# Whole Slide Pathology Image Patch Based Deep Classification: An Investigation of the Effects of the Latent Autoencoder Representation and the Loss Function Form

Ana Lomacenkova and Ognjen Arandjelović

School of Computer Science, University of St Andrews, Scotland, United Kingdom

*Abstract*— **The analysis of whole-slide pathological images is a major area of deep learning applications in medicine. The automation of disease identification, prevention, diagnosis, and treatment selection from whole-slide images (WSIs) has seen many advances in the last decade due to the progress made in the areas of computer vision and machine learning. The focus of this work is on patch level to slide image level analysis of WSIs, popular in the existing literature. In particular, we investigate the nature of the information content present in images on the local level of individual patches using autoencoding. Driven by our findings at this stage, which raise questions about the us of autoencoders, we next address the challenge posed by what we argue is an overly coarse classification of patches as tumorous and non-tumorous, which leads to the loss of important information. We showed that task specific modifications of the loss function, which take into account the content of individual patches in a more nuanced manner, facilitate a dramatic reduction in the false negative classification rate.**

## I. Introduction

In broad terms, the focus of this work is on the analysis of images of pathology slides in medicine, which is an increasingly important task in the realm of digital pathology [1], [2], [3], [4]. The information contained in a whole-slide image (WSI) of a patient's tissue can aid in the early identification, prevention or treatment of various diseases, not the least important of which is cancer. However, analysis of WSIs by human pathologists is very time consuming and could greatly benefit from automation, be it partial or full. Due to the nature of WSI, ordinary computer vision techniques are poorly suitable for such analysis. In particular large volume of the images makes the application of off-the-shelf machine learning computationally infeasible, whereas the large image size poses challenges to deep learning. Some of the existing techniques addressing this problem implement downsampling of WSIs, producing lower-resolution and smaller volume image sets suitable for analysis by convolutional neural networks (CNNs). Yet this approach inevitably leads to the information loss due to the loss of high frequency, local detail. An alternative, also widely adopted in the existing literature, consists of breaking up WSIs into image patches, making patch based inferences, and then from these a whole slide level inference [5]. This approach also effects a loss of information, albeit in a different manner than downsampling. In particular, spatial information relating different patches, and any global information is lost – the focus is on local (patch level) information, and the assumption is that this information is sufficient for slide level analysis.

### A. Previous work

Computer-aided diagnostic systems are a major area of development in digital pathology. The volume and complexity of analysis tasks often make unassisted human interpretation inefficient: apart from being time-consuming, human interpretation uses only a small fraction of morphological information presented on pathology slides [6]. Therefore, a more flexible and robust automated approach is required.

Deep learning techniques started to gain popularity in the digital pathology field in roughly in 2015, following the advances in optimising the training algorithms for deep learning models. While widely ranging in design details, CNN-based solutions became a common choice for histopathological image analysis [7], [8].

The aforementioned problem of high-dimensionality of the input is mitigated with segmenting the WSIs into low-dimensional patches and processing them separately. This solution implies a trade-off: while the segmented input is easier to process, the global context of the image is lost. Nevertheless, it remains the most viable and efficient option when processing whole slide images with CNNs, since the segmentation preserves more morphological information than straightforward downsampling of the image. Another issue faced by the researchers in this area is the difficulty in obtaining a representative data set. For supervised training, WSIs must be annotated by human pathologists; the manual process of labelling gigapixel images is time-consuming and in addition introduces a certain level of labelling noise [9]. While this issue can be leveraged by using weakly supervised models which employ slide-level annotations, fully-supervised models require fine labelled data large enough to provide a sufficient amount of class balanced training data.

