HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 2 (2024)

MTWSI-Net: Towards Improved Multi-Task Whole Slide Image Classification with Contrastive Learning

Sean Hwang

The Hotchkiss School, USA

ABSTRACT

Pathology is a digital technology for the capture and analysis of high-resolution Whole Slide Images (WSIs). WSI entails scanning glass pathology slides, typically containing tissue samples, into a digital format. This facilitates comprehensive digital viewing, analysis, and interpretation by pathologists. Despite the undeniable utility of this method, the inherent time-consuming and labor-intensive nature of traditional pathology image analysis, which is heavily reliant on expert pathologists, is well-known. Recent years have seen a surge in research aiming to address these challenges through the development of automated systems using machine learning approaches. While promising, such systems often exhibit bias towards specific datasets and struggle to transition effectively to real-world scenarios. For this reason, there is a need to establish a uniform machine learning training approach for generating more robust and consistent results. In this paper, I introduce a contrastive learning-based multi-task whole slide image classification system. The proposed system excels at extracting consistently reliable features for identical cancer categories, thereby enhancing accuracy in downstream tasks. Through extensive experiments, the results demonstrate that the proposed system outperforms pre-existing state-of-the-art machine learning models. I expect that the proposed system can significantly contribute to pathologists by offering valuable cancer screening capabilities in WSIs.

Introduction

Pathology has witnessed a revolutionary shift with the adoption of Whole Slide Imaging (WSI). This innovative technology involves the digitization of traditional glass pathology slides which enables pathologists to engage in comprehensive digital analysis of tissue samples. Despite the benefits of WSIs, the traditional pathology process is a time-consuming and labor-intensive endeavor, as pathologists manually analyze WSIs. This reliance on subjective visual inspection poses challenges which hinders the efficiency of pathology workflows.

In response to the inherent limitations of traditional methods, recent years have seen the emergence of machine learning-based techniques designed to automate and expedite pathology image analysis. These approaches hold the promise of alleviating the burdensome nature of manual assessments and improving diagnostic accuracy. However, a critical challenge emerges in the application of these machine learning models to real-world scenarios. Many existing techniques exhibit a propensity to be biased towards specific training datasets. This limitation poses a significant barrier to the seamless integration of machine learning solutions into the real-world.

In this research paper, I propose a solution to tackle the bias issue in machine learning-based pathology image analysis. I introduce a novel contrastive learning-based multi-task WSI analysis system. The proposed system aims to not only enhance the efficiency and accuracy of pathology image analysis but also ensure its applicability in diverse and real-world clinical settings.

The subsequent chapters are structured as follows: Chapter 2 provides background knowledge essential for a comprehensive understanding of the proposed approach. In Chapter 3, every detail of the proposed system is elaborated upon, with comprehensive experimental results presented and analyzed in Chapter 4. The concluding insights are summarized in Chapter 5.

Related Work

Image Classification

Image classification is a task in computer vision and machine learning which aims to train a model to categorize images or objects into predefined classes or labels. The goal is to enable the system to automatically recognize and assign a label to a given image based on its visual content. Convolutional neural networks are commonly used for image classification due to their ability to automatically learn hierarchical features from images. They consist of convolutional layers that capture spatial hierarchies and pooling layers for feature reduction.

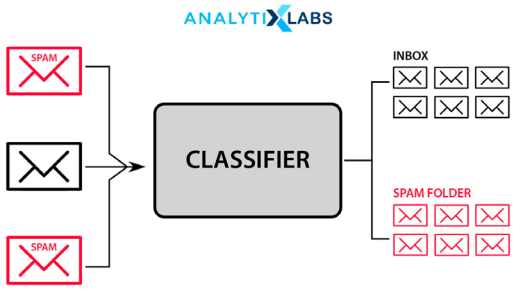

Figure 1. Example of object classification (Bansal 2023)

Due to the success of convolutional neural network-based image classification, its application in analyzing Whole Slide Images (WSI) in pathology has experienced rapid growth. These models effectively learn to discern visual patterns indicative of specific cancer types within WSI, yielding promising results for the detection of cancer cells across diverse human organs. Further details on this topic will be elaborated in the subsequent section, Chapter 2.1.

