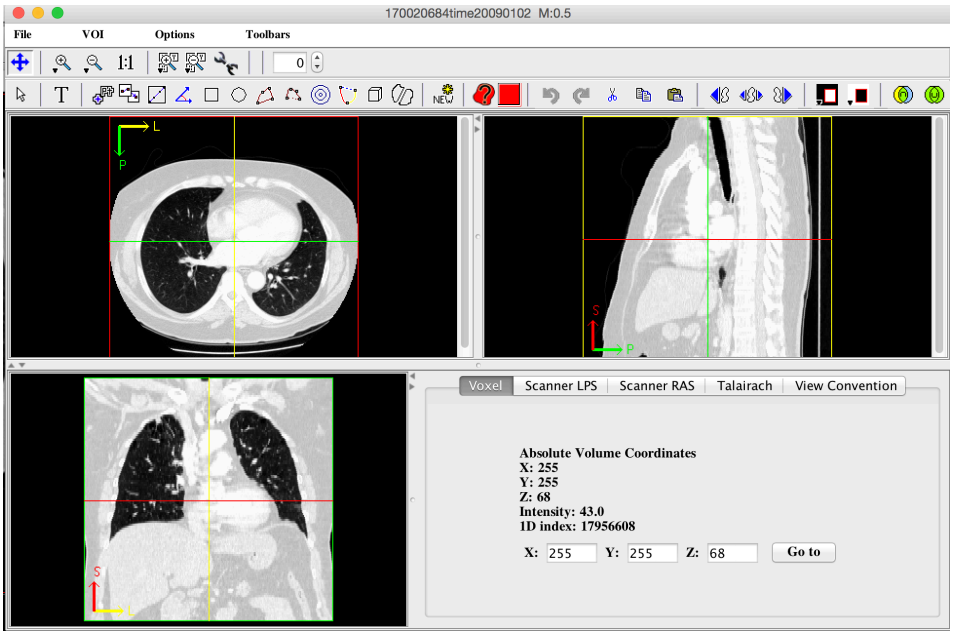

Figure 2 Tri-planar view of a scan before preprocessing

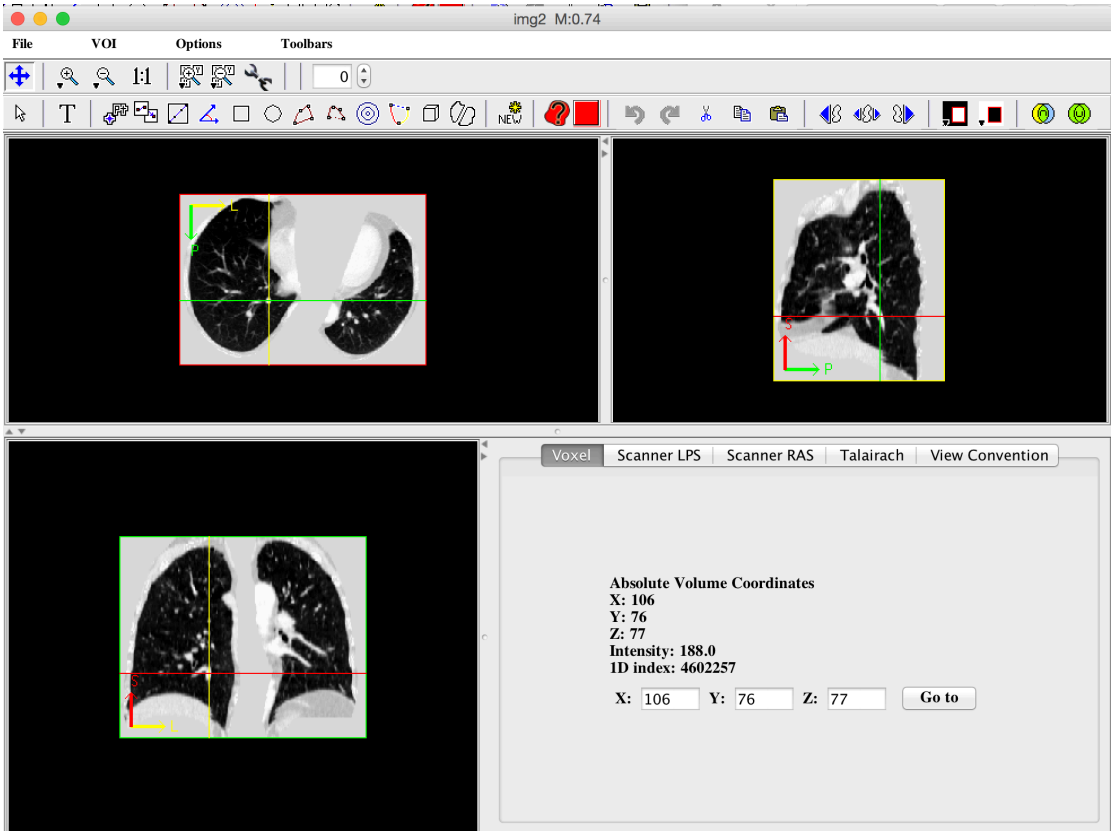

Figure 3 Tri-planar view of the same scan after preprocessing

All data downloaded and passing QA process were processed but the preprocessing steps do not run successfully on all images. Two more quality problems on some NLST data were detected in preprocess step. The first problem is some of the images contain 2 volumes like figure 4. The second problem is that a few images are not in HU shown in figure 5. 61 of 16054 images from patients with biopsy cannot be preprocessed and these images were removed.

