2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Robust Classification of Histology Images Exploiting Adversarial Auto Encoders

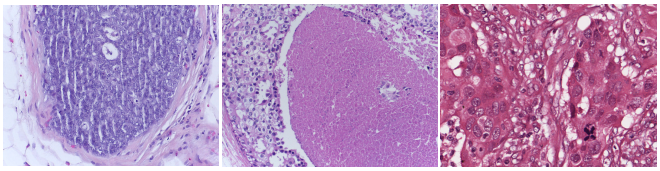
Nikhil Cherian Kurian, Gurparkash Singh, Poorvi Hebbar, Shreekanya Kodate, Swapnil Rane, Amit Sethi

***Abstract*— Deep learning (DL) thrives on the availability of a large number of high quality images with reliable labels. Due to the large size of whole slide images in digital pathology, patches of manageable size are often mined for use in DL models. These patches are variable in quality, weakly supervised, individually less informative, and noisily labelled. To improve classification accuracy even with these noisy inputs and labels in histopathology, we propose a novel method for robust feature generation using an adversarial autoencoder (AAE). We utilize the likelihood of the features in the latent space of AAE as a criterion to weigh the training samples. We propose different weighting schemes for our framework and evaluate the effectiveness of our methods on the publically available BreakHis and BACH histopathology datasets. We observe consistent improvement in AUC scores using our methods, and conclude that robust supervision strategies should be further explored for computational pathology.**

I. INTRODUCTION

Supervised deep learning (DL) models have consistently shown promising results for automated image analysis in medical image analysis over the last eight years [1], [2]. However, the success of DL models depends on the availability of large datasets of high quality and correctly labelled images for training. When the quality of the training images or the accuracy of their labels degrade, the accuracy of the DL models trained using them reduces drastically [3]. Consequently, for automated medical image analysis in general, and computational pathology in particular, medical experts on a research team need to carefully label, annotate, and curate whole slide images (WSIs) to prepare training and testing datasets. This process often involves precise annotations of regions of interest (ROIs) so that high quality and homogeneous patches (sub-images) of anatomical structures can be mined. These patches then inherit the same label for as the ROI from which these are mined. This data preparation process is time-consuming and expensive.

On the other hand, weakly supervised labels for WSIs (or large images, in general) are much easier to obtain by simply mining their associated electronic medical records (EMRs) for overall diagnosis. Such weak supervision disregards the heterogeneity in the quality and anatomical content of patches mined from a single slide. For example, image quality can vary with tissue preparation, staining, and slide preparation methods [4], as shown in Figure 1. Additionally, intra-tumoral heterogeneity is a natural phenomenon. For example, tumoral and benign structures occur side by side and there is spatial variations in disease grade or mutational landscape, often in a single slide. Spatial heterogeneity in anatomy and quality combined with the gigapixel size of WSIs means that weakly supervised label propagation from the WSI to ts patches leads to mislabeling of a certain unknown proportion of the individual patches.



(a) (b) (c)

Fig. 1: Diversity in the quality of histopathology images: (a) Venetian blinds artifact due to improper cut, (b) tissue degradation due to improper fixation, and (c) over-staining of eosin dye.

In this paper, we address the problem of learning robust models for image classification in the face of label noise and weak supervision. We use adversarial autoencoders to get sample-wise weights for each training image. We assume and confirm that the samples that can deteriorate the model training will fall into the lesser likelihood regions of the class-specific distribution priors. This assumption eliminates the need for additional optimization steps to calculate the sample-wise weights. We also explore different weighting strategies such as binary weighting, binary normalized weighting and normalised weighting schemes that can be used to weigh variants of cross entropy loss function for robust supervision. We use the publically available BACH [5] and BreakHis [6] breast cancer histopathology to evaluate the effectiveness of the proposed weighting scheme in the face of labelled noise.

