

Denoising AutoEncoder-Based Representation Learning for Multi-Task Whole Slide Image Analysis

JooHa Lee¹ and Sherrie Lah[#]

¹Crean Lutheran High School, USA

[#]Advisor

ABSTRACT

Along with the advancement of artificial intelligence, there have been significant improvements in the field of whole slide images (WSI). WSI in machine learning is mainly utilized for pathological analysis, consisting of diverse tasks such as classification of normal versus tumor patches, segmentation of precise areas of potential tumor, or object detection indicating tumor sites. However, training these distinct models for each individual task is both time intensive and inefficient. Therefore, there is a high demand for developing a unified learning algorithm capable of concurrently handling multiple WSI tasks. To address the aforementioned problem, a representation learning based transfer learning method is proposed to process multiple downstream tasks including classification, segmentation, and object detection. Synthesizing two stages, the proposed method utilizes the reconstruction of images from representation learning and pretrained parameters from transfer learning method to create a more accurate and time-efficient model for analyzing WSI. Overall, the proposed method offers a better representation of WSI, which leads to enhanced accuracy in analysis and interpretation. Through extensive experiments, I have found that the proposed method outperforms previous state-of-the-art networks in various downstream tasks including classification, segmentation, and object detection. I expect the proposed method to be applied in real world scenarios with increased practicality and accuracy.

Introduction

Pathology Image

Pathology images provide valuable insights into the cellular and tissue-level changes that occur in the human body due to diseases. These images play a crucial role in characterizing a patient's condition, guiding treatment decisions, and forecasting prognosis. Current methods of obtaining pathology images utilize microscopic slides analyzed and annotated by pathologists. Pathologists examine diverse annotations that span from providing regional diagnoses of broader areas to creating finer segmentations that highlight precise portions within an image. However, the method utilizing manual analysis has multiple flaws of being time consuming, labor intensive, and subject to the inter-observer. Alternatively, a machine learning based automated system with superior accuracy and efficiency offers a better prospect in disease detection.

Previous Method

In response to the preceding problem, there has been numerous research studying the use of computerized architecture in detecting tumors in various regions of the human body. Liu et al. proposed a framework using the convolutional neural network to localize and detect tumors in the lymph nodes near the breast (Liu et al.

2017). This method tended to produce False Negatives in smaller sized tumors, and were unable to tune hyperparameters due to nearly perfect area under curve scores.

Halicek et al. proposed a digitized architecture in order to predict primary head and neck squamous cell carcinoma (SCC) and thyroid carcinoma (Halicek et al. 2019). Their system was limited to detecting a certain range of SCC cell size due to the application of down-sampled resolution and required further investigation as the type of trained images was different from the ones that are usually used in clinical settings.

Wang et al. presented a transfer learning method based on convolutional neural network architecture that classifies WSI patches of liver cancer to improve the efficiency of clinical diagnosis (Wang et al. 2021). Their network did not provide external validation on real-world clinical datasets, which may result in the variance of performance in real-world scenarios.

Propose Method

To tackle this challenge, I propose a representation learning based transfer learning method to process multiple downstream tasks including classification, segmentation, and object detection for whole slide images (WSI). The proposed method includes two stages: an AutoEncoder based representation learning and the transfer learning method. In the first stage, an image is inputted into an encoder to produce an activation map that compresses and extracts some features of the WSI. The activation map then goes through a decoder, which results in a reconstructed WSI based on the features of the activation map. The overall goal of the first stage is to extract the most important features of the map while carrying the ability to reconstruct the original image with high accuracy. I also proposed denoising, which adds image noise to the WSI to robustly extract better features of the WSI. The second stage utilizes transfer learning, where the pretrained parameters trained in the first stage are brought to efficiently train downstream networks such as classification, segmentation, and object detection.

There are various features of an WSI that the convolutional neural network extracts from, such as regions of possible tumor. In order to improve a training model's accuracy and performance, it is important for the convolutional neural network to extract only the most important features of an image. The first stage utilizing representation learning ensures that as the model is trained, it will produce an efficient activation map and reconstructed image. As opposed to the supervised approach that randomly initializes parameters during the learning process, the proposed method utilizes transfer learning with pretrained parameters that will find better representations of the WSI. Furthermore, it effectively trains multiple downstream tasks, which speeds up the training process with higher accuracy.

This research paper is structured into the following parts: chapter 2 describes the background knowledge of the topic, chapter 3 explains the proposed approach, chapter 4 shows the experimental results of the proposed method, and chapter 5 concludes the research paper.

Background Knowledge and Related Work

Whole Slide Image (WSI)

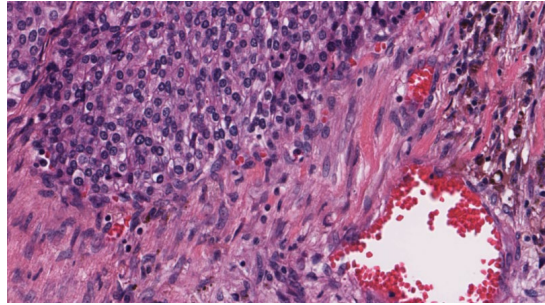


Figure 1. Example of an whole slide image (NVIDIA 2023)

Whole slide imaging is the process of obtaining high-resolution digital images at the microscopic level. These slides are obtained by extracting tissue samples from various parts of the body, then placing them under a microscopic scanner. The produced digital images, called whole slide images (WSI), are in high resolution and magnification. Furthermore, they contain multiple features such as being able to enlarge certain regions for a better view or adding direct annotations and labels to the image. These aspects allow a time efficient and accurate process of analysis compared to other methods of tissue evaluation.