One of the factors that highly facilitated the advances in computer vision applications in digital pathology were the grand challenges introduced in 2012. Among the contests targeted at evaluating and comparing algorithms for the analysis of pathological images were EM segmentation challenge [10], mitosis detection challenges (2012 and 2013), GLaS gland segmentation challenge (2013), TUPAC (2016)

and Camelyon Grand Challenge (2016 and 2017) aimed at breast cancer detection [7]. Camelyon16 and Camelyon17 provided the participants with a fine-labelled dataset, which is freely available for download. The deep learning architectures such as GoogLeNet, AlexNet, VGG16 and others were all evaluated during the Camelyon16 contest, with GoogLeNet achieving the best result and VGG16 placing second. The winning team utilized a 27-layer GoogLeNet model and achieved an AUROC (area under the receiver operating characteristic curve) of 0.9250, compared with human pathologists' 0.9664 [11]. These and other results presented in response to the aforementioned challenges suggest that there is significant potential of the deep learning in the field of digital pathology.

In this work, we adopt the use of the Camelyon17 data set and the VGG16 architecture as the baseline for our investigations.

## II. Experiments and discussion

### A. Data

Experiments and analysis in the present work were performed on the Camelyon17 data set. It is a cancer metastases in lymph nodes detection challenge organised by the Diagnostic Image Analysis Group and Department of Pathology of the Radboud University Medical Centre. The data provided to the participants at the time of the contest is currently open access and therefore has high potential for reuse [12]. Camelyon17 consists of a total of 1000 WSIs across 5 different medical centres, converted into TIFF format. Slide level labels are available. Additionally, 10 WSIs from each medical centre are exhaustively annotated.

### B. Latent space representation of patches

Each $224 \times 224$ pixel RGB patch extracted from a WSI represents a 150528-element input for the CNN. Undoubtedly, not all of the information conveyed in such input is useful for the classification task at hand. Eliminating confounding information is helpful in several ways. Firstly, by reducing the dimensionality of data, it facilitates easier and more efficient learning. Secondly, it has the potential of representing the data in a more meaningful manner, thereby also aiding in the learning process [12]. Thus, our first goal in the present work is to investigate the impact of the size of this latent representation on patch based classification.

To preserve as much information as possible with a compact embedding, we decided to use convolutional autoencoders which have proven as highly successful in the task. A basic autoencoder is a simple network transforming an input into an output while minimising the error, where the error is the difference between input and output [13]. Between the input and output layers there are one or more hidden layers, consisting of encoder and decoder parts. The former extracts the latent representation from input data, whereas the latter decodes the feature representation in the latent space, thus producing output in the same space as input. Autoencoders have been successfully used for dimensionality reduction, information retrieval, and numerous other tasks [14]. A convolutional autoencoder differs from more basic autoencoder models, such as linear, in preserving the spatial locality of the input in a manner similar to other convolutional networks [15].

*1) Experiments & findings:* To examine the effect of the latent space dimensionality on performance, and thereby gain insight into the information content of patches, we designed a custom autoencoder architecture used for every compression level. Specifically, our models were trained with the embedding dimensionality being respectively 2, 8, 32, 65, 125, 250, 500, and 1000 times lower than the dimensionality of raw input.

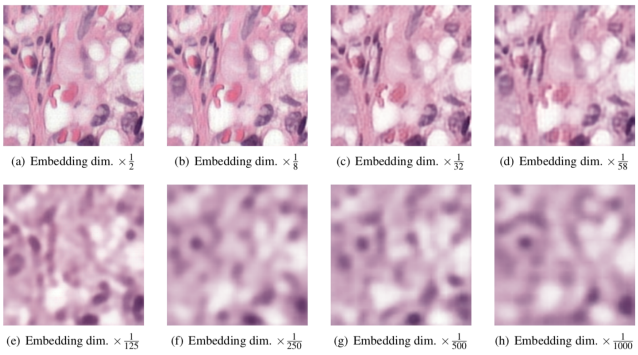
For lower compression levels (up to 65 times) the internal autoencoder architecture features six layers (3-layer encoder and 3-layer decoder), whereas for more severe compressions the architecture was deeper, comprising ten layers (5-layer encoder and 5-layer decoder).