II. RELATED WORK

Previous attempts to prevent overfitting on noisy outliers includes curriculum learning, self-paced learning, and robust loss functions. In curriculum learning the model is trained gradually using easier samples first, similar to how human are taught [7]. Unsupervised measures, such as entropy of classification output, are used to calculate the hardness of the training samples. Self-paced learning schemes incorporate label information by including the loss of a sample as a measure of hardness [8]. These two ideas have been extended in self-paced curriculum learning [9], self-paced boosting [10] and diversity-based self-paced learning [11]. Meta-

N.C Kurian,S. Kodate and A. Sethi are with the Department of Electrical Engineering, IIT Bombay, India (corresponding author e-mail address: nikhilkurian@iitb.ac.in)

G.Singh and P.Hebbar are with the Department of Computer Science and Engineering, IIT Bombay, India

S.Rane is with the Tata Memorial Centre, Mumbai, India

learning based sample weighting has also been explored [12]. Loss functions that are not over-eager to fit (have high gradient) on the outliers have been demonstrated to be robust to sample noise, such as the L1 loss [13] and the generalized cross-entropy loss [14].

Adversarial autoencoder (AAE), which we use in this work, is an extension of regular autoencoder that induces a prior distribution in the bottleneck layer [15]. Sampling from the prior distribution leads to a generative model that can be used for feature extraction [16] and anomaly detection [17].

Robust learning for histopathology image classification has been explored only in a few works. For instance, a novel loss function and graph-based ensemble boosters to enhance the strength of training samples has been proposed [18]. Recently [19] proposed a modification of generalized cross entropy and empirically verified the advantage of such loss functions on large histology dataset. Self-similarity between multiple patches has also been used to smoothen label noise [20].

III. Methodology

Our proposed robust supervision technique falls in the class of sample weighting schemes. Unlike other techniques that involve additional optimization steps or predefined curricula, we utilize the likelihood of the features of a sample in the latent space of AAE to derive its dynamic weight. We hypothesized that the less informative or the noisy samples will fall into the low likelihood regions of the regularized latent space.

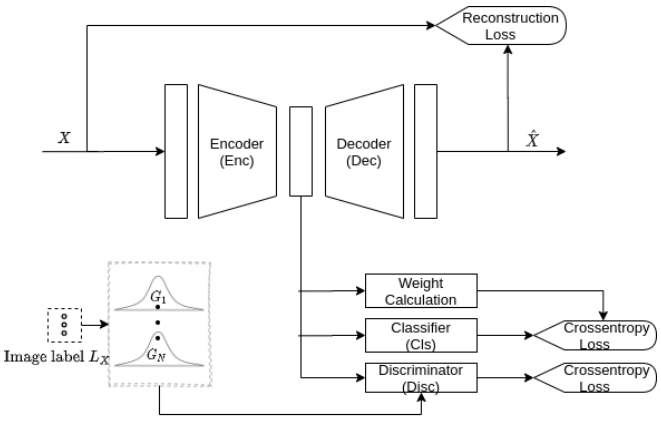


Fig. 2: Adversarial auto-encoder based architecture used for robust classification.

As shown in figure 2, our model has an encoder block that acts as a feature generator for the task-specific classifier as well as for the AAE. The adversarial training for the generated features in the d-dimensional latent space is performed with the aid of the discriminator block. The discriminator compares the features generated by the encoder with a random vector sampled from one of the N corresponding class specific prior distribution, which we assume to be a d-dimensional Gaussian distribution. Since the task-specific classifier is also needs to be optimized, the encoder block in this adversarial task has to generate samples that are optimized both for the classifier as well as to fool the discriminator. Here the role of decoder block in our architecture is to ensure that all images belonging to a particular class are pulled towards a mean feature vector of its Gaussian prior. During the training phase, feature generator and discriminator will perform the following min-max game to generate the samples:

$$\arg\max_{Disc} \arg\min_{Enc,Dec,Cls} [C(X, L_X) + \lambda_1 R(X) - \lambda_2 D(X, P)]$$

where C is the classification loss, R is the reconstruction loss, and D is the discriminator loss. These losses are sample-wise weighted cross-entropy, mean square error, and cross entropy respectively in our scheme. Further, X is a training sample and L_X is its label, and P is a sample from the prior distribution. Hyperparameters λ_1 and λ_2 were decided based on validation.