WSI is extracted from various parts of the body including different organs, tissues, and skin. They are widely utilized in pathological analysis in order to detect possible abnormalities such as tumors. The slides can be viewed from multiple magnifications and segmented areas, where various features of focus are interpreted in close detail. This feature enables pathologists to examine certain regions of the tissue in depth and thus increase the accuracy of tumor detection.

Types of WSI Analysis

WSIs, as described above, are widely utilized in different types of pathological analysis, including classification, object detection, and segmentation.

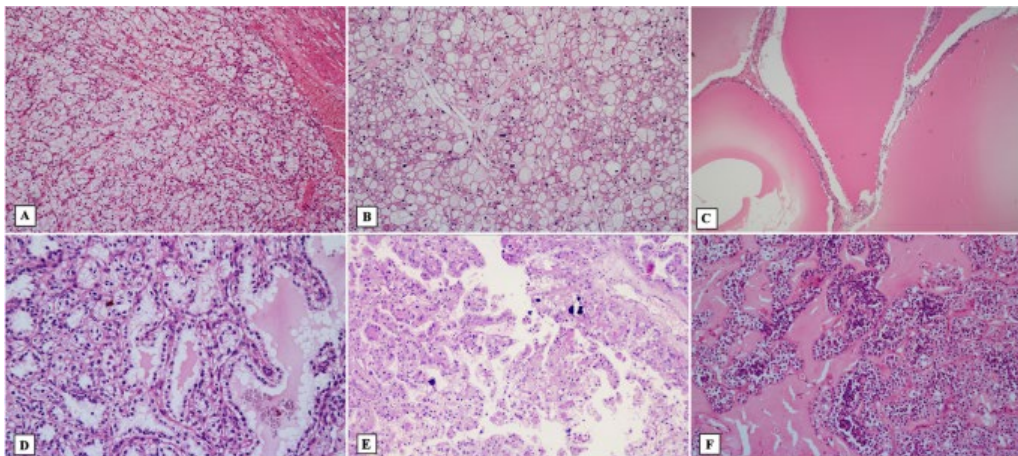


Figure 2. Example of classification task (Athanzio et al. 2021)

(a) clear cell carcinoma, (b) wrinkled nuclei and perinuclear halos, (c) cystic spaces with delicate septae, (d) clear cell papillary carcinoma, (e) MiT family translocation carcinoma, and (f) biphasic tumor.

Classification task of WSI is determining whether a specific feature is contained in the WSI or not, usually in distinct categories. With WSI, artificial networks are trained to produce a prediction value from the specific categories. An example is determining if a metastatic tissue, which is a tissue that is able to spread to different regions of the body, is included in specific regions of the WSI. Figure 2 illustrates different features in WSI that are utilized in the task.

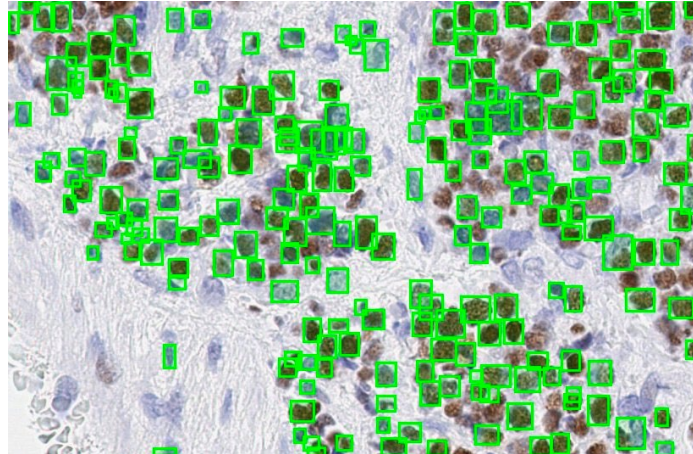


Figure 3. Example of object detection task (Chandradevan et al. 2019)

Object detection is the task of detecting specific features within a WSI. These features are marked in colored boxes that emphasize the indication. This task can be applied to a wide variety of WSI, ranging from locating specific types of cells to detecting abnormalities within the extracted tissue. Figure 3 features the detection of cell nuclei within a WSI. All of the green boxes represent the network's predictions.

