Pancreatic Cancer Detection in Whole Slide Images Using Noisy Label Annotations

Han Le[1], Dimitris Samaras[1], Tahsin Kurc[2], Rajarsi Gupta[2,3], Kenneth Shroyer[3], and Joel Saltz[2]

[1] Department of Computer Science, Stony Brook University, NY, USA
hdle@cs.stonybrook.edu
[2] Department of Biomedical Informatics, Stony Brook Medicine, NY, USA
[3] Department of Pathology, Stony Brook University Hospital, NY, USA

Abstract. We propose an approach to accurately predict regions of pancreatic cancer in whole-slide images (WSIs) by leveraging a relatively large, but noisy, dataset. We employ a noisy label classification (NLC) method (called the *NLC model*) that utilizes a small set of clean training samples and assigns the appropriate weights to training samples to deal with sample noise. The weights are assigned online so that the network loss approximates the loss for the clean samples. This method results in a 9.7% performance improvement over the baseline non-NLC method (the *Baseline-Noisy model*). We use both methods in an ensemble setup to generate labels for a large training dataset to train a classifier. This classifier outperforms a classifier trained with manually annotated data by 2.94%-3.74% in terms of AUC for testing patches in WSIs.

Keywords: Pancreas · Pancreatic Cancer · Whole Slide Image

1 Introduction

We target the problem of automatically detecting regions of pancreatic cancer in WSIs. Segmentation of cancer regions is a fundamental operation in digital pathology image analysis [7,9]. Pancreatic cancer segmentation is particularly important since it can be utilized to characterize immune responses that have been shown to affect survival outcomes and treatment response in pancreatic cancer patients [3]. A challenge to using deep learning in this task is the difficulty of generating detailed and large training datasets. In pancreatic cancer, malignant cells are typically arranged in irregularly shaped and poorly formed glands that infiltrate surrounding tissues. There is a wide spectrum of heterogeneity in the appearance of tumor cells, combined with the fact that a majority of the cancer region is comprised of non-cancer stromal and immune cells [5]. This morphologic complexity significantly limits highly detailed and fine-grained annotations because the annotation process is too laborious and time consuming.

In order to address this challenge, we propose an approach that formulates the cancer region detection problem as a noisy label classification problem. Annotated cancer regions are considered noisy due to the lack of specific delineation and

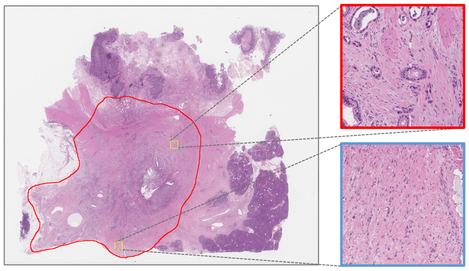

Fig. 1: A WSI with human annotation. The red box indicates a true positive patch; the blue box indicates a false positive patch extracted from the tumor annotated regions.(Please zoom in on a digital device).

labeling of the cancer and non-cancer components within the cancer region. Our approach uses a small amount of high-quality ground truth data (*clean data*), a larger volume of *noisy data*, and an ensemble of deep learning models to generate a large training dataset for a deep learning classifier.

Studies [1] have shown that the performance of a deep learning network can be adversely affected when it is trained with a noisy dataset. Numerous methods have been proposed to cope with noisy label classification for natural images [11,6,10,12,14]. Ren et al.[11] propose a technique to assign weights to training samples by using an additional clean validation set. Their intuition is to apply smaller weights to noisy samples and increase the weights of clean training samples to improve the gradient update. There is relatively limited work on the development and application of noisy label classification methods in medical imaging data [2,4,13]. Dgani et al. [4] model label noise as a part of the deep learning network to recover true labels of noisy samples for the task of classifying breast micro-calcifications in multi-view mammograms.

We make the following contributions: **(1)** Our approach is the first method for detection of pancreatic cancer regions in WSIs by using a large, but noisy, training dataset combined with the noisy label classification (NLC) technique of Ren et al. [11]; **(2)** We propose a pipeline to generate a large training dataset from moderately-sized and noisy annotated data; **(3)** Using this pipeline, we have generated a training dataset of 353,000 patches from 190 WSIs in The Cancer Genome Atlas (TCGA) repository. Our experiments show that a classifier trained with this larger noisy dataset outperforms a classifier trained with fewer clean ground truth data only. Our approach provides a viable mechanism for generating a large training dataset from moderately-sized and noisy annotated data. The training dataset and our prediction results on 190 TCGA WSIs are publicly available for use in other imaging studies [1].