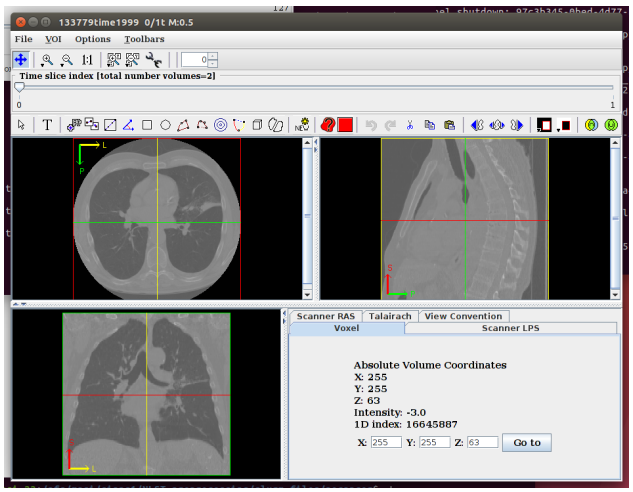

Figure 4 NLST image with 2 volumes

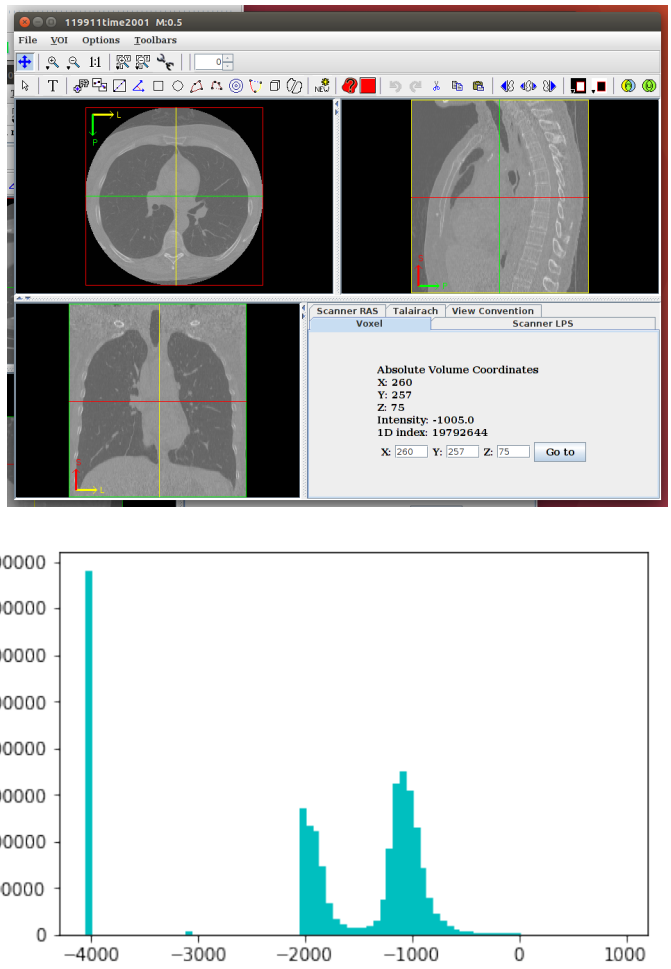

Figure 5 Preview and histogram of a Scan not in HU

4. Kaggle Method

The cancer detection pipeline proposed by Liao et al. (2019) which won Kaggle DSB2017 challenge (referred as Kaggle method) were used on all preprocessed data. The pipeline consists of 2 steps, detection and classification.

First, a 3D CNN is built to detect suspicious nodules and predict bounding boxes. 3D small patches of size 128x128x128x1 are extracted from the scans and input to the next work and patches going beyond the range of lung scans is padded with value 170. The structure of detector network with a U-Net backbone to compare multi-scale information and an RPN output layer to generate proposals. Output nodule proposals includes coordinate of center of proposal, radius and confidence score of the proposal (Liao et al., 2019).

After detection nodules, five proposals with highest confidence scores are picked and the network in detection is reused for classification. Five 64D features were extracted from a fully connected layer to get cancer probability for each proposal. The probabilities for each proposal were integrated to get a final score for the scan (Liao et al., 2019).

The Kaggle pipeline was applied on all data downloaded without being removed in previous steps.

5. Distance-LSTM Method

As the Kaggle pipeline makes prediction on scan level, the DLSTM method proposed by Gao et al. (2019) is used to predict longitudinal scans like in NLST data. The method models the global time interval between previous time points to last scan as a global multiplicative function to input gate and forget gate (Gao et al., 2019).

In the experiments, the DLSTM network is trained as a post prepossessing network for features extracted from the Kaggle method (Liao et al.,2019). Five proposals with highest risks were selected with the pre-trained Kaggle model and a 64D feature was extracted for each proposal. For each scan, a 5x64 feature is extracted and fed into the DLSTM network. Learning rate is set to 0.01 initially and decreased at 50^{th}, 70^{th} and 80^{th} epoch. The maximum training epoch is 100 (Gao et al.,2019).

The DLSTM network was trained on all 7 subsets of data to explore what would happen if more negative data is fed into the model.