Each convolutional layer except the last layer of encoder and decoder use the Rectified Linear Unit (ReLU) activation function, and the last layers use the Parametric Rectified Linear Unit (PReLU) instead of a more common sigmoid activation function. Sigmoid activation function is often successfully applied in binary classifiers, yet it gives rise to the problem of vanishing gradients: if a neuron's activation is saturated (is either 0 or 1), the gradient will be too close to zero for the network to learn efficiently [16]. ReLU has no upper bound for the output and therefore eliminates the problem of vanishing gradients [17]. However, ReLU must be used with caution with reconstruction units because of the hard saturation below the threshold of 0 [18]. PReLU (Parametric Rectified Linear Unit) alleviates this issue by introducing a learned parameter [19]. The use of ReLU and PReLU in the autoencoders architecture used in this work is in large part motivated by the fact that the input data does not contain any negative values and the empirical evidence that autoencoders with ReLU and PReLU often show better performance than the more common configuration with hyperbolic tangent and sigmoid functions.

Mean Squared Error loss function with Adam optimiser was used for training the autoencoders.

*a) Findings & discussion:* We start with qualitative, visual analysis. Examples of the patches reconstructed from the respective embeddings are shown in Figure 1. It can be readily seen that a severe reduction in the patch representation dimensionality effects major distortion both with respect to colour and texture, as well as the shape of cells. Nevertheless, significant compression is possible without virtually any observable loss of perceptual information. Overall, out findings suggest that the intrinsic dimensionality (or intuitively, complexity) of patches is indeed much lower than of their raw form.

However, perceptual change is of rather tangential interest and importance. It is possible that relevant, class discriminative information (non-tumorous vs tumorous) is retained even with major distortion to appearance. Hence, we next sought to examine quantitatively the effects of autoencoding, by looking at the ultimate goal: patch level classification



(a) Embedding dim. $\times\frac{1}{2}$ (b) Embedding dim. $\times\frac{1}{8}$ (c) Embedding dim. $\times\frac{1}{32}$ (d) Embedding dim. $\times\frac{1}{58}$

(e) Embedding dim. $\times\frac{1}{125}$ (f) Embedding dim. $\times\frac{1}{250}$ (g) Embedding dim. $\times\frac{1}{500}$ (h) Embedding dim. $\times\frac{1}{1000}$

Fig. 1: Example of autoencoded pathological WSI patch, for different ratios of latent to input space dimensionality (2–100).

| | | Ground truth | |
|---|---|---|---|
| | | Non-tumorous | Tumorous |
| Predicted | Non-tumorous | 95.12% | 17.20% |
| | Tumorous | 4.88% | 82.80% |

TABLE I: Baseline patch classification performance, using raw image data.

| | | Ground truth | |
|---|---|---|---|
| | | Non-tumorous | Tumorous |
| Predicted | Non-tumorous | 88.54% | 41.93% |
| | Tumorous | 11.46% | 58.07% |

TABLE II: Baseline patch classification performance, using autoencoding with latent to input space dimensionality ratio of 250.

performance. The corresponding confusion matrices using original and autoencoded (with the latent space dimensionality 250 times lower than that of the original input space) are shown in Tables I and II.

It can be immediately observed that the autoencoding process significantly affects performance – negatively so. Considering that the latent space dimensionality in the case shown is still rather high, approximately 600, the findings suggests a number of conclusions. The first one is that the dimensionality reduction effected by this autoencoder is not sufficiently sophisticated to capture the kind of appearance variation as can be observed in pathological WSI patches. A more important insight pertains to the very nature of the dimensionality reduction approach. In particular, our findings suggest that the focus on best *describing* local appearance in a latent space may not be the best approach to take when the end goal is that of discrimination.