2 Noisy Label Classification Approach

We propose a patch-based classification model to detect and classify cancer and non-cancer regions. This method partitions a WSI into tiles (or patches) of $P \times P$ pixels and predicts a class label for each tile. The classification model is trained

[1] https://github.com/SBU-BMI/quip_paad_cancer_detection

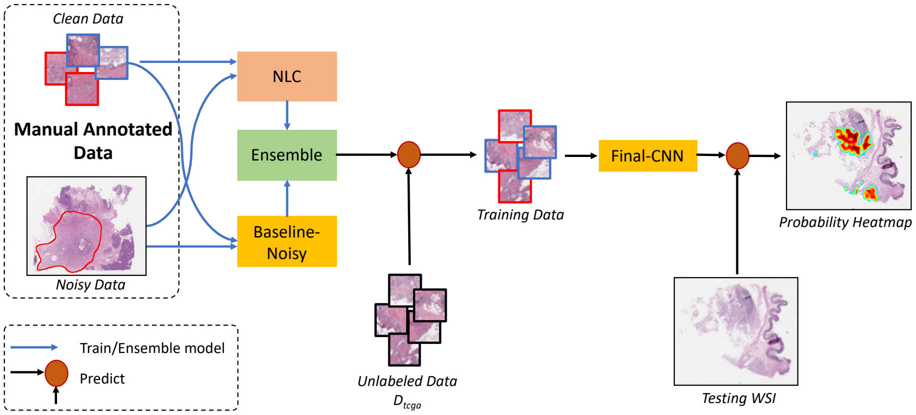

Fig. 2: Proposed training data generation pipeline. The NLC and the Baseline-Noisy models are trained with limited manually annotated data. The Ensemble model generates labels for patches extracted from 190 TCGA WSIs, D_{tcga}.

with a set of tiles from cancer and non-cancer regions. In this section, we describe the process of generating a set of training tiles from a relatively small amount of high-quality annotated data (*clean data*) and a larger set of weakly annotated data (*noisy data*). The overall framework is illustrated in Figure 2.

2.1 Tumor Region Annotation and Tile Extraction

Cancer and non-cancer regions in WSIs are manually segmented by pathologists. Pathologists are normally asked to carefully draw accurate contours around all of the cancer and non-cancer regions after histologic examination at intermediate and high-magnification. As a result, they often have to spend hours to generate high-quality and error-free ground-truth training datasets. In this work, the pathologist was asked to mark the boundaries of the cancer region in each WSI at low- to intermediate-magnification. This reduced manual annotation time but introduced noise because non-cancer components within the cancer regions could not be delineated at low-magnification. Note that the regions that lie outside of the annotated cancer regions were guaranteed to be non-cancer regions and did not introduce noise. Figure 1 shows an example annotation to identify a cancer region (within the red lines) with true positive and false positive patches extracted from the annotated region.

After manual annotation, image tiles were extracted from the annotated cancer and non-cancer regions. To determine the best tile size for classification, several tiles from different annotated regions were presented to the pathologist. In our experiments, tiles were extracted at 1,000×1,000 pixels (equivalent to $500\times500\mu m$) at 20x magnification and then resized to 224×224 pixels. A patch was labeled positive if at least 50% of its area intersects with a cancer region. Negative patches were determined by the patch being *fully* from outside the area of the cancer regions. All other tiles were removed from the dataset. In order to generate a set of high-quality tile annotations (i.e., clean annotation data), a set of tiles from the cancer regions was selected randomly from the WSIs and presented to the pathologist for classification as cancer, non-cancer, or undecided. If the pathologist could not classify a tile, the tile was labeled as *undecided* and removed from the clean annotation dataset.

2.2 NLC Model: Noisy Label Classification Model

The manual annotation process ensures that tiles outside of cancer regions are true negative samples (non-cancer tiles). Tiles extracted from cancer regions are labeled positive, but this set contains both true and false positive samples (noisy training samples, i.e., non-cancer tiles that represent immune and stromal cells within the cancer region). To address this issue, we have adapted the noisy label classification method proposed by Ren et al. [11] with a modification on how to construct the subset of clean samples. Instead of selecting random clean samples from all regions in the WSIs, we choose samples in cancer regions only. In our experiments, we generated 100 clean samples per class via this strategy.