6. Imbalanced Data

Adjustments were made on the pipeline to deal with the problem of data imbalance. As the minority class is of more importance in this case, adjustments were made in data domain and cost space to overcome the consequences of data imbalance.

The first approach is oversampling positive class. The class torch.utils.data.WeightedRandomSampler in Pytorch (Paszke et al., 2019) is used to load data so that each batch contains approximately same number of positive and negative patients. The class distribution in each batch is not tuned to save time and resource, and to compare the results for each dataset fairly.

The second approach is giving weights to the loss function. Binary cross entropy is loss is used in the original D-LSTM network. The loss for each class is multiplied by a weight argument and the losses are averaged across observations for each mini batch (Paszke et al., 2019).

To achieve best performance, the weight argument request tuning for each data set. For the same reason as in oversampling, I kept the weight for each class proportional to 1/class size so that more weight is given to the minority class.

McNemar's Test (McNemar,1947) is used to evaluate if the models are significantly different because all models used the same test sets at decision threshold of 0.5. For each model, area under ROC curve(AUC), true positive rate (TP/(TP+FN), the percentage of positive cases correctly classified as belonging to the positive class), false positive rate(FP/(FP+TN), the percentage of negative cases misclassified as belonging to positive class), true negative rate (TN/(FP+TN), the percentage of negative cases correctly classified as belonging to the negative class), false negative rate (FN/(TP+FN), the percentage of positive cases misclassified as belonging to the negative class) for test set at decision threshold 0.5 are recorded.

Figure 6 below summarizes the entire pipeline.

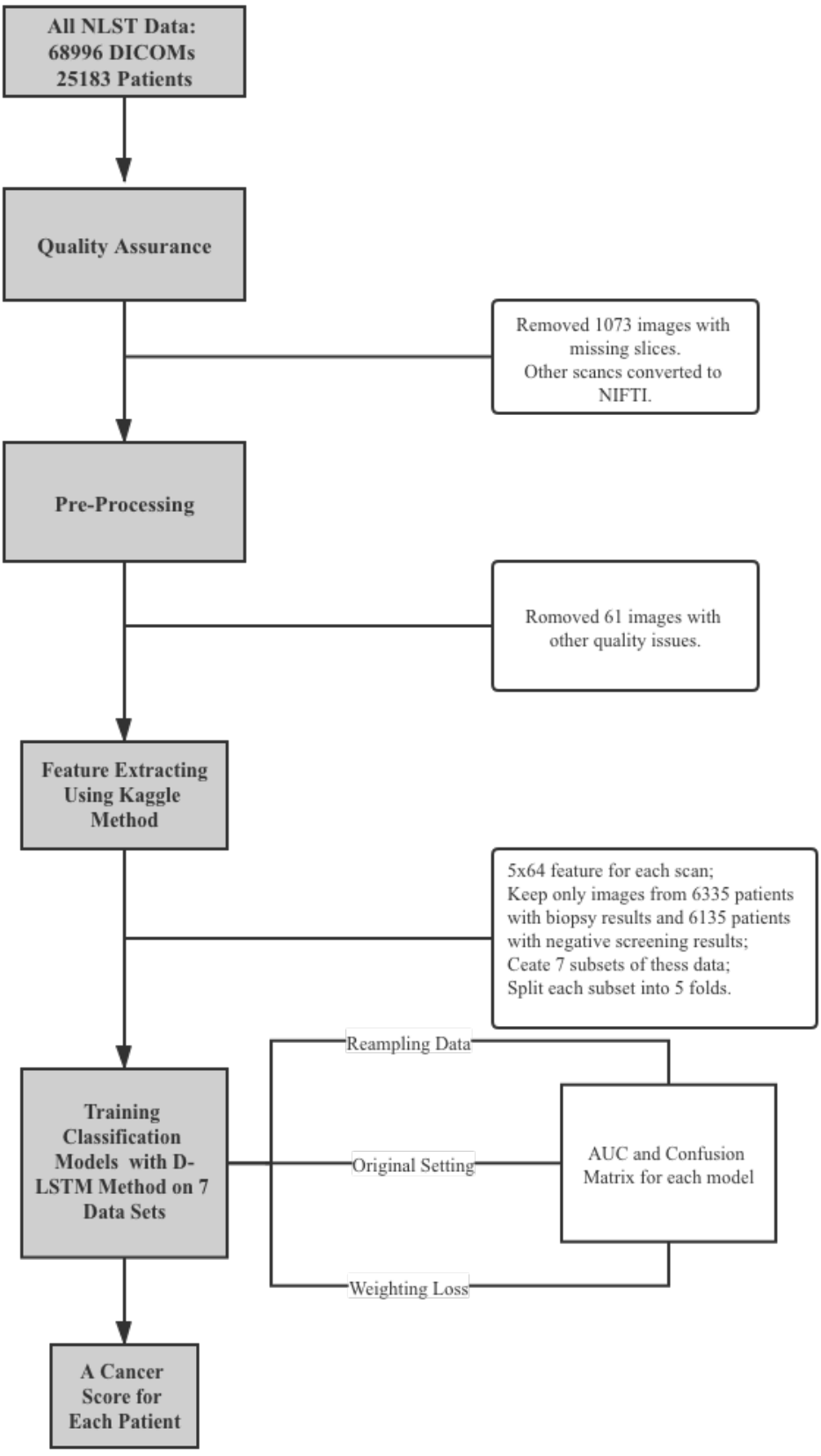

Figure 6 Flowchart of process for the pipeline

Chapter III. Results

1. Kaggle Pipeline

The Kaggle Pipeline (Liao et al.,2019) is run on all data passing QA and a score (referred as Kaggle score) is given to each image. Histograms of the Kaggle scores were drawn for both patients with or without biopsy. Figure 7 shows the histograms for patients with biopsy results and Figure 8 shows histograms for patients without biopsy results of different screening results. For a patient with multiple scans, the score of only the latest scan is recorded.