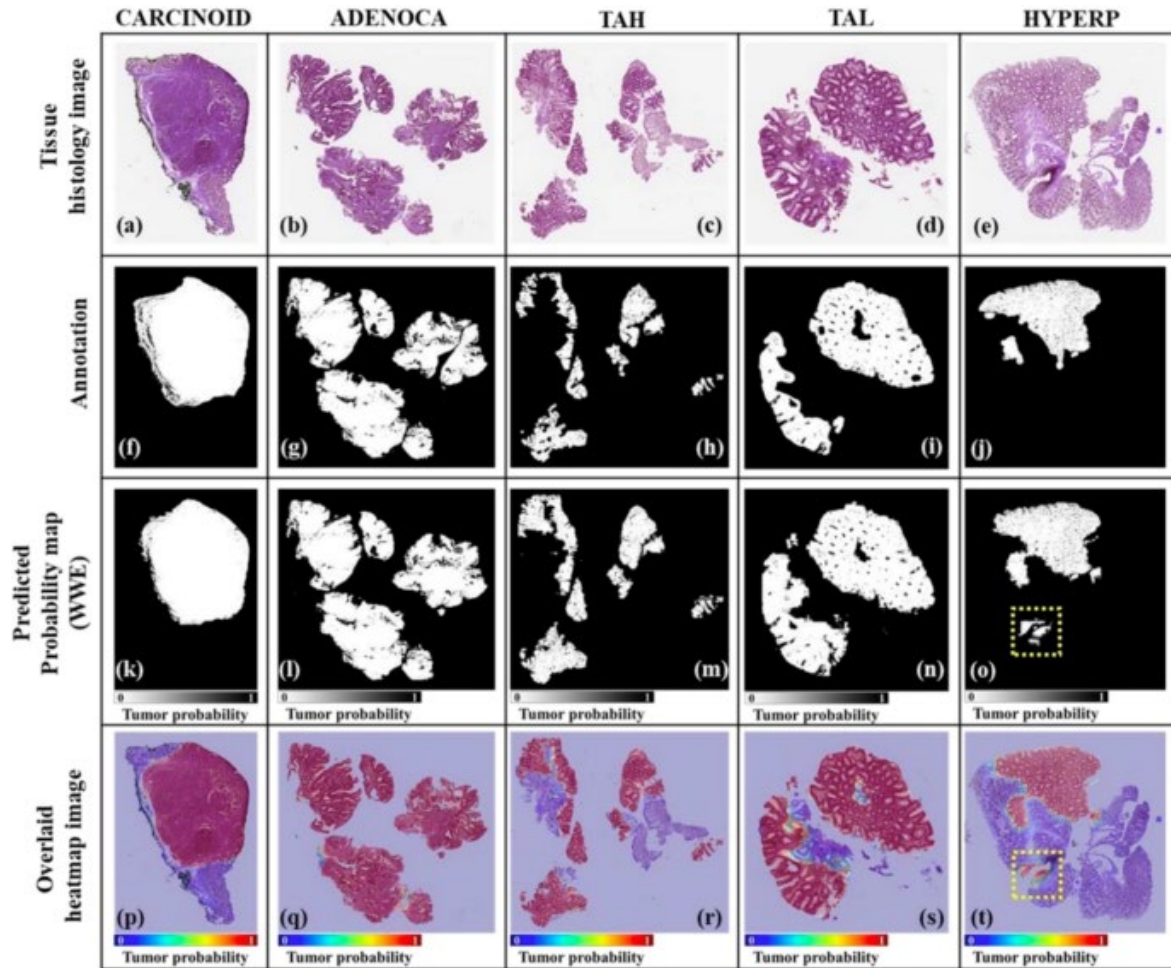
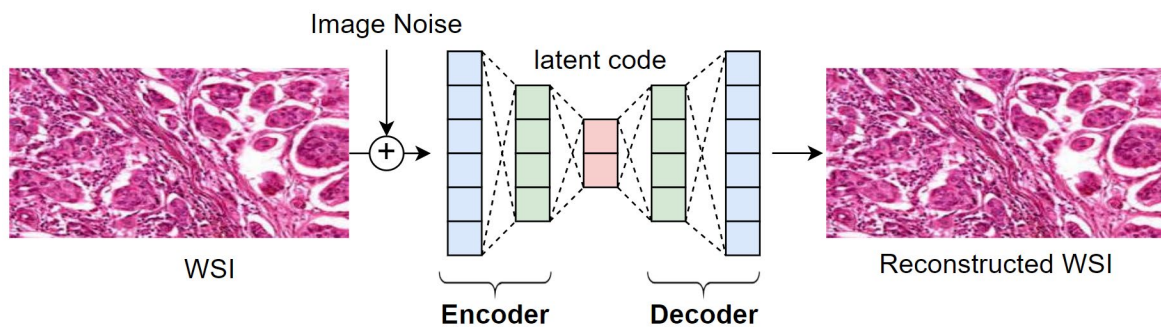


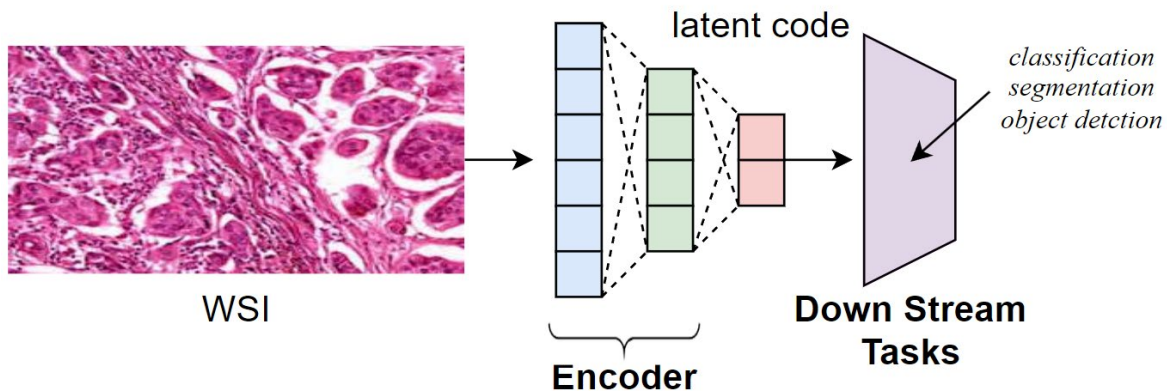
Figure 4. Example of segmentation task (Kim et al. 2021)