Whole Slide Images

Whole Slide Images (WSIs), also known as digital slides, refer to high-resolution digital representations of entire pathology slides. Traditional pathology involves the examination of tissue samples under a microscope, but with WSI, the entire slide is scanned and digitized. This digital transformation allows pathologists to view and analyze the entire specimen at varying magnifications on a computer screen, offering several advantages over conventional microscopy.

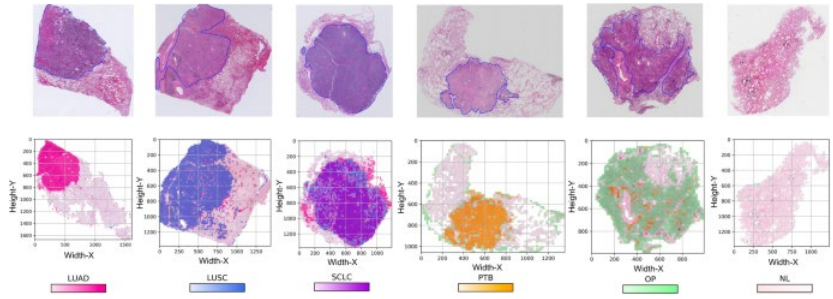

Figure 2. Example of WSI classification (Yang et al. 2021)

Classification system for categorizing types of lung cancer (Yang et al., 2021). Similarly, Sun et al. presented a cancer type classification system for the liver using a similar approach (Sun et al., 2019). Kuntz et al. developed a machine learning model specifically for the classification and prognostication of gastrointestinal cancer (Kuntz et al., 2021). While these methods have demonstrated promising results, a notable limitation lies in their pronounced bias toward specific datasets, rendering them less applicable to real-world scenarios. Therefore, there is a need to study and develop an uniform training approach to enhance the generalizability of trained models.

In this research, to solve this problem, I introduce a novel contrastive learning-based cancer type classification system. The detailed information of the proposed approach will be explained in Chapter 3.

Proposed Method

The proposed method comprises two distinct training phases. The initial phase focuses on representation learning which aims to train the feature extraction process to extract consistent features from diverse cancer images. In the second phase, this trained feature extractor is leveraged through transfer learning to enhance the efficiency and accuracy of downstream tasks. Both representation learning and transfer learning phases are illustrated in Figure 3 and Figure 4.

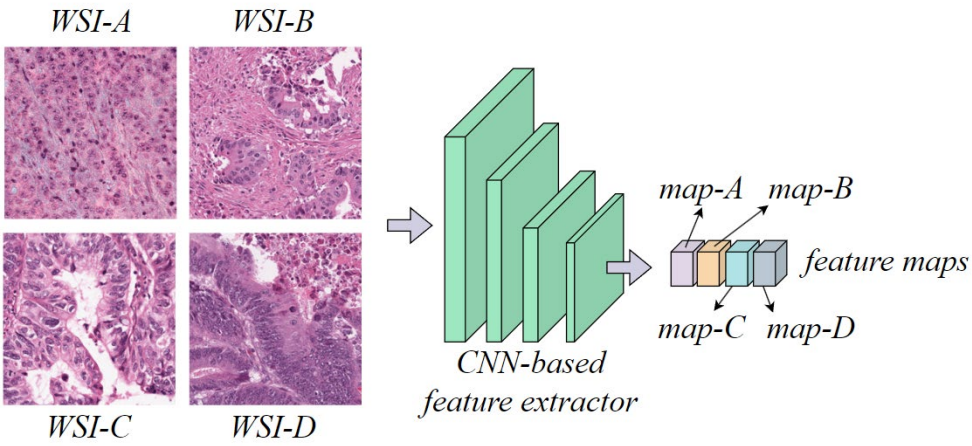

Figure 3. Architecture of the proposed representation learning approach

The convolutional neural network-based feature extractor processes the input of four pairs of Whole Slide Images (WSIs). These images are transformed into four distinct feature maps labeled as *map-a*, *map-b*,