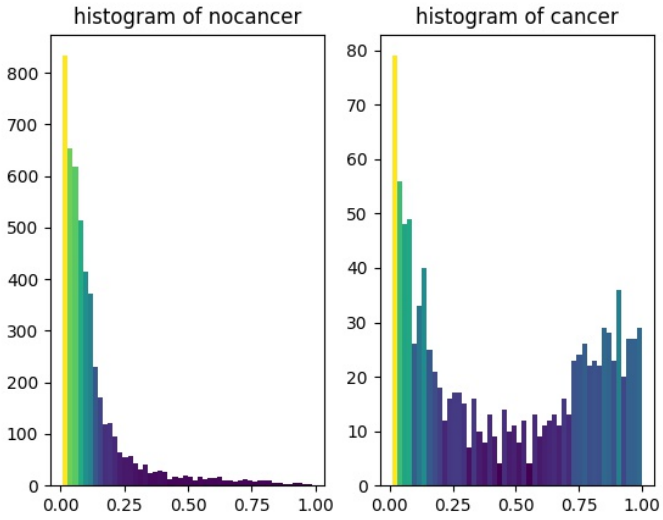

Figure 7 Kaggle score histogram for patients with biopsy results

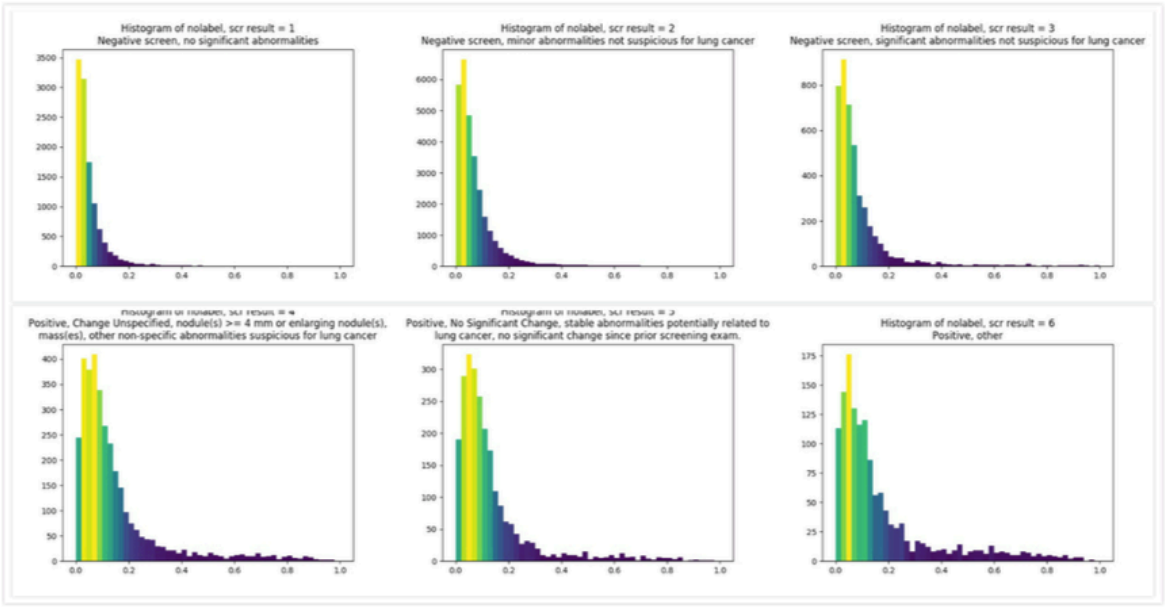

Figure 8 Kaggle score histogram for patients without biopsy results

2. Statistical Analysis of the Pipeline

McNemar's Test (McNemar,1947) is conducted between two classifiers too test again the null hypothesis that None of the two models performs better than the other. Thus, the alternative hypothesis is that the performances of the two models are not equal. Models trained with the same data with different approaches and models trained with the same approach on different data sets are compared. Table 2 to Table 4 showed the range of p-values for these tests.

Table 2 Results of McNemar's Tests on models trained with same data on using different approaches

**: $p < 0.01$;

*: $0.01 <= p < 0.05$;

NS: $P >= 0.05$

	original setting VS resampling	original setting VS weighted loss	resampling VS weighted loss
subset1: 1252	NS	NS	NS
subset2: 2383	**	**	NS
subset3: 3505	**	**	NS
subset4: 5894	**	**	NS
subset5: 7485	**	**	NS
subset6: 9014	**	**	NS
subset7: 12029	**	**	NS

Table 3 Results of McNemar's Tests on models trained with different training data on original D-LSTM setting

**: p < 0.01;

*: 0.01 <= p < 0.05;