The task of segmentation is to segment or identify specific regions of the WSI. For example, the task of obtaining an approximate location for a cell's nucleus involves segmentation, where a network produces WSI highlighting specific segmented areas as its prediction. Figure 4 displays the segmentation task of colorectal cancer, where multiple regions of a single WSI is segmented in order to highlight different parts of the slide and predict the probability of tumor.

Proposed Approach



(a)



(b)

Figure 5. Overall structure of proposed network (a): Stage 1, (b): Stage 2

In this chapter, I provide a comprehensive review of the proposed method. The proposed network consists of two stages: representation learning and transfer learning, as depicted in figure 5. Figure 5 (a) shows an Auto-Encoder based representation learning, and Figure 5 (b) shows a transfer learning method that processes multiple downstream tasks. Chapter 3.1 and 3.2 describe the architecture and operation of each proposed method along with its concept and underlying assumption, and Chapter 3.3 outlines the training procedure and hyperparameters used in the experiments.

WSI Representation Learning

Figure 5 (a) shows the first stage of the proposed method, which is an AutoEncoder-based representation learning. The ultimate goal of this process is to find a better representation of WSI through the intensive learning process.

The input is a whole slide image (WSI) denoted as I , where $I \in \mathbb{R}^{HW}$ (H and W denote the height and width of the input image, respectively). After image noise is applied, the noise added image, denoted as \tilde{I} , goes through an Encoder. The Encoder is shown as $E : \tilde{I} \rightarrow Z$, where Z represents the latent code, or the output of the Encoder. The latent code Z is denoted as $Z \in \mathbb{R}^L$ (L denotes the dimension of the latent code). Next, the latent code goes through a Decoder ($D : Z \rightarrow \hat{I}$), where $\hat{I} \in \mathbb{R}^{HW}$ is the output, or the reconstructed WSI.

Through representation learning, the network is able to learn to extract the most important features of the WSI into the latent code, or activation map. During the learning process, the network can improve in the generalization of patterns depicted in the WSI, such as detecting or regionalizing possible tumor locations with better accuracy. Furthermore, the latent code is reduced in the size of pixel dimensions, which will reduce the learning speed of the network. Overall, the proposed network will have higher accuracy with shortened learning speed, therefore increasing the efficiency of diverse tumor detection tasks.

Downstream Task (Transfer Learning)

The second stage of the proposed network, a transfer learning method that is able to train multiple downstream tasks, is shown in Figure 5 (b). The stage starts with an input WSI, denoted as $I \in \mathbb{R}^{HW}$. The input goes through the pretrained parameters of the Encoder from the previous stage to produce a latent code. This step is repre-

sented as $E: I \rightarrow Z$, where Z , or the latent code, is able to process multiple downstream tasks including classification, segmentation, and object detection. The downstream tasks are denoted as P_i , where each downstream task of classification, segmentation, and object detection are marked as P_0 , P_1 , and P_2 , respectively.

Using transfer learning enables the network to gain better representations of the input WSI. As opposed to randomly initialized parameters, pretrained parameters from the Encoder reduces the learning speed of the network with higher accuracy of processing downstream tasks.

Loss Function

Equation 1: L1 loss function

$$L1 = \frac{1}{HW} \sum_x^W \sum_y^H |I(x, y) - \hat{I}(x, y)|$$

Here, H and W denote the dimensions of the input image, and (x, y) denotes the specific pixel value of the original and reconstructed WSI. I and \hat{I} denote the original image and reconstructed image respectively.

The input and output of the first stage of representation learning is a WSI and a reconstructed WSI, respectively. An effective way to evaluate the performance of the network would be to compare the difference between the two images. The minimal difference between the original and reconstructed WSI would be the ultimate objective of the learning process. To calculate the difference, an L1 loss function can be utilized. It shows the pixelwise difference between the original and reconstructed image iterated for the whole training process, divided by the pixel number of the input in order to produce an average. The loss function measures the difference in each corresponding pixels' intensity.

Equation 2: Cross entropy loss function

$$L_{CE} = -\log_e P$$

Here, P denotes the probability produced as the output of the classification task. The second stage of the proposed network includes multiple downstream tasks. Among them, the classification task produces a probability of a specific category as an output. Therefore, a loss function that is able to well evaluate the accuracy of the prediction is needed. The cross entropy loss function puts the output probability through a minus natural logarithmic function, which helps better compare the loss of the proposed network in relation to the desired score.

Equation 3: Intersection over Union (IoU) loss function

$$IoU = \frac{B_{gt} \cap B_p}{B_{gt} \cup B_p}$$

Here, B_{gt} and B_p denote ground truth and prediction of the specific area of the WSI respectively.