Let (x, y) be a (tile, label) tuple, and let D_n and D_c be the set of noisy and clean samples, respectively. The network parameters, θ, can be computed by minimizing the training loss over the training data: $\min_\theta \sum_{d_i \in D_n} w_i \mathcal{L}_i(d_i, \theta)$, where w_i is the importance weight of sample x_i and $\mathcal{L}_i$ is the loss function associated with x_i. The weights $\{w_i\}_{i=1}^N$ are treated as hyper-parameters. They are computed by minimizing the loss over the clean dataset: $\min_{w \geq 0} \sum_{d_i \in D_c} \mathcal{L}_i(d_i, \theta^*(w))$. For computational efficiency, the update of the weights is computed in an online manner for each batch of training samples.

2.3 Ensemble Model of NLC and Baseline-Noisy models

The Baseline-Noisy model is the same CNN architecture used for the NLC model, but it is trained with the noisy and clean samples without NLC. In our experiments (see Section 3.2), we observed that the NLC model is better than the Baseline-Noisy model at classifying patches in cancer regions, whereas the Baseline-Noisy model is better at classifying patches in non-cancer regions. To utilize the strengths of both models, an Ensemble Model computes the final prediction for a tile by averaging the prediction probabilities from the NLC and Baseline-Noisy models. We used the Ensemble model to generate labels for a large training dataset for the Final-CNN model in Figure 2.

3 Experimental Evaluation

3.1 Experimental Setup

Datasets. We used high-resolution WSIs of pancreatic adenocarcinoma (PAAD) scanned at 40x magnification (approximately 0.25 microns per pixel) from SEER[2]

[2] https://seer.cancer.gov/

Purpose	ID	#WSIs	#Positive	#Negative	#Total
Noisy Training	D_n	50	21,805	47,640	69,445
Clean set	D_c	14	100	100	200
Unlabeled Data	D_{tcga}	190	-	-	353,000
Testing	T_{seer}	14	1,700	1,700	3,400
Testing	T_{seer2}	14	850	2,550	3,400
Testing	T_{tcga}	190	1,051	2,003	3,054

Table 1: Dataset Statistics. D_n and D_c were used in training the NLC and the Baseline-Noisy models. The Unlabeled Data, D_{tcga}, was used to generate the training set for the Final-CNN model. T_{seer}, T_{seer2}, and T_{tcga} are test sets.

(64 WSIs) and TCGA[3] (190 diagnostic WSIs). A pathologist manually annotated cancer and non-cancer regions in 50 WSIs that were randomly selected from the SEER dataset to generate the *noisy* annotation data. This process yielded a total of 69,445 tiles; 21,805 positive/cancer tiles and 47,640 negative/non-cancer tiles. We generated a manually annotated *clean* dataset of 100 positive and 100 negative tiles from the remaining 14 SEER images. The noisy and clean data comprised the "Manually Annotated Dataset" in Figure 2. This dataset is used to train the NLC and the Baseline-Noisy models. We randomly extracted 353,000 tiles from 190 TCGA WSIs. This dataset, D_{tcga}, was used as part of the training dataset for the Final-CNN in Figure 2.

We created three test datasets: T_{seer}, T_{seer2}, and T_{tcga}. T_{seer} consists of 1700 positive tiles and 1700 negative tiles from 14 SEER WSIs. We initially extracted a total of 3,960 patches from cancer regions in these images for pathologist review and classification as positive or negative. The pathologist labeled 1,829 patches (46.2%) as positive, 1,833 patches (46.3%) as negative, and 298 patches (7.5%) as undecided. From the clean samples (1,829 positive and 1,833 negative patches), we randomly selected 3,400 patches to create T_{seer} and 200 patches for D_c. The second set, T_{seer2}, contains a subset of 850 negative and 850 positive samples from T_{seer}, and 1,700 negative samples randomly extracted in the non-tumor regions from 14 SEER WSIs. The third test dataset, T_{tcga}, is made up of 3,054 patches from 190 TCGA WSIs. Table 1 shows the number of patches extracted for the training and test datasets.

Baseline-Clean: Baseline Model on Clean Data. To evaluate the contribution of the clean dataset to the performance of the network, we trained the Baseline-Clean model by using the clean set only. The model is optimized by minimizing the following training loss: $\min_\theta \sum_{d_i \in D_c} \mathcal{L}_i(d_i, \theta)$.

Implementation. We used the Preact-Resnet-34 architecture [8] for all of the models: NLC, Baseline-Noisy, Baseline-Clean and the Final-CNN model. Preact-Resnet is a common CNN that is used in many medical imaging applications.