NS: P>= 0.05

	subset1: 1252	subset2: 2383	subset3: 3505	subset4: 5984	subset5: 7485	subset6: 9014	subset7: 12029
subset1: 1252	NS	**	**	**	**	**	**
subset2: 2383	**	NS	**	**	**	**	**
subset3: 3505	**	**	NS	**	**	**	**
subset4: 5894	**	**	**	NS	**	NS	NS
subset5: 7485	**	**	**	**	NS	**	**
subset6: 9014	**	**	**	NS	**	NS	NS
subset7: 12029	**	**	**	NS	**	NS	NS

Table 4 Results of McNemar's Tests on models trained with different training data trained by resampling data

**: p < 0.01;

*: 0.01 <= p < 0.05;

NS: P>= 0.05

	subset1: 1252	subset2: 2383	subset3: 3505	subset4: 5984	subset5: 7485	subset6: 9014	subset7: 12029
subset1: 1252	NS	NS	NS	*	NS	NS	NS
subset2: 2383	NS	NS	NS	*	NS	NS	NS
subset3: 3505	NS	NS	NS	**	NS	NS	NS
subset4: 5894	*	*	**	NS	NS	**	**
subset5: 7485	NS	NS	NS	NS	NS	NS	*
subset6: 9014	NS	NS	NS	**	NS	NS	NS
subset7: 12029	NS	NS	NS	**	*	NS	NS

Table 5 Results of McNemar's Tests on models trained with different training data trained with weighted loss

**: p < 0.01;

*: 0.01 <= p < 0.05;

NS: P>= 0.05

	subset1: 1252	subset2: 2383	subset3: 3505	subset4: 5984	subset5: 7485	subset6: 9014	subset7: 12029
subset1: 1252	NS	NS	NS	NS	NS	NS	NS
subset2: 2383	NS	NS	NS	NS	NS	NS	NS
subset3: 3505	NS	NS	NS	NS	NS	NS	NS
subset4: 5894	NS	NS	NS	NS	NS	NS	NS
subset5: 7485	NS	NS	NS	NS	NS	NS	NS
subset6: 9014	NS	NS	NS	NS	NS	NS	NS
subset7: 12029	NS	NS	NS	NS	NS	NS	NS

3. DLSTM Pipeline on Imbalance Data

The original DLSTM method without resampling or weighting loss, original approach with only resampling data, original approach with only weighted loss were used on each of the 7 subsets of data. Not only the AUC is calculated, other 4 metrics, true positive rate(TPR), false positive rate(FPR), true negative rate(TNR), false negative rate are recorded(FNR).

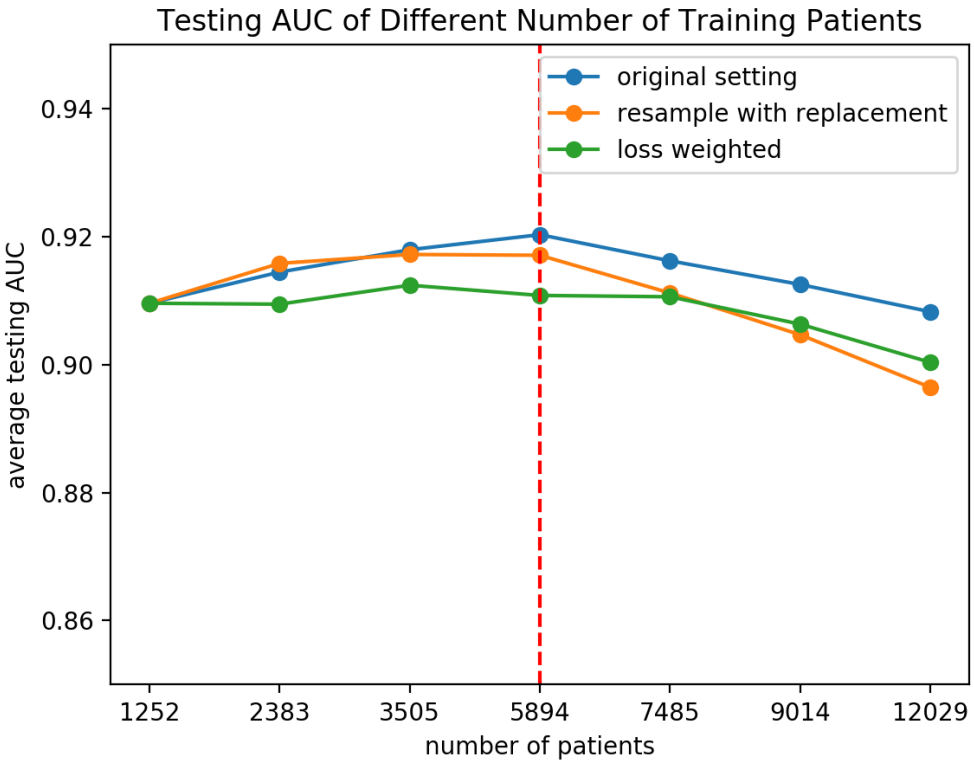

Figure 9 Testing AUC for DLSTM pipeline.

The x axis shows the number of patients in used to train the model. The y axis is the average test AUC for all five folds. The blue curve represents the AUC for the models trained without any approaches to deal with data imbalance. The orange curve shows the AUC for models trained by resampling training data with replacement. The green curve shows the AUC for models trained with weighted loss.