When the discriminator reports low confidence in distinguishing a real Gaussian sample against the feature vector, the adversarial training is declared successful. Training the classifier separately on top of a well-trained AAE generator gave poor classifier performance because such a feature generator was agnostic to the classification task a priori.

We explored the use of following schemes for sample-wise weighting in the loss C to train the adversarial autoencoder based classifier (AAEC):

- **Binary weighting (BW):** The sample's class-specific likelihood is compared to a global threshold, which is a tunable hyperparameter, to decide on its inclusion or exclusion (binary weight).
- **Binary normalized weighting (BNW):** Binary weighting is computed separately within each training mini-batch by normalizing the likelihood within the batch and comparing that to a threshold.
- **Normalised weighting (NW):** Continuous weights are obtained by normalizing the likelihood within each batch.

The binary weighting schemes described above are similar to curriculum and self-paced learning frameworks that allows only easy samples to appear in the training phase based on the age (state of learning iterations) of the model. On the other hand, the NW scheme ensures that all samples are represented in training, albeit with different weights. We further extended BW and BNW schemes, choosing the best of the models as the initial weights and continue training on these without any explicit weighting. We call these set of schemes that continue their training from BW and BNW without any weights as **Binary Weighting-No Weighting (BWNW)** scheme and **Binary Normalised Weighting-No Weighting (BNWNW)** scheme respectively.

IV. Experiments

We used a the same network architectures, as shown in Table I for two experiments – one where we artificially added label noise, and another where we worked with an unknown level of noise. Both of our experiments were

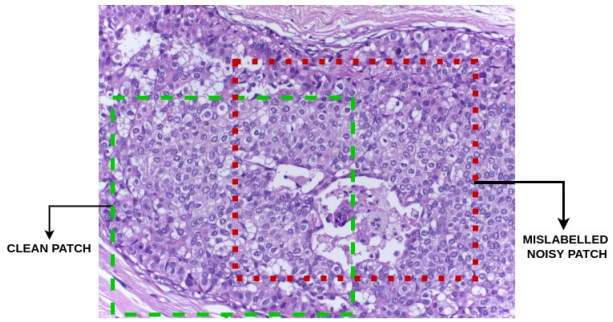


Fig. 3: Patch extraction from DCIS region can result in noisy samples (red box) where the basement membrane, which is its tell-tale feature, is not visible, or a good sample (green box) where it is visible.

based on binary classification ($N = 2$), with the centres of the two prior multivariate Gaussian distributions located at the opposite ends of a d-dimensional hypersphere in the AAE latent space. For both the experiments, a clean subset of the datasets that constituted around 30% of the total number of images were used for validation and testing. The rest 70% of the images that had corrupted labels were used for training the model.

TABLE I: Network architectures used in our experiments

Network ↓	Layer Type	Kernels	Dropout	Activation
Encoder	Conv	4	0.5	Leaky-ReLU
	Conv	8	0.5	Leaky-ReLU
	Conv	16	None	None
Decoder	Trans.-Conv	8	0.3	Leaky-ReLU
	Trans-Conv	4	0.3	Leaky-ReLU
	Trans-Conv	36	None	None
Classifier and Discriminator	Dense	32	0.5	Leaky-ReLU
	Dense	16	0.5	Leaky-ReLU
	Dense	2	None	Softmax

For convolution and transpose-convolution (upsampling) layers, we used kernels of size 5×5 with a stride of 3. Batch-normalization was used after each layer of all four segments of the model. The activation function used throughout the network is leaky-ReLU with a slope of 0.2 for the negative inputs. We used Adam optimizer with a learn rate of 0.001 and a batch size of 32 for both the experiments. Data imbalance in our experiments was accounted using a proportionate weighted sampling for each mini-batch updates.

Tumor versus non-tumor: For our first set of experiments, we use the BreakHis dataset [6], which is available at different magnifications. We took 400x magnification consisting of 1,450 images divided between tumor and non-tumor images. To test the robustness of our method, we synthetically added label noise by randomly flipping the labels of a pre-determined percent of training samples. We varied this percentage from 0 to 20 in steps of 5 and we show the AUC values on a clean (without label noise) test (held-out) dataset in Table II and plot the ROC curves for 20% noise in Figure 4. We further added color jitter based data-augmentation on the fly to simulate the staining variations found in the real world. We compared our models with the proposed weighting schemes by training a ResNet18 network (transfer learning from a pre-trained model) that does not use any sample-wise weighting.