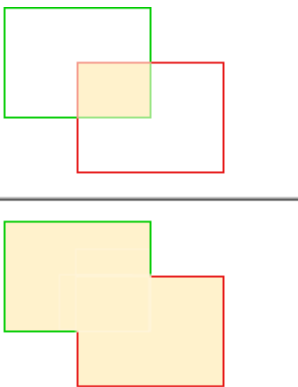
$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of Union}} = \frac{\text{Diagram 1}}{\text{Diagram 2}}$$


Figure 6. Visual representation of IoU

(The green box represents the ground truth, and the red box represents the prediction of the proposed network.)

For other downstream tasks such as object detection and segmentation, a IoU loss function is utilized. The main objective of these two tasks are to align with the ground truth region as closely as possible, so that the overlapping region between the ground truth and prediction is maximal. The IoU loss function is the area of overlap divided by the area of union, which produces a number between 0 to 1. A number closer to 1 implies that the prediction has more overlapping regions with the ground truth, and therefore the network has high accuracy.

Experimental Results

Dataset

Four main datasets are utilized to train and test the proposed network. The Metastatic Tissue Classification dataset is used in the classification of metastatic tissue. This dataset is composed of 327,680 samples extracted from lymph node sections, and is used to analyze whether the possible tumor will spread to different parts of the body or not. The CryoNuSeg Dataset, composed of tissues of ten different organs of the body, is involved in the segmentation of nuclei. The diversification of organs in the dataset reduces the probability of a biased training of the network.

Similar to the CryoNuSeg dataset, the PanNuke dataset is also used in the segmentation of nuclei. However, as pathology images are rare and hard to obtain in a large quantity, this dataset is made up of artificially generated images. These artificial operations produce tissue images from various organs, which contributes to assessing a model's performance. Object Detection signet ring cell dataset is utilized in the object detection of signet ring cells (SRC). SRCs are a type of tumor commonly found in the gastric mucosa and intestine with a highly poor prognosis, which makes early detection essential. Using this dataset, an output image with different colored boxes is generated. Green boxes represent the True Positives, or the correct predictions of the network, while the yellow boxes represent the False Negatives, or the areas of tumor the network failed to detect.

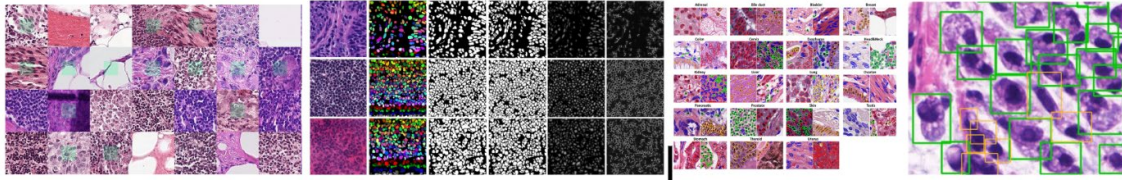


Figure 7. Example images of datasets used by the proposed network

(The images above are Metastatic Tissue Classification dataset, CryoNuSeg dataset, PanNuke dataset, and Object Detection signet ring cell dataset respectively.)

Evaluation Metric

IoU

For downstream tasks of object detection and segmentation, an Intersection over Union (IoU) evaluation metric is utilized.

Equation 4: Intersection over Union (IoU)

$$IoU = \frac{B_{gt} \cap B_p}{B_{gt} \cup B_p}$$

Here, B_{gt} and B_p denote ground truth and prediction of the specific area of the WSI respectively. The IoU evaluation metric is the area of overlap divided by the area of union, which produces a number between 0 to 1. A number closer to 1 implies that the prediction has more overlapping regions with the ground truth, and therefore has high accuracy.

Dice Coefficient Score

Equation 5: Dice Coefficient Score

$$Dice = \frac{2 \|B_{gt} \cap B_p\|}{\|B_{gt}\| + \|B_p\|}$$

Here, B_{gt} and B_p denote the ground truth and prediction of the proposed network respectively. In medical datasets, the area of overlap between the ground truth and prediction, or the True Positive area is the most important in the evaluation of a network. In order to emphasize this area of overlap, a Dice Coefficient Score is utilized as an evaluation metric. Similar to the IoU, the Dice Coefficient Score is double the area of overlap divided by the sum of the areas of ground truth and the network's prediction. The score will better evaluate the accuracy of the network according to its True Positive value.

Confusion Matrix

| | | |
|----------|---------------------|---------------------|
| | Positive | Negative |
| Positive | True Positive (TP) | False Negative (FN) |
| Negative | False Positive (FP) | True Negative (TN) |

Figure 8. Example of a confusion matrix

For the downstream task of binary classification, a confusion matrix is used. Binary classification is when the network produces either a positive or negative result, which is compared to the ground truth values of positive or negative. A correctly predicted positive is considered True Positive (TP), a correctly predicted negative is considered True Negative (TN), an incorrectly predicted positive is considered False Positive (FP), and an incorrectly predicted negative is considered False Negative (FN). These four values reflect the performance of the network as probabilities, and make up the confusion matrix.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - score = \frac{2 \cdot R \cdot P}{R + P}$$

Figure 9. Visual representations of accuracy, recall, precision, and F1-score values. (*R* denotes the recall value while *P* denotes the precision value.)