[3] https://portal.gdc.cancer.gov/

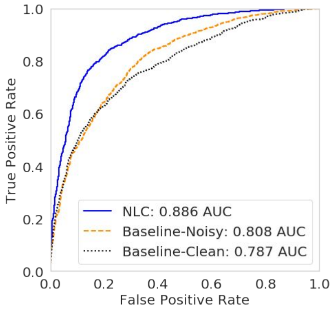

Fig. 3: Area Under the Curve (AUC) of the NLC, the Baseline-Noisy, and the Baseline-Clean evaluated on T_{seer}.

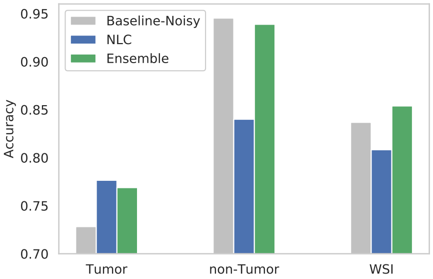

Fig. 4: Accuracy of the proposed models in tumor and non-tumor regions, and in WSIs as evaluated on T_{seer2}.

We trained the NLC, Baseline-Noisy, and Baseline-Clean models with the same training process starting with an initial learning rate of 0.001, a momentum of 0.9 and a weight decay of 0.0001. The learning rate was decreased by a factor of 10 at the 100^{th} and 125^{th} epochs. The network weights of the models were initialized randomly and the models were trained until convergence (which took 150 epochs). We used the cross entropy loss function to compute the loss for each training sample: $\mathcal{L}_i = -y_i log(\hat{y}_i) - (1 - y_i) log(1 - \hat{y}_i)$, where $y_i = 1$ if the sample is positive and $y_i = 0$ otherwise. $\hat{y}_i$ is the prediction score of the network after the sigmoid function is applied.

We trained the two Final-CNN classification models with D_{tgca}: one model with labels generated by the Baseline-Noisy model and the other with labels by the Ensemble model. We used the same training procedure as for the Baseline-Noisy model apart from starting with a learning rate of 0.01. We decreased the learning rate by 10 at the 10^{th} epoch. We initialized the CNNs with the weights of the Baseline-Noisy model and trained for 15 epochs.

3.2 Results

We used the area under the ROC (Receiver Operating Characteristic) curve, or simply AUC, as our performance metric. Figure 3 shows the AUC values of the NLC, Baseline-Noisy, and Baseline-Clean models tested against T_{seer}. The Baseline-Clean model shows the worst performance with an AUC of 0.787. We attribute this to the fact that it was trained with a small training dataset. The NLC model outperforms the Baseline-Noisy model by 9.7% in the tumor regions, where the performance improvement is due to the use of the NLC method.

The experiments show that the NLC model generally performs well in tumor regions, whereas the Baseline-Noisy model performs better than the NLC model in non-tumor regions. We believe that this is because of the training process. The NLC model is regularized by clean data (however limited) that guides the network to better distinguish between positive and negative samples

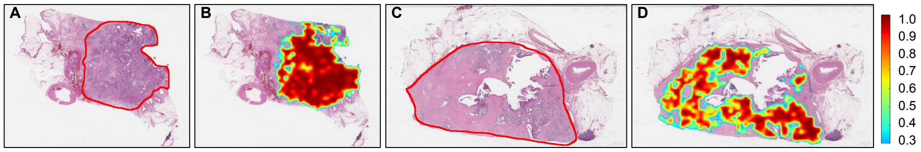

Fig. 5: Prediction probability maps with example WSIs generated using the Final-CNN model trained with D_{tcga} with labels generated by the Ensemble model. Images A, C show the ground truth cancer regions (red lines) segmented by pathologists. Images B, D display the probability maps as heatmaps on two unseen testing SEER WSIs.

in tumor regions. In contrast, there is no clear guidance from tumor regions for the Baseline-Noisy model during training. The performance suffered in tumor regions because negative and positive patches inside a tumor region are not explicitly distinguished from one another. Because the number of negative samples is larger than the number of positive samples, the Baseline-Noisy model likely learned to better detect negative samples in non-tumor regions. This observation led to the implementation of the Ensemble as shown in Figure 2. Figure 4 shows the performance of the NLC, Baseline-Noisy, and Ensemble models with test patches in tumor regions, test patches in non-tumor regions, and all of the test patches in the test dataset T_{seer2}.