An important observation to make is the poor performance in the classification of patches deemed tumorous. An understandable first reaction to this finding could attempt to explain it by the heterogeneity of tumour, noted earlier. However, we believe that a more considered examination suggests an alternative etymology. Firstly, the aforementioned heterogeneity is exhibited as much, if not more, in the non-tumorous tissue and is reflected in the immune system response. Hence, were heterogeneity the principal driver behind the phenomenon, it would have been seen in the classification performance of non-tumorous patches too. Rather, we hypothesised that the reason lies in the manner patches are labelled in the Camelyon17 data set – the presence of *any* amount of tumour within a patch results in the patch being labelled as tumorous. This sharp discontinuity, which does not take into account fully the content of the patch, that is the amount of healthy vs tumorous tissue, fails to facilitate good learning. To test our hypothesis, we examined the probability of false negative classification of a tumour patch as a *function of the proportion of tumour within the patch*, see Figure 2. The plot clearly supports our explanation. In particular, it can be seen that misclassification probability is low when (approximately) at least half of the patch is covered by tumour, but rises dramatically when there is little (but nevertheless, some) tumour presence. Motivated by this insight, in the next section we investigate how this challenge may be overcome.

### C. Improving patch analysis

As argued and evidenced in the previous section, labelling all patches which contain *any* tumour as tumorous is an impediment to the learning process. Indeed, the serious consequence of this is that the baseline VGG16 network based model results in an unacceptably high false positive rate of nearly 42%. Hence we now describe different approaches we investigated as a way of tackling this problem. A summary of the results follows.

*1) Non-binary, class weighted approach 1:* To address the previously highlighted, overly coarse labelling of patches as tumorous vs non-tumorous, we introduce a finer set of classes for use in the training stage. In particular, we classify patches as either having (i) 0% (no tumour content), (ii) more than 0% but less than 20% (low tumour content), (iii) between 20% and 60% (medium tumour content), or (iv) over 60% of tumorous content (high tumour content). In addition, to account for the greater seriousness of misclassifying patches with greater amount of tumour within them, we

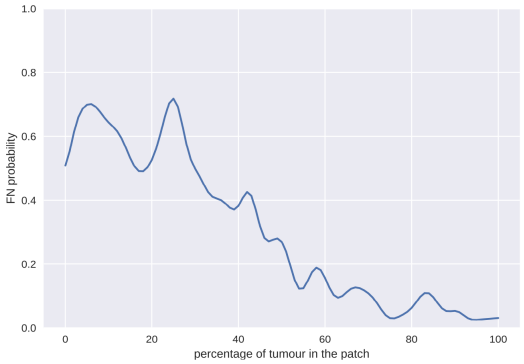



Fig. 2: Probability of false negative classification of patches deemed tumorous, using the baseline VGG16 network.

| Approach | False positive rate | False negative rate |
|---|---|---|
| Baseline | 11.46% | 41.93% |
| Class weighted 1 | 14.41% | 33.33% |
| Class weighted 2 | 13.24% | 22.96% |
| Error weighted | 27.72% | 5.38% |
| Decoderless | 24.84% | 9.49% |

TABLE III: Comparison of the effects of different training strategies, all using the same dimensionality of the autoencoder latent space (250 lower than the original input, i.e. approximately 600).

adjust the loss function so that the errors associated with the misclassification of the four classes are weighted in proportion 0.15:0.05:0.3:0.5 (n.b. in principle this weighting can be learned using the standard cross-validation methodology).

*2) Non-binary, class weighted approach 2:* The second approach we investigated is in substance nearly identical to the previous one, with the exception that previous classes (i) and (ii) were merged into one class. Hence, training was done using three classes, namely (i) less than 20% (no or low tumour content), (ii) between 20% and 60% (medium tumour content), or (iii) over 60% of tumorous content (high tumour content). The corresponding error weighting was 0.15:0.35:0.5.

