

Tendentious Noise-rectifying Framework for Pathological HCC Grading — Supplementary Materials —

Xiaotian Yu*¹
yuxiaotian@zju.edu.cn

Zunlei Feng*¹
zunleifeng@zju.edu.cn

Mingli Song, Yuexuan Wang¹
brooksong, amywang@zju.edu.cn

Xiuming Zhang¹³
1508056@zju.edu.cn

Thomas Li²
liktt@hku.hk

¹ Zhejiang University
Zhejiang, China

² University of Hong Kong,
Hong Kong, China

³ The First Affiliated Hospital,
College of Medicine,
Zhejiang University,
Zhejiang, China

A. Model Architectures

The proposed Tendentious Noise-rectifying Framework (TNF) is designed for pathological HCC grading with noisy annotations. The detail of TNF architecture is provided for better illustrating the framework.

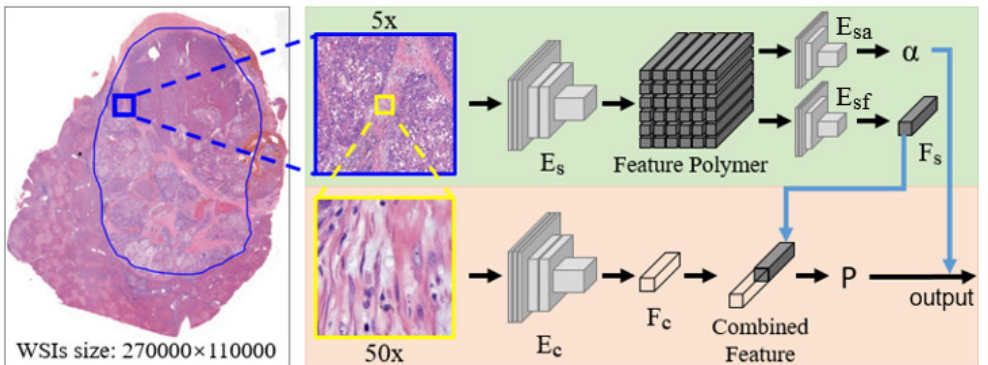


Figure 1: The flow diagram of Tendentious Noise-rectifying Framework (TNF). The encoder E_c is adopted in the main branch (the bottom one) to extract cellular features, and the encoder E_s is adopted in the auxiliary branch (the top one) to generate corresponding feature polymer representing structural features.

As described in Section 3 of the original paper, TNF contains two branches in charge of classification and rectification. The Resnet18 [10] pretrained on ImageNet [11] is adopted as the encoder E_c of the main branch to extract cellular features F_c for HCC grading. For the auxiliary branch, E_s is the encoder of the pretrained AE on the training dataset to generate the feature polymer. The E_{sa} and E_{sf} is followed to output the α and structural features F_s , respectively. Here, the size of 50x input patch is $448 \times 448 \times 3$, and the dimension of output features F_c and F_s is 512×1 . These two features will be contacted to combine cellular and structural information, and inputted into the classifier to get a 5 dimension vector, representing the prediction confidence on five grades. The architecture of E_s , E_{sf} , and E_{sa} is shown in Table 1. Batch Normalization is adopted after each convolutional layer.

Table 1: The network architectures for encoder E_s , E_{sf} , and E_{sa} . Batch Normalization is adopted after each convolutional layer.

Encoder E_s in the auxiliary branch
ReLU, kernel_size= 4×4 , output_channel=16, conv, stride=2, padding=1
ReLU, kernel_size= 4×4 , output_channel=32, conv, stride=2, padding=1
ReLU, kernel_size= 4×4 , output_channel=32, conv, stride=2, padding=1
ReLU, kernel_size= 4×4 , output_channel=64, conv, stride=2, padding=1
ReLU, kernel_size= 4×4 , output_channel=64, conv, stride=2, padding=1
ReLU, kernel_size= 4×4 , output_channel=64, conv, stride=2, padding=1
Encoder E_{sf} in the auxiliary branch
ReLU, kernel_size= 3×3 , output_channel=128, conv, stride=2, padding=3
ReLU, kernel_size= 3×3 , output_channel=256, conv, stride=2, padding=1
ReLU, kernel_size= 3×3 , output_channel=256, conv, stride=1, padding=1
ReLU, kernel_size= 3×3 , output_channel=512, conv, stride=2, padding=1
ReLU, kernel_size= 3×3 , output_channel=512, conv, stride=1, padding=1
Average Pooling, ourput_size= 1×1
Encoder E_{sa} in the auxiliary branch
ReLU, kernel_size= 3×3 , output_channel=128, conv, stride=2, padding=3
ReLU, kernel_size= 3×3 , output_channel=256, conv, stride=2, padding=1
ReLU, kernel_size= 3×3 , output_channel=256, conv, stride=1, padding=1
ReLU, kernel_size= 3×3 , output_channel=512, conv, stride=2, padding=1
ReLU, kernel_size= 3×3 , output_channel=512, conv, stride=1, padding=1
Average Pooling, ourput_size= 1×1
ReLU, output_channel=128, dense
Sigmoid, output_channel=1, dense

B. Qualitative Comparison

The results of comparative experiment are shown in the original paper, containing the metrics of accuracy, sensitivity, and specificity. Additionally, the Receiver Operating Characteristic (ROC) curve is widely used to analyze diagnostic results in medical tasks [6, 7]. The ROC curve can better evaluate the robustness with different thresholds, and helps to find the best threshold to classification.