map-c, and *map-d*. The objective is to maximize the similarity between two feature maps derived from WSIs belonging to the same cancer category. Despite potential variations in pixel-level visual patterns, the latent features are expected to exhibit similarity, given that they pertain to identical cancer types. To accomplish this, I utilize the cosine similarity function to quantify the similarity between each pair of feature maps and calculate the corresponding loss value. The step-by-step process is explained in Equations 1-3.

Equation 1: Cosine similarity function

$$S_{(\alpha,\beta)} = \frac{map_{\alpha} \cdot map_{\beta}}{|map_{\alpha}| \times |map_{\beta}|}$$

In the Equation 1, $S_{(a,b)}$ denotes the similarity index or score of two feature maps *map-a*, and *map-b*. The similarity score ranges from 0 to 1. A score of 0 signifies that two feature maps are diametrically opposed, whereas a score of 1 indicates that the two feature maps are identical. Three total score values are computed from the four feature maps. These scores are fed to the softmax function to generate probabilities, as explained in Equation 2.

Equation 2: Softmax function

$$P_{(\alpha,\beta)} = \frac{e^{S_{(\alpha,\beta)}}}{e^{S_{(\alpha,\beta)}} + e^{S_{(\alpha,c)}} + e^{S_{(a,d)}}}$$

In Equation 2, $P_{(a,\ b)}$ represents the converted probability of the similarity score $S_{(a,\ b)}$. Among the probabilities derived from all three score values, the probability associated with the score calculated from the same category should be maximized. To quantify the disparity between the prediction and ground truth, I employ the cross-entropy loss function commonly utilized in various classification tasks. The cross-entropy loss function is detailed in Equation 3.

Equation 3: Cross-entropy loss function

$$L = -log_e^{\square} P$$

Here, P denotes the predicted probability, and L denotes the calculated loss value. Throughout the training process, the model is optimized to minimize the loss value, thereby refining the parameters of the feature extractor to extract more consistent features from inputs of the same category. Following training, the pre-trained feature extractor is employed in the downstream task.

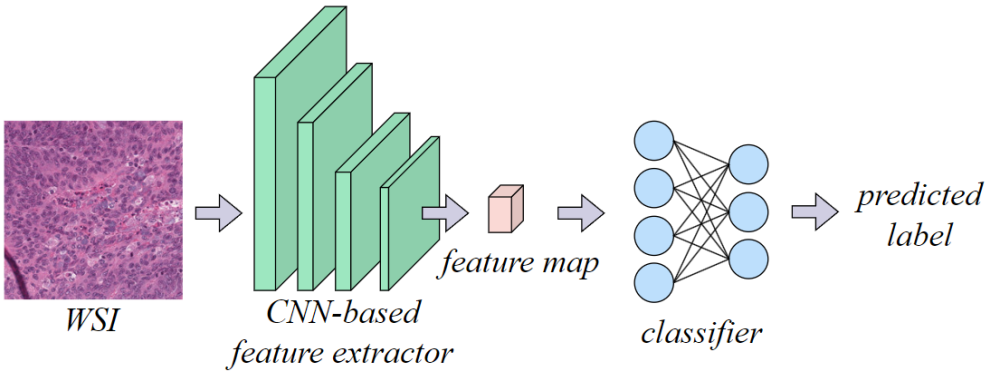

Figure 4. Architecture of the proposed transfer learning approach

Figure 4 illustrates the architecture of the proposed transfer learning approach. The pre-trained feature extractor receives a single WSI as input and generates a feature map, which is then input into the subsequent cancer type classifier. The classifier outputs the predicted cancer type. Instead of training this classification network from randomly initialized weights, I employ transfer learning, where the pre-trained feature extractor serves as the starting point for training weights. This training approach facilitates faster training and yields more accurate results. Additionally, the cross-entropy loss function is applied in training the model during transfer learning, with the mathematical equation being identical to Equation 3. The effectiveness of the proposed approach will be studied in Chapter 4.