*3) Non-binary, error weighted approach:* In this approach we retain the three classes of the previous method, but change the loss function in a substantially different manner. In particular, we adjust the form of the cost function so that instead of penalizing all misclassifications equally, more severe errors contribute to the loss more heavily. Specifically, the weighting is proportional to the difference of the true proportion of tumour within a patch and the nearest boundary of the incorrectly predicted class.

*4) Decoderless, non-binary, error weighted approach:* Lastly, we also examined the effects of removing the decoder from the process. In other words, we directly connected the output of the autoencoder encoding stage, i.e. the latent representation, to the classification convolutional neural network. In other aspects, the methodology is identical to the previously described three class, error weighting approach.

*5) Results & discussion:* Our findings are summarized in Table III. The first observation that can be readily made is that all of the proposed approaches aimed at reducing the false negative rate of the baseline method are successful in achieving this. Recall that our focus on false negatives is driven by clinical needs – an error in the form of a missed cancer detection is a far more serious one than an unnecessary alert to a healthy person. The best performing method in this context can be seen to employ error weighting in the loss function during the training stage, reducing the false negative rate from the original 41.93% to only 5.38% (nearly eightfold). At the same time, it is also insightful to observe the associated proverbial cost, in the form of the increased false positive rate. While this is of course undesirable, considering the aforementioned asymmetry in error type importance, the trade-off provides a net overall benefit.

## REFERENCES


[1] P. D. Caie, N. Dimitriou, and O. Arandjelovic, "Precision medicine in digital pathology via image analysis and machine learning," *Artificial Intelligence and Deep Learning in Pathology E-Book*, p. 149, 2020.

[2] I. P. Nearchou, D. A. Soutar, H. Ueno, D. J. Harrison, O. Arandjelovic, and P. D. Caie, "A comparison of methods for studying the tumor microenvironment's spatial heterogeneity in digital pathology specimens," *Journal of Pathology Informatics*, vol. 12, 2021.

[3] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, "Digital pathology and artificial intelligence," *The Lancet Oncology*, vol. 20, no. 5, pp. e253–e261, 2019.

[4] C. G. Gavriel, N. Dimitriou, N. Brieu, I. P. Nearchou, O. Arandjelović, G. Schmidt, D. J. Harrison, and P. D. Caie, "Assessment of immunological features in muscle-invasive bladder cancer prognosis using ensemble learning," *Cancers*, vol. 13, no. 7, p. 1624, 2021.

[5] X. Yue, N. Dimitriou, P. D. Caie, D. J. Harrison, and O. Arandjelovic, "Colorectal cancer outcome prediction from H&E whole slide images using machine learning and automatically inferred phenotype profiles," in *Proceedings of International Conference on Bioinformatics and Computational Biology*, vol. 60, 2019, pp. 139–149.

[6] A. BenTaieb and G. Hamarneh, "Deep learning models for digital pathology," *arXiv*, p. arXiv:1910.12329, 2019.

[7] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[8] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep learning for whole slide image analysis: an overview," *Frontiers in Medicine*, vol. 6, p. 264, 2019.

[9] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[10] "Segmentation of neuronal structures in EM stacks challenge – ISBI 2012," http://tinyurl.com/d2fgh7g, accessed: 2021-04-30.

[11] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv*, p. arXiv:1606.05718, 2016.

[12] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. van de Loo, R. Vogels, *et al.*, "1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset," *GigaScience*, vol. 7, no. 6, p. giy065, 2018.

[13] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML workshop on Unsupervised and Transfer Learning*, 2012, pp. 37–49.

[14] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[15] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proceedings of International Conference on Artificial Neural Networks*. Springer, 2011, pp. 52–59.

[16] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv*, p. 1811.03378, 2018.

[17] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv*, p. arXiv:1710.05941, 2017.

[18] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of International Conference on Artificial Intelligence and Statistics*. In Proceedings of JMLR Workshop and Conference, 2011, pp. 315–323.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of International Conference on Computer Vision*, 2015, pp. 1026–1034.