Based on the confusion matrix, values of accuracy, recall, precision, and F1-score are produced. The accuracy is the correct prediction of samples all over all predictions, which indicates a ratio of how many of all predictions of the network were correct. Recall is the TP value over the sum of TP and FN. In other words, it is the actual correct positive predictions over all positive ground truth values. However, the value may be misleading in some cases as only the true positive values are considered in the equation. Considered somewhat inversely proportional, the precision value is described as TP over the sum of TP and FP. This value captures the recall value's error, and aids in better evaluating the proposed network. The F1-score is calculated as two times the product of the recall and precision value all over the sum of the recall and precision value. This is also known as the harmonic mean of recall and precision.

ROC Curve



Figure 10. Example of an ROC Curve

The ROC Curve utilizes a one hot vector, which is a list of values of either 0 or 1 that represents probability. With the data of lists containing numbers, each representing a probability of whether the value is positive or negative, a threshold is set. Numbers over the threshold are considered positive with the value of 1, while numbers under the threshold are considered negative with the value of 0. A stress test of changing values of the threshold is conducted in order to analyze the performance of the classifier. From the test, the change in TP and FP rate is evaluated.

During the evaluation, threshold values are changed in order to assess the change in TP and FP rate. Graphical interpretations of these changing rates are shown as the ROC Curve. As the curve gets closer to the positive y axis, accuracy increases. A desired network would have a curve with an angle close to 90 degrees, adhering to the y axis.

Performance Comparison

Classification

Table 1. Experimental results for accuracy, recall, precision, and F1-score

| | Accuracy | Recall | Precision | F1-score |
|--------------------------------------|-----------------------------------|-------------------------|-------------------------|-------------------------|
| VGG19 (Simonyan et al. 2014) | 0.9292 (±0.0010) | 0.9494 (±0.0016) | 0.9142 (±0.0013) | 0.9281 (±0.0011) |
| EfficientNet-B7 (Tan et al. 2019) | 0.9329 (±0.0011) | 0.9532 (±0.0008) | 0.9111 (±0.0009) | 0.9320 (±0.0008) |
| HRNet-40 (Wang et al. 2020) | 0.9380 (±0.0009) | 0.9528 (±0.0011) | 0.9150 (±0.0015) | 0.9358 (±0.0012) |
| Resnet-50 (He et al. 2016) | 0.9377 (±0.0008) | 0.9616 (±0.0010) | 0.9396 (±0.0008) | 0.9457 (±0.0009) |
| Proposed Method (Resnet-50 based) | 0.9646 (±0.0007) | 0.9852 (±0.0009) | 0.9430 (±0.0011) | 0.9636 (±0.0013) |