Experimental Results

Dataset and Evaluation Metric

To train and evaluate the performance of the proposed model, I employed three distinct Whole Slide Image (WSI) classification datasets, including the Microsatellite Instable (MSI) vs. Microsatellite Stable (MSS) dataset from Nicolas (2019), the Metastatic Tissue dataset from Kaggle (2020), and the Breast Cancer dataset from Mooney (2017).

The Metastatic Tissue dataset utilized in this study comprises a total of 327,680 WSIs. The dataset is curated from histopathologic scans of lymph node sections. The primary focus of this dataset is the annotation of images with binary labels indicating the presence or absence of metastatic tissue.

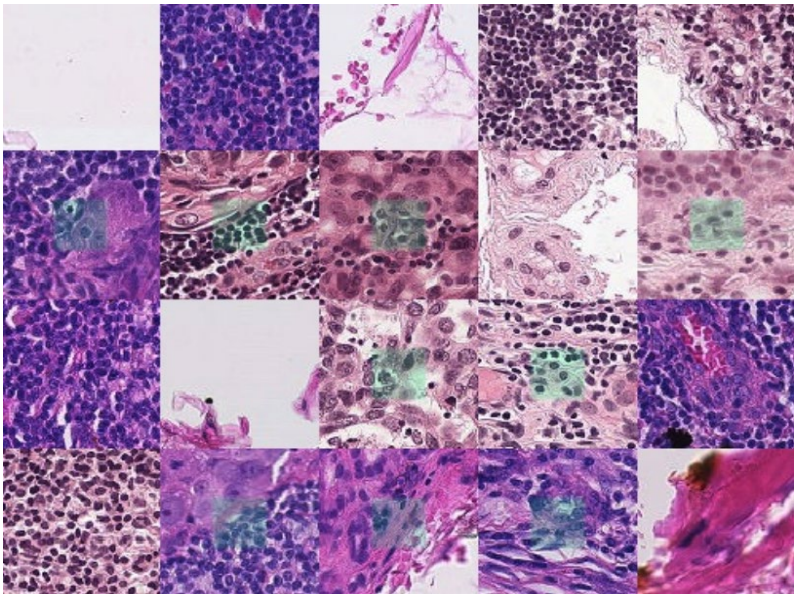

Figure 5. Metastatic Tissue dataset (Nicolas 2019)

The Breast Cancer dataset consists of 162 WSIs, each scanned at a magnification of 40x. A total of 277,524 patches were extracted from these whole mount slides. These patches are categorized into two classes: Invasive Ductal Carcinoma (IDC) negative (198,738 patches) and IDC positive (78,786 patches). IDC is the most common type of breast cancer, accounting for approximately 70-80% of all breast cancer diagnoses. It originates in the milk ducts of the breast but has the potential to invade surrounding tissues in the breast.

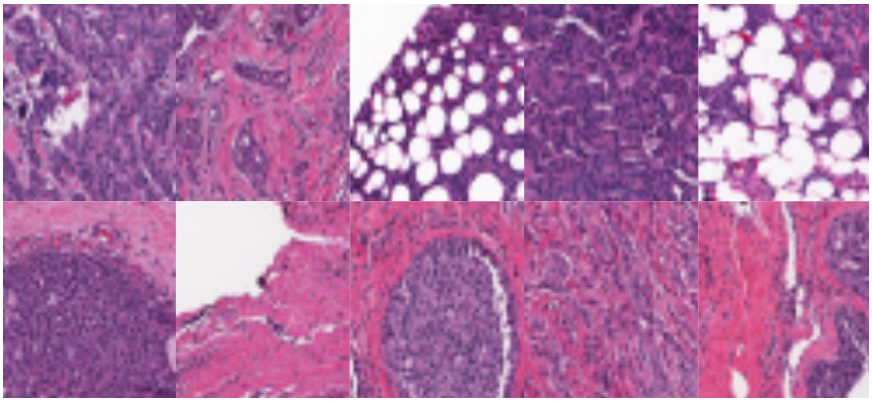

Figure 6. Breast Cancer dataset (Mooney 2017)