The ROC curve graphs of different methods are shown in Fig. 2. Here the best model of each method is chosen. It can be seen that TNF achieves largest AUC (0.9670) among all existing methods, which proves the high robustness of our framework. It's also obvious that, although most methods like GCE, SL, and LSR get improvement in the accuracy of classification, the results on AUC are not as good as CE (0.9564). It indicates these anti-noise methods are not robust enough to HCC dataset, which makes them inapplicable in this task. In summary, the proposed TNF has comprehensive ability in HCC grading, and that is the reason why TNF is superior to existing methods and achieves state-of-the-art performance.

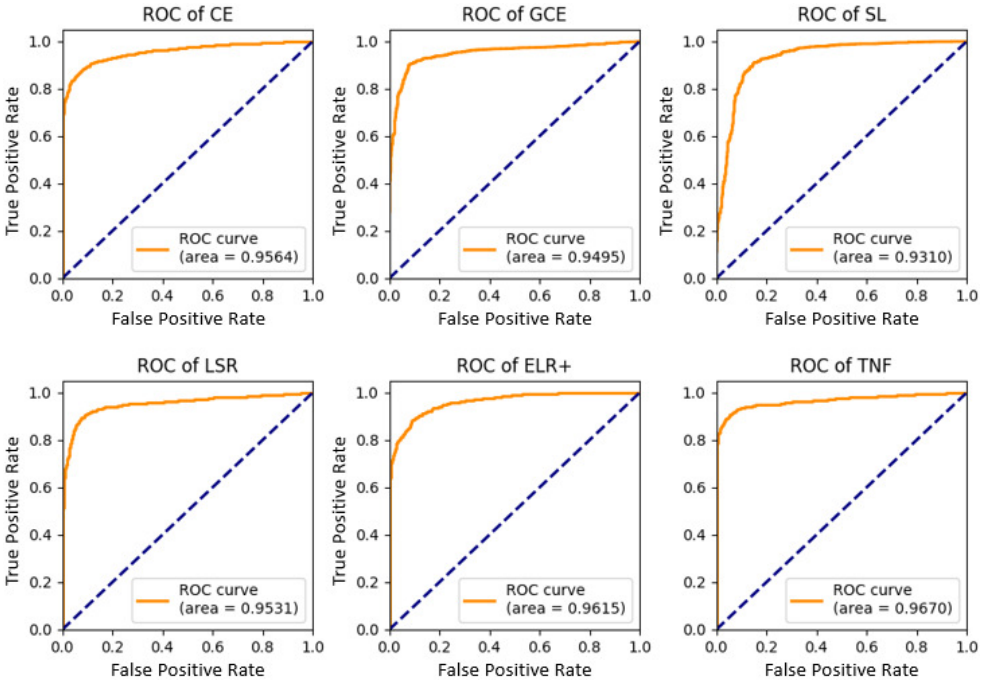


Figure 2: The ROC curve graphs of different methods on the test dataset. Here all the cancerous grades are considered as the same class to evaluate the correlation between true positive rate and false positive rate. The best model of each method is chosen to get the results. The Area Under Curve (AUC) is shown in the bottom right.

To further demonstrate the effectiveness of the proposed method, additional experiments on CAMELYON16 dataset [8] are conducted, which consists of pathological slides of breast cancers with binary labels. CAMELYON16 dataset contains 111 tumor slides and 160 normal slides, from which we crop 36,000 patches for each category as the training and validation set. And 20,000 patches are also cropped from 121 testing slides to evaluate the model

Table 2: The classification results of different methods on CAMELYON16. The metrics contain accuracy (*Acc.*), sensitivity (*Sen.*), specificity (*Spe.*). All these results averaged over ten experiments and the best results are marked in **bold** (All scores are in %).

	CE	CM [■]	GCE [■]	SCE [■]	LSR [■]	ELR+ [■]	TNF
<i>Acc.</i>	86.25 ±0.006	86.57 ±0.007	86.60 ±0.007	87.73 ±0.006	86.19 ±0.008	88.04 ±0.008	88.61 ± 0.009
<i>Sen.</i>	89.92 ±0.008	88.67 ±0.010	90.24 ±0.008	90.05 ±0.009	89.58 ±0.007	90.61 ±0.009	90.93 ± 0.007
<i>Spe.</i>	82.67 ±0.05	84.19 ±0.005	83.03 ±0.004	85.41 ± 0.007	82.89 ±0.006	84.52 ±0.006	85.20 ±0.006

performance. As shown in Table 2, the proposed TNF still achieves the best performance among all existing methods, demonstrating the generalization of our method.