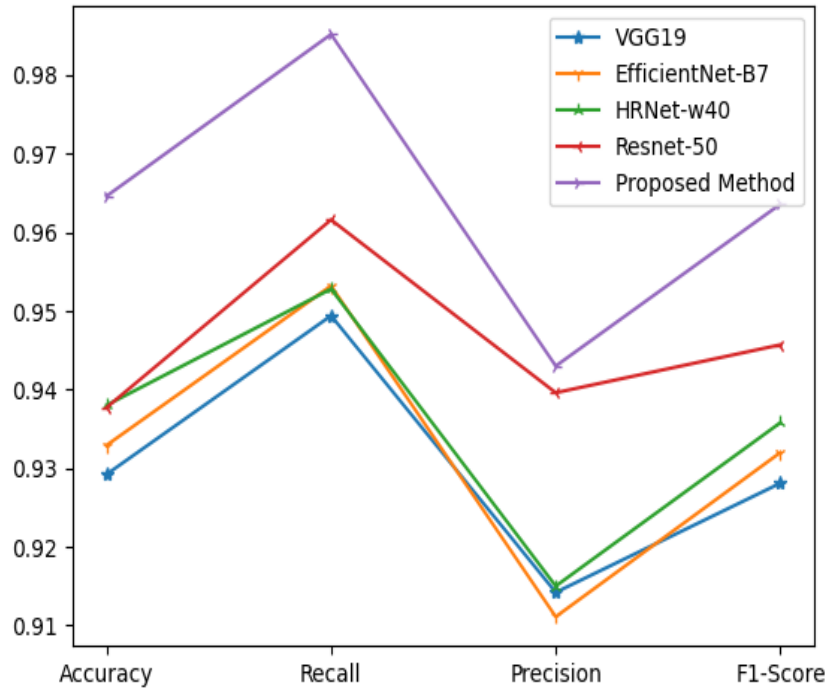


Figure 11. Graphical representation of Table 1

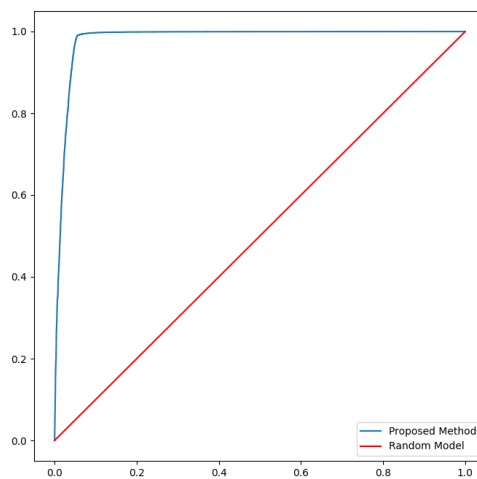


Figure 12. ROC Curve of the proposed network and random model

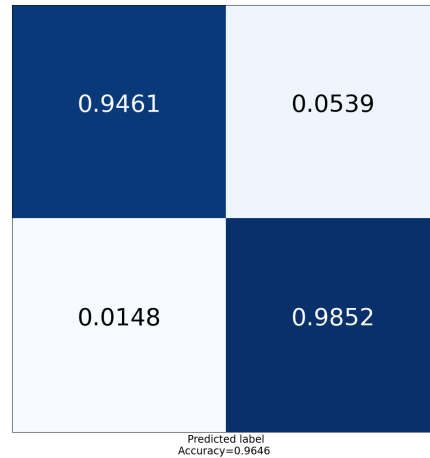


Figure 13. Confusion matrix derived from the experimental results

For the downstream task of classification of metastatic tissue, our proposed method produced the highest scores in all categories of accuracy, recall, precision, and F1-Score, as shown in Table 1. Figure 11 is a graphical representation of this data which emphasizes the difference in performance between different networks from the experiment, which also demonstrates the outperformance of the proposed network. Figure 12 shows the ROC curve, where the proposed network is comparatively closer to the positive y axis compared to a random model, which proves its high accuracy. The confusion matrix of Figure 13 also indicates that the proposed network generated a high level of accuracy, as shown in the darkness of the colors.

Object Detection

Table 2. Experimental results from the object detection task

| | Backbone | AP |
|--------------------------------------|--------------|------|
| YOLOv2 (Redmon and Farhadi 2017) | Darknet-19 | 21.4 |
| SSD (Liu et al. 2016) | Resnet-50 | 29.8 |
| EfficientDet-D0 (Tan et al. 2020) | Efficient-B0 | 32.9 |
| Faster R-CNN (Ren et al. 2015) | Resnet-50 | 34.8 |
| YOLOv3 (Redmon et al. 2018) | Darknet-53 | 35.4 |
| RetinaNet (Lin et al. 2017) | Resnet-50 | 38.1 |
| Proposed Method (RetinaNet based) | Resnet-50 | 46.9 |

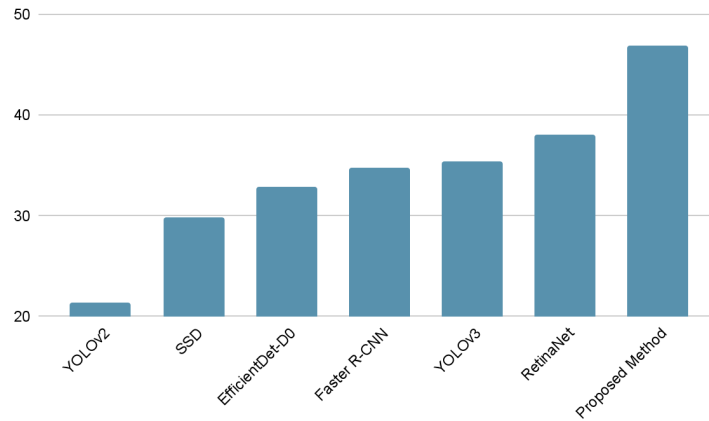


Figure 14. Graphical representation of Table 2

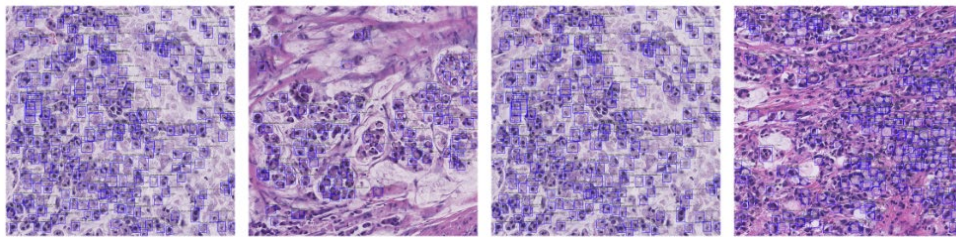


Figure 15. Example images of object detection from the proposed network

Object detection of SRCs also exceeded the state-of-the-art methods, producing an average precision in the IoU of 46.9, as shown in Table 2. Figure 14 illustrates the graphical depiction, where the bar for the proposed method is taller than any other bars. The whole slide images in Figure 15 depict the performance of the method, where blue boxes are the predictions of the network to be SRCs.

Segmentation

Table 3. Experimental results for the segmentation task

| | mIoU |
|--|------|
| PSPNet (Zhao et al. 2017) | 77.9 |
| Multipath-RetinaNet (Lin et al. 2017) | 80.2 |
| Resnet-38-MS-COCO (Wu et al. 2019) | 83.9 |
| DeepLabv3 (Chen et al. 2017) | 84.8 |
| Proposed Method | 86.4 |