DCIS versus IDC: In this set if experiments we used ICIAR 2018 Grand Challenge dataset called Breast Cancer Histology (BACH) [5]. We took on the challenging problem of classifying between ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC). The distinction between these classes is the presence (in DCIS) or the absence (in IDC) of a basement membrane around tumorous cells that otherwise look quite similar. To generate the training data, we sampled patches of size 512x512 pixels from the original images of the size of 2048x1536. The sampling resulted in the dataset that contained around 1540 patches. Such sub-sampling presents a more realistic scenario than the previous experiment for medical image classification, although we do not have much control over the percent of labels that are noisy. For instance, a patch sampled from a DCIS image may not include a basement membrane, which makes it indistinguishable from an IDC sample. Further, some samples may contain no tumor region at all – neither DCIS nor IDC. Additionally, some patches may have other artifacts as shown in Figure 1. Once again, we evaluated the robustness of our models on clean test (held-out) samples by manually curating the test cases before training the models. We show the AUC values in Table III and the ROC curves are shown in Figure 4. In this experiment, the noise level was unknown, as we only used the native label noise and did not synthetically add additional noise.

V. Results and Conclusion

The AUC values for both of our experiments are shown in table II and III. The worst case ROCs curves for two experiments are shown in Figure 4. From the results of the tumor versus non-tumor experiment, we observe that the performance of a regular CNN network worsens drastically as label noise level is increased. Further the DCIS versus IDC experiment shows that the robust weighting strategies perform much better than a conventional CNNs that implicitly overfits on noisy samples. The weighing strategies we found to be most robust in both the experiments. A visual sample of high and patches from DCIS and IDC classes shown in Figure 5 confirms that low-likelihood patches are less informative as these contain substantial non-tumoral portions for both classes, and fail to capture the basement membrane for DCIS. These experiments support the direction of further developing strategies for more robust histopathology, especially when the quality in image or strong supervision cannot be ensured.

TABLE II: Test AUCs by noise levels for tumor vs. non-tumor

Label noise level → Training Scheme ↓	0%	5%	10%	15%	20%
Conv-Net	**0.903**	**0.851**	0.802	0.790	0.796
AAEC-BWNW	0.815	0.802	0.802	0.806	0.763
AAEC-BW	0.823	0.819	0.819	0.805	0.802
AAEC-BNWNW	0.810	0.808	0.808	0.797	0.790
AAEC-BNW	0.836	0.828	**0.827**	0.808	0.803
AAEC-NW	0.838	0.829	0.826	**0.821**	**0.814**

TABLE III: Test AUCs for DCIS vs. IDC with native noise

Training Scheme ↓	AUC score
Conv-Net	0.809
AAEC-BWNW	0.810
AAEC-BW	0.836
AAEC-BNWNW	0.841
AAEC-BNW	0.843
AAEC-NW	**0.855**