The MSI vs. MSS Classification dataset is sourced from histological images of colorectal cancer and gastric cancer patients. The dataset comprises 411,890 unique image patches. Microsatellites, also known as short tandem repeats or simple sequence repeats, are repetitive sequences of DNA scattered throughout the genome. They consist of short sequences of nucleotides repeated in tandem, and their stability is crucial for maintaining genomic integrity. MS) and MSS refer to conditions where there is a deviation from the normal stability of these repetitive sequences.

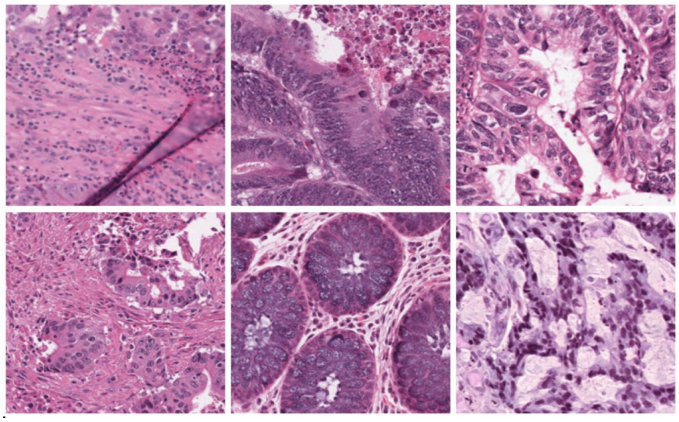

Figure 7. MSI vs. MSS dataset (Nicolas 2019)

To assess the effectiveness of the proposed method, I quantified its performance using four widely employed evaluation metrics: accuracy, recall, precision, and F1-score, as commonly applied in various classification tasks (Yacouby et al., 2020).

Evaluation

MSI vs. MSS Dataset

Table 1. Evaluation results on various network architectures (MSI vs. MSS dataset)

	Accuracy	**Recall**	**Precision**	**F1-Score**

VGG19	0.8983 (±0.0013)	0.8898 (±0.0011)	0.8659 (±0.0015)	0.8777 (±0.0014)
MobileNetV2	0.9079 (±0.0007)	0.8924 (±0.0007)	0.8712 (±0.0008)	0.8807 (±0.0013)
EfficientNet-B7	0.9087 (±0.0010)	0.8953 (±0.0008)	0.8699 (±0.0009)	0.8820 (±0.0011)
Xception	0.9134 (±0.0012)	0.8993 (±0.0015)	0.8801 (±0.0010)	0.8868 (±0.0012)
HRNet-w32	0.9200 (±0.0006)	0.9021 (±0.0011)	0.8834 (±0.0015)	0.8890 (±0.0009)
Resnet-50	0.9212 (±0.0012)	0.9055 (±0.0015)	0.8808 (±0.0016)	0.8935 (±0.0014)
Densenet-121	0.9208 (±0.0008)	0.9074 (±0.0011)	0.8831 (±0.0012)	0.8952 (±0.0011)
ResNeXt-101	0.9346 (±0.0005)	0.9203 (±0.0008)	0.8957 (±0.0008)	0.9078 (±0.0010)