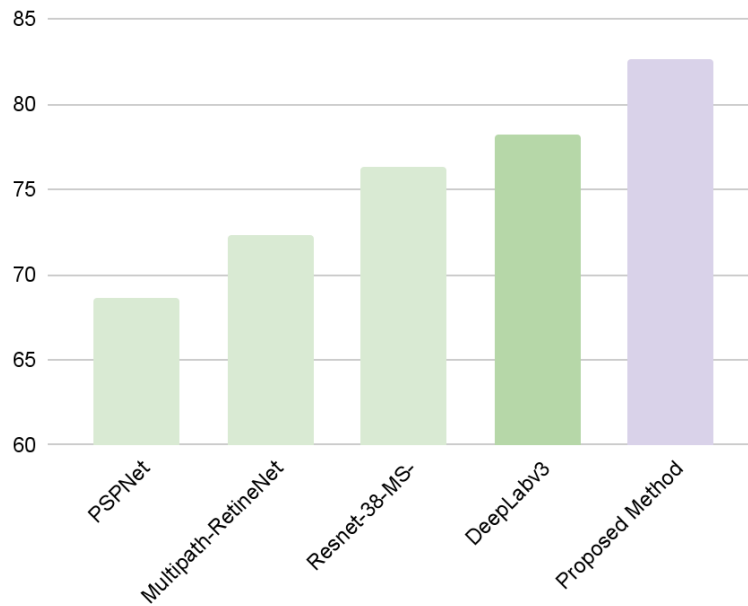


Figure 16. Graphical representation of Table 3

Table 3 shows that the downstream task of segmenting nuclei resulted in the mean IoU precision of 86.4, again surpassing previous methods. Figure 16 illustrates the difference of accuracy between different networks, where the bar in purple shows the proposed network with comparably high results.

Conclusion

In this paper, I proposed an artificial network that improves the efficiency and accuracy of the pathological analysis of tumors. The proposed method consists of a denoised AutoEncoder based representation learning and the transfer learning method. This allows the network to process multiple downstream tasks such as classification, segmentation, and object detection in a decreased time and increased accuracy compared to previous methods of pathological analysis performed by a human pathologist. From the four datasets of pathological images from various organs of the body, I carried out four distinct experiments addressing multiple downstream tasks. From the experiments, the proposed method achieved state-of-the-art results, outperforming previous approaches. In the future, I aim to apply this proposed network in real-world clinical settings, potentially reducing the strains and time consuming labor of pathologists with a higher time efficiency and precision.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Athanazio, D. A., Amorim, L. S., da Cunha, I. W., Leite, K. R. M., da Paz, A. R., de Paula Xavier Gomes, R., ... & Bezerra, S. M. (2021). Classification of renal cell tumors—current concepts and use of ancillary tests: recommendations of the Brazilian Society of Pathology. *Surgical and Experimental Pathology*, 4(1), 1-21.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587. <https://doi.org/10.48550/arXiv.1706.05587>

Halicek, M., Shahedi, M., Little, J. V., Chen, A. Y., Myers, L. L., Sumer, B. D., & Fei, B. (2019). Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. *Scientific reports*, 9(1), 14043.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). <https://doi.org/10.48550/arXiv.1512.03385>

Kim, H., Yoon, H., Thakur, N., Hwang, G., Lee, E. J., Kim, C., & Chong, Y. (2021). Deep learning-based histopathological segmentation for whole slide images of colorectal cancer in a compressed domain. *Scientific reports*, 11(1), 22520.

Lin, G., Milan, A., Shen, C., & Reid, I. (2017). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1925-1934). <https://doi.org/10.48550/arXiv.1611.06612>

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988). <https://doi.org/10.48550/arXiv.1708.02002>

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). Springer International Publishing. <https://doi.org/10.48550/arXiv.1512.02325>

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., ... & Stumpe, M. C. (2017). Detecting cancer metastases on gigapixel pathology images. arXiv preprint arXiv:1703.02442.

NVIDIA. (2023, Jul 13), “Whole Slide Image Analysis in Real Time with MONAI and RAPIDS”: NVIDIA. <https://developer.nvidia.com/blog/whole-slide-image-analysis-in-real-time-with-monai-and-rapids/>

R. Chandradevan, D. Chittajallu, L. Cooper, D. Gutman, M. McCormick, and A. Enquobahrie (2019, June 25). “Cell Nuclei Detection on Whole-Slide Histopathology Images Using HistomicsTK and Faster R-CNN Deep Learning Models”: Kitware. <https://www.kitware.com/cell-nuclei-detection-on-whole-slide-histopathology-images-using-histomicstk-and-faster-r-cnn-deep-learning-models>

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788). <https://doi.org/10.48550/arXiv.1506.02640>

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767. <https://doi.org/10.48550/arXiv.1804.02767>

- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28. <https://doi.org/10.48550/arXiv.1506.01497>
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. <https://doi.org/10.48550/arXiv.1409.1556>
- Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR. <https://doi.org/10.48550/arXiv.1905.11946>
- Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10781-10790). <https://doi.org/10.48550/arXiv.1911.09070>
- Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10), 3349-3364. <https://doi.org/10.48550/arXiv.1908.07919>
- Wang, J., Xu, Z., Pang, Z. F., Huo, Z., & Luo, J. (2021). Tumor detection for whole slide image of liver based on patch-based convolutional neural network. *Multimedia Tools and Applications*, 80, 17429-17440.
- Wu, Z., Shen, C., & Van Den Hengel, A. (2019). Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90, 119-133. <https://doi.org/10.48550/arXiv.1611.10080>
- Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881-2890). <https://doi.org/10.48550/arXiv.1612.01105>