#	Trainset	Label Source	Testset	AUC
1	$D_n \cup D_c$	Human	T_{tcga}	0.829
2	D_{tcga}	Baseline-Noisy	T_{tcga}	0.832
3	D_{tcga}	Ensemble	T_{tcga}	**0.860**
4	$D_n \cup D_c$	Human	T_{seer2}	0.917
5	D_{tcga}	Baseline-Noisy	T_{seer2}	0.928
6	D_{tcga}	Ensemble	T_{seer2}	**0.944**

Table 2: AUC values of CNNs trained with SEER data with human annotation (#1 and #4), and with D_{tcga} with labels generated by the Baseline-Noisy model (#2 and #5), or by the Ensemble model (#3 and #6). Models are evaluated on 2 test sets: T_{tcga} (#1 – #3) and T_{seer2} (#4 – #6).

To further evaluate the proposed methods, we generated 2 sets of (tile, label) pairs for the 353,000 patches extracted from the 190 TCGA WSIs: one with labels generated by the Baseline-Noisy model and the other with labels generated by the Ensemble model. As shown in Table 2, the CNNs trained with labels generated by the Ensemble model (#3 and #6 in Table 2) outperform the CNNs trained with manually generated labels (#1 and #4) by 3.74% on T_{tcga} and by 2.94% on T_{seer2} testset in terms of AUC. They also slightly outperform the CNNs trained with labels generated by the Baseline-Noisy model. Figure 5 shows probability heatmaps of two SEER WSIs classified by the Final-CNN model trained with D_{tcga} and patch labels generated by the Ensemble model.

4 Conclusions

Generating large training sets for pancreatic cancer region detection is very challenging due to the complexity and heterogeneity of tumor regions. Our approach involves collecting a relatively small set of clean data in cancer regions and applying a technique for assigning weights to training samples. Our results show that this approach can generate large training sets from noisy datasets. Given the high cost of generating ground truth data, we believe that methods which work with weakly-labeled, noisy data will be crucial to the broader adoption of deep learning in digital pathology. We plan to investigate additional sampling and noise reduction techniques to improve the quality of weakly-labeled training datasets and cancer region detection accuracy.

Acknowledgement. This work was supported in part by 1U24CA180924-01A1, 3U24CA215109-02, and 1UG3CA225021-01 from the NCI, R01LM011119-01 and R01LM009239 from the NLM. This work leveraged resources from XSEDE, which is supported by NSF ACI-1548562 grant, including the Bridges system (NSF ACI-1445606) at the Pittsburgh Supercomputing Center. We thank Vu Nguyen for meaningful discussions.

References

1. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 233–242. JMLR. org (2017)
2. Azizi, S., Yan, P., Tahmasebi, A., et al.: Learning from noisy label statistics: Detecting high grade prostate cancer in ultrasound guided biopsy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 21–29. Springer (2018)
3. Balachandran, V.P., Łuksza, M., Zhao, J.N., Makarov, V., Moral, J.A., Remark, R., Herbst, B., Askan, G., Bhanot, U., Senbabaoglu, Y., et al.: Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. Nature **551**(7681), 512 (2017)
4. Dgani, Y., Greenspan, H., Goldberger, J.: Training a neural network based on unreliable human annotation of medical images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 39–42. IEEE (2018)
5. Feig, C., Gopinathan, A., Neesse, A., Chan, D.S., Cook, N., Tuveson, D.A.: The pancreas cancer microenvironment. Clin. Cancer Res., 18 (2012), pp. 4266-4276
6. Ghosh, A., Kumar, H., Sastry, P.: Robust loss functions under label noise for deep neural networks. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
7. Golatkar, A., Anand, D., Sethi, A.: Classification of breast cancer histology using deep learning. In: 15th International Conference on Image Analysis and Recognition. pp. 837–844. Springer (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016)

9. Kong, B., Sun, S., Wang, X., Song, Q., Zhang, S.: Invasive cancer detection utilizing compressed convolutional neural network and transfer learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 156–164. Springer (2018)
10. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1944–1952 (2017)
11. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. Proceedings of the 35th International Conference on Machine Learning (2018)
12. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5552–5560 (2018)
13. Xue, C., Dou, Q., Shi, X., Chen, H., Heng, P.A.: Robust learning at noisy labeled medical images: Applied to skin lesion classification. The IEEE International Symposium on Biomedical Imaging (2019)
14. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Advances in neural information processing systems. pp. 8778–8788 (2018)