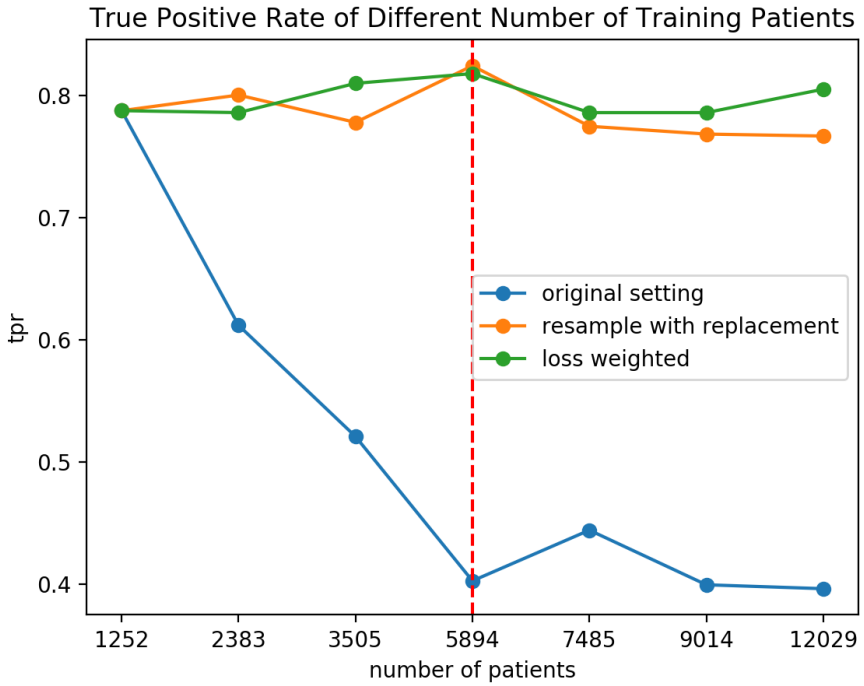

Figure 10 TPR for DLSTM pipeline

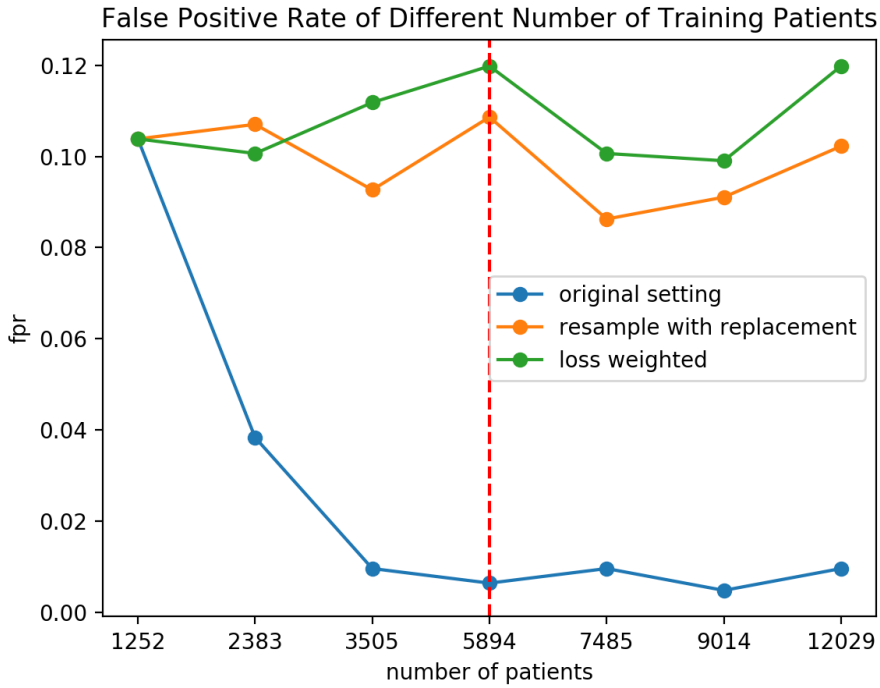

Figure 11 FPR for DLSTM pipeline

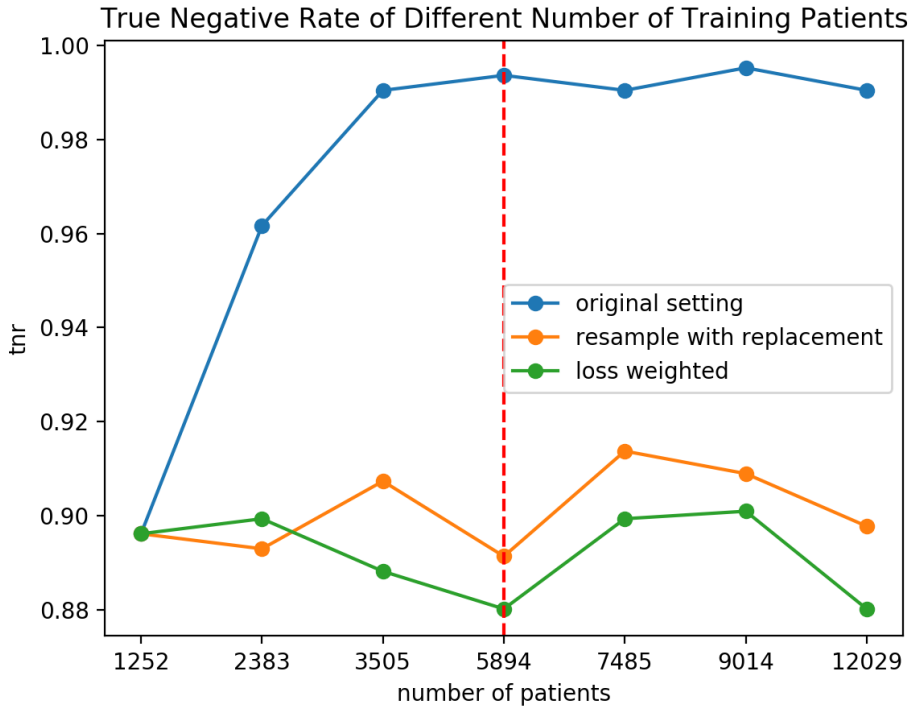

Figure 12 TNR for DLSTM pipeline

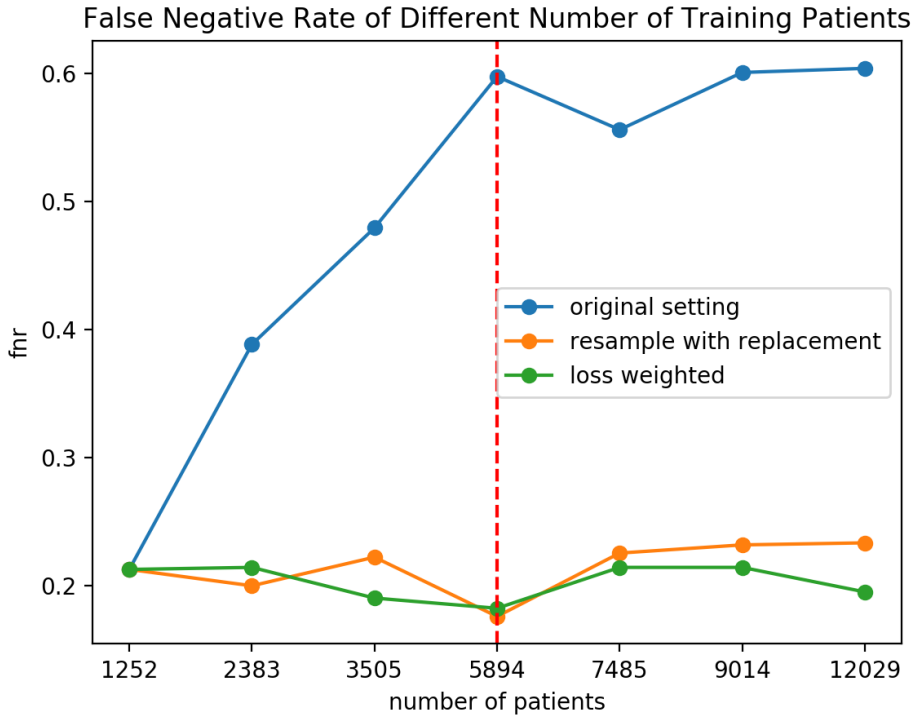

Figure 13 FNR for DLSTM Pipeline

Figure 9 to 13 shows the 4 confusion matrix metrics for the data set at decision threshold 0.5. The x axis is number of patients in used to train the model. The y axis is the average test metric for all five folds. The blue curve represents the metric for the models trained without any approaches to deal with data imbalance. The orange curve shows the metric for models trained by resampling training data with replacement. The green curve shows the metric for models trained with weighted loss.

Chapter IV. Discussion

1. Imbalanced Data

From Table 3, it is clear that adding more data to classifier makes a great difference in model performance. Table 4 and 5 show that approaches of resampling data and weighting the loss would eliminate most of the differences between models trained with different degrees of class imbalance. In table 2, both approaches would make the model performs significantly different from the original DLSTM model but the difference between two approaches are not great.

As we can see from the figures above, AUC first increased when adding negative data part B but decreased after adding negative data part C and none of the approaches supposed to overcome the consequences of imbalance seem to worsen the condition. From left to right on x-axis, the size of training data increases, which should benefit the classifier, but the degree of imbalance class distribution also increases which is supposed to damage the result. One of the reason for the curve to look like that is 5894 patients is the threshold that adding more negative data is beneficial the classifier, where large data could overcome the damage caused by imbalance.

The conclusion above is based on the assumption that the data quality of part C is as good as part A and B. However, this is not guaranteed because data in part C did not go through biopsy.

A reason that neither of the approaches seem to might be that the metric AUC did not reflect the change made by the approaches. To figure out whether the approaches worked, true positive rate, true negative rate, false positive rate and false negative rate at decision threshold 0.5 were looked at. As is seen in these curves, the two techniques have similar results of lowering false negative rate and keeping true positive rate from dropping. The results show that these two approaches did reduce the consequences of adding more negative sample.

However, these approaches have limitations. Both approaches introduced additional computational costs and oversampling positive class would be overwhelming in cases of very large scale training data. Tuning class distribution in each batch and weight in the loss are required if a better performance is looked at.

2. AUC as a Classification Metric

The ROC curve is the curve that plots the true positive rates against the false positive rate as decision threshold varies. It is a widely used way to visualize the classifier's performance to select an operation point or decision threshold. The area under the ROC curve is a feature of the curve and is a widely-used metric in machine learning domain because of advantages like decision threshold independent; and it is invariant to a priori class probabilities (Bradley,1996). However, when dealing with highly skewed data, AUC does not provide all information， as it is hard for the ROC curve to capture the effect large number of negative examples has on the algorithm's performance.