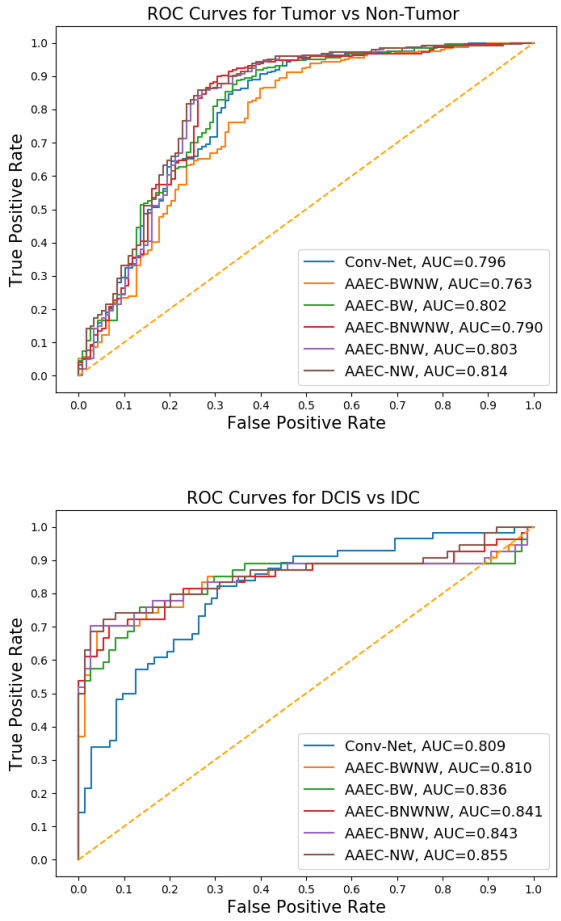


Fig. 4: ROC curves of the experiments

VI. Compliance with Ethical Standards

This research study was conducted retrospectively using human subject data made available in open access by [5]. Ethical approval was not required as confirmed by the license attached with the open access data.

References

[1] K. Yasaka, H. Akai, A. Kunimatsu, S. Kiryu, and O. Abe, "Deep learning with convolutional neural network in radiology," *Japanese journal of radiology*, vol. 36, no. 4, pp. 257–272, 2018.

[2] N. Cherian Kurian, A. Sethi, A. Reddy Konduru, A. Mahajan, and S. U. Rane, "A 2021 update on cancer image analytics with deep learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 4, p. e1410, 2021.

[3] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, 2020.

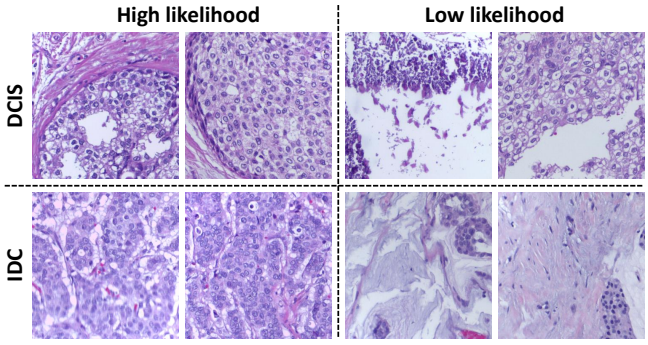


Fig. 5: Sample patches that appear in the high and low likelihood regions for DCIS and IDC

[4] A. Patil, M. Talha, A. Bhatia, N. C. Kurian, S. Mangale, S. Patel, and A. Sethi, "Fast, self supervised, fully convolutional color normalization of h&e stained images," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1563–1567.

[5] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan *et al.*, "Bach: Grand challenge on breast cancer histology images," *Medical image analysis*, vol. 56, pp. 122–139, 2019.

[6] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[7] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[8] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in neural information processing systems*, 2010, pp. 1189–1197.

[9] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning." in *AAAI*, vol. 2, no. 5.4, 2015, p. 6.

[10] T. Pi, X. Li, Z. Zhang, D. Meng, F. Wu, J. Xiao, and Y. Zhuang, "Self-paced boost learning for classification." in *IJCAI*, 2016, pp. 1932–1938.

[11] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *Advances in Neural Information Processing Systems*, 2014, pp. 2078–2086.

[12] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4334–4343.

[13] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[14] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8778–8788.

[15] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[16] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," in *Proc. Interspeech 2017*, 2017, pp. 1243–1247. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1421

[17] N. Li and F. Chang, "Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder," *Neurocomputing*, vol. 369, pp. 92–105, 2019.

[18] X. Shi, H. Su, F. Xing, Y. Liang, G. Qu, and L. Yang, "Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis," *Medical Image Analysis*, vol. 60, p. 101624, 2020.

[19] N. C. Kurian, P. S. Meshram, A. Patil, S. Patel, and A. Sethi, "Sample specific generalized cross entropy for robust histology image classification," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1934–1938.

[20] H.-T. Cheng, C.-F. Yeh, P.-C. Kuo, A. Wei, K.-C. Liu, M.-C. Ko, K.-H. Chao, Y.-C. Peng, and T.-L. Liu, "Self-similarity student for partial label histopathology image segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 117–132.