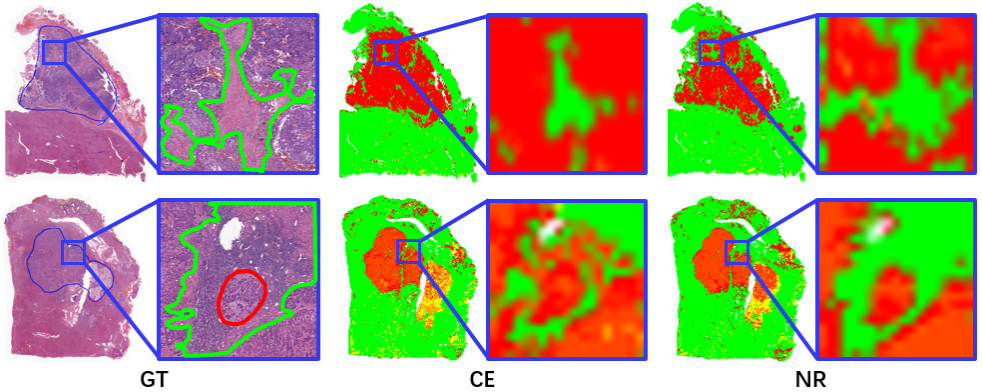


Figure 3: The visualization results of models trained by different losses. Here the left figure is the thumbnails of the whole slides, and the right figure is local region to show the details. In the figure of Ground Truth, accurate annotations are supplemented for better evaluation, where green loop denotes the healthy region, and the red loop denote the cancerous region. The size of the local region is 15000×15000 , containing about 900 patches with 50x magnification.

C. Visualization Results

This paper has shown exhaustive results to demonstrate the high accuracy and robustness in test dataset. The fundamental idea of our framework, on the other hand, is rectifying noisy labels during training to avoid overfitting. So in order to evaluate the rectification ability of the proposed NR loss, visualization figures of models trained by CE and NR are shown in Fig. 3. Additional accurate annotations are added in the enlarged local regions. From Fig. 3, we can see that these regions are roughly annotated as tumors by pathologists. The model trained by CE will be confused and misguided by the noisy labels. Conversely, the prediction of model trained by NR is much more similar to the Ground Truth (GT). It indicates that with the proposed NR loss, the model will be well optimized by the rectified

labels even the annotated labels of these patches are incorrect, which makes our framework robust in HCC grading with noisy annotations.

D. Generalization of NR Loss

The NR loss proposed in this paper is derived based on the special characteristics of HCC grading. Since each sample only contains one kind of HCC for most samples, the NR loss for HCC grading contains binary targets. Nevertheless, this NR loss can be generalized to multi-class labels by replacing the item of the reversed label with the product of items on other categories, which is defined as:

$$\mathcal{L}_{NR-multi} = -\log(1 - (1 - p_j)^{1-\alpha} \prod_{k \neq j}^K (1 - p_k)^\alpha), \quad (1)$$

where p_j denotes the prediction on the annotated label, and p_k denotes the prediction on other labels. This multi-class NR loss is evaluated on MNIST and CIFAR-10 with 40% random noise. The results are shown in Table 3. Here, the α is set as constant 0.01, and all these methods use ResNet-18 as the backbone. The results demonstrate that our proposed NR loss can also be generalized into multi-class for rectifying the noise.

Table 3: The classification accuracy of different methods on MNIST and CIFAR-10 with 40% random noisy labels. All these results averaged over ten experiments and the best results are marked in **bold** (All accuracies are in %).

	CE	CM [■]	SCE [■]	ELR+ [■]	NR
MNIST	98.28 ± 0.002	98.36 ± 0.002	98.64 ± 0.001	98.69 ± 0.002	98.90 ± 0.002
CIFAR-10	80.06 ± 0.018	81.77 ± 0.014	84.38 ± 0.018	84.92 ± 0.022	85.84 ± 0.017

References

- [1] Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA: The Journal of the American Medical Association*, 318(22):2199–, 2017.
- [2] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, and Fei Fei Li. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision Pattern Recognition*, 2009.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [4] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning Regularization Prevents Memorization of Noisy Labels. *arXiv preprint arXiv:2007.00151*, 2020.

- [5] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020.
- [6] Charles E Metz. Some practical issues of experimental design and data analysis in radiological roc studies. *Investigative Radiology*, 24(3):234–245, 1989.
- [7] Nancy A Obuchowski. Receiver operating characteristic curves and their use in radiology. *Radiology*, 229(1):3–8, 2003.
- [8] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *IEEE Conference on Computer Vision Pattern Recognition*, 2017.
- [9] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. Learning From Noisy Labels by Regularized Estimation of Annotator Confusion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [10] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric Cross Entropy for Robust Learning with Noisy Labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019.