Weiss et al. (2003) conducted researches to explore relationship between class distribution on training set and indicates that AUC preformed relatively better on a relatively balanced training set but the optimal class distribution for AUC is not necessarily 1:1 and hard to determine (Sun et al.,2009). So the reason that subset 4

has highest AUC might be that the class distribution in this subset is closest to the optimal class distribution for this classifier.

Chapter V. Conclusion

In this thesis, we proposed a deep learning pipeline with approaches to deal with class imbalance in the NLST data set. CT scans from NLST data set are downloaded and quality assured (Gao, 2019) before applying a pre-processing procedure (Liao et al., 2019) to segment the lung. The Kaggle method (Liao et al.,2019) is used to extract features on the scan level and the D-LSTM method (Gao et al., 2019) predicts the cancer score for each patient. The pipeline is applied on 7 subsets of the data set with different degrees of imbalance in class distribution and the performance of the pipeline reduces as more than 5268 patients with negative class used in training. Two approaches, resampling data and weighting the loss are applied to address the problem of imbalanced class distribution in the data set. Though the effect of these approaches is hard to observe in AUC, the confusion matrix shows the influence.

References

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Centers for Disease Control and Prevention. (2019). *Basic Information About Lung Cancer*. https://www.cdc.gov/cancer/lung/basic_info/

Donahue, J., Lisa, A. H., Hendricks, A., Guadarrama, S., Rohrbach, M., Venugopala, S., Saenko, K., & Darrell, T. (2015). *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*.

Ertekin, S., Huang, J., Bottou, L., & Lee Giles, C. (2007). Learning on the border: active learning in imbalanced data classification. *International Conference on Information and Knowledge Management, Proceedings*, 127–136. https://doi.org/10.1145/1321440.1321461

Farag, A., Elhabian, S., Graham, J., Farag, A., & Falk, R. (2010). Toward precise pulmonary nodule descriptors for nodule type classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6363 LNCS*(PART 3), 626–633. https://doi.org/10.1007/978-3-642-15711-0_78

Gao, R. (2019). *GitHub - RiqiangGao/QA_tool*. https://github.com/RiqiangGao/QA_tool

Gao, R., Huo, Y., Bao, S., Tang, Y., Antic, S. L., Epstein, E. S., Balar, A. B., Deppen, S., Paulson, A. B., Sandler, K. L., Massion, P. P., & Landman, B. A. (2019). Distanced LSTM: Time-Distanced Gates in Long Short-Term Memory Models for Lung Cancer Detection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11861 LNCS*, 310–318. https://doi.org/10.1007/978-3-030-32692-0_36

Gatsonis, C. A., Aberle, D. R., Berg, C. D., Black, W. C., Church, T. R., Fagerstrom, R. M., Galen, B., Gareen, I. F., Goldin, J., Gohagan, J. K., Hillman, B., Jaffe, C., Kramer, B. S., Lynch, D., Marcus, P. M., Schnall, M., Sullivan, D. C., Sullivan, D., Zylak, C., … Dyer, S. (2011). The national lung screening trial: Overview and study design. *Radiology*, *258*(1), 243–253. https://doi.org/10.1148/radiol.10091808

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*.

Hua, K. L., Hsu, C. H., Hidayati, S. C., Cheng, W. H., & Chen, Y. J. (2015). Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and Therapy*, *8*, 2015–2022. https://doi.org/10.2147/OTT.S80733

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *INTELLIGENT DATA ANALYSIS*, 429--449. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.711.8214

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. http://code.google.com/p/cuda-convnet/

Lin, P. L., Huang, P. W., Lee, C. H., & Wu, M. T. (2013). Automatic classification for solitary pulmonary nodule in CT image by fractal analysis based on fractional Brownian motion model. *Pattern Reco*gnition

Liao, F., Liang, M., Li, Z., Hu, X., & Song, S. (2019). Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-OR Network. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(11), 3484–3495. https://doi.org/10.1109/TNNLS.2019.2892409

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

National Cancer Institute. (2014). *Nlst Participant : Data Dictionary*. https://cdas.cancer.gov/datasets/nlst/

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. http://arxiv.org/abs/1912.01703

Ren, S., He, K., Girshick, R., & Sun, J. (2015). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. https://github.com/

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015* (pp. 234–241). Springer International Publishing.

Santeramo, R., Withey, S., & Montana, G. (2018). Longitudinal detection of radiological abnormalities with time-modulated lstm. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11045 LNCS*, 326–333. https://doi.org/10.1007/978-3-030-00889-5_37

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). CLASSIFICATION OF IMBALANCED DATA: A REVIEW. In *International Journal of Pattern Recognition and Artificial Intelligence* (Vol. 23, Issue 4). www.worldscientific.com

Tekade, R., & Rajeswari, K. (2018, July 2). Lung Cancer Detection and Classification Using Deep Learning. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*. https://doi.org/10.1109/ICCUBEA.2018.8697352

Weiss, G. M. (2003). Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. In *Journal of Artificial Intelligence Research* (Vol. 19).

Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., Mak, R. H., & Aerts, H. J. W. L. (2019). Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical Cancer Research*, *25*(11), 3266–3275. https://doi.org/10.1158/1078-0432.CCR-18-2495