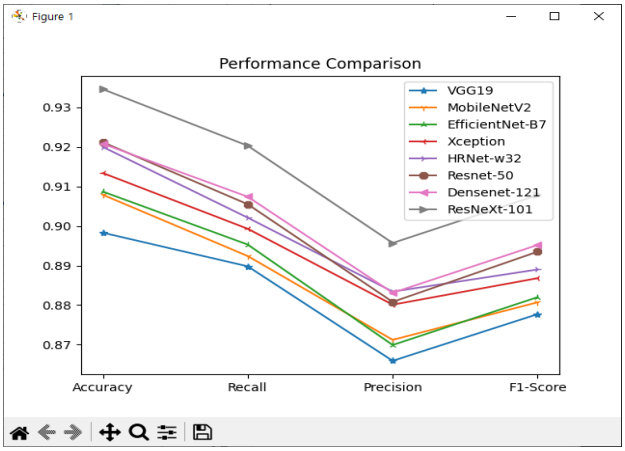

Figure 8. Evaluation results on various network architectures (MSI vs. MSS dataset)

Table 1 and Figure 8 provide a comprehensive summary of the evaluation results obtained across various network architectures using the MSI vs. MSS dataset. In the experimentation phase, I trained multiple network architectures including VGG19 (Simonyan et al. 2014), MobileNetV2 (Sandler et al. 2018), EfficientNet-B7 (Tan et al. 2019), Xception (Fran et al. 2017), HRNet-w32 (Wang et al. 2020), Resnet-50 (He et al. 2016), Densenet-121 (Huang et al. 2017), and ResNeXt-101 (Xie et al. 2017), leveraging the proposed representation learning approach.

VGG19, MobileNetV2, and EfficientNet-B7 exhibited relatively inaccurate results, possibly due to their shallower network depth. In contrast, both HRNet-32 and ResNet-50, equipped with deeper convolutional layers, demonstrated significantly improved efficacy. Notably, all evaluation results (accuracy) surpassed the 89% threshold, marking a remarkable achievement. The effectiveness of the proposed representation learning is elaborated in detail in Chapter 4.2.3.

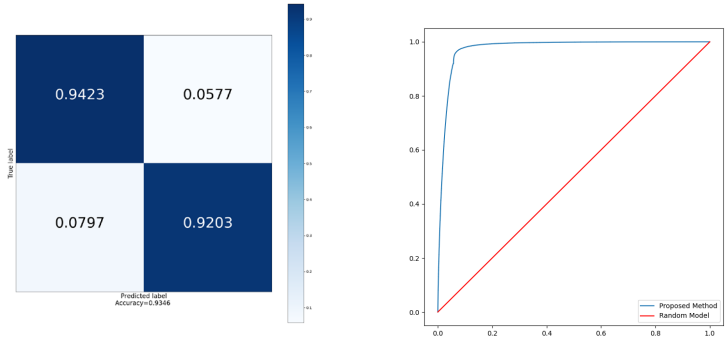

Figure 9. Confusion matrix and ROC curve (MSI vs. MSS dataset)

Figure 9 showcases the confusion matrix and the receiver operating characteristic (ROC) curve for the methods proposed in the MSI vs. MSS classification dataset. A close look at the diagonal elements of the confusion matrix offers compelling evidence of the stability and consistency realized by our proposed approach. Furthermore, the ROC curve's pronounced near-rectangular contour in the upper-left quadrant suggests that the method has outstanding results.

Metastatic Tissue Dataset

Table 2. Evaluation results on various network architectures (Metastatic Tissue dataset)

	Accuracy	**Recall**	**Precision**	**F1-Score**
VGG19	0.9359 (±0.0014)	0.9299 (±0.0013)	0.9024 (±0.0015)	0.9153 (±0.0011)
MobileNetV2	0.9374 (±0.0013)	0.9319 (±0.0011)	0.9009 (±0.0012)	0.9168 (±0.0010)
EfficientNet-B7	0.9381 (±0.0007)	0.9321 (±0.0009)	0.9022 (±0.0007)	0.9166 (±0.0006)
Xception	0.9424 (±0.0006)	0.9369 (±0.0008)	0.9060 (±0.0011)	0.9210 (±0.0007)
HRNet-w32	0.9442 (±0.0006)	0.9385 (±0.0007)	0.9149 (±0.0009)	0.9228 (±0.0011)
Resnet-50	0.9481 (±0.0012)	0.9424 (±0.0010)	0.9142 (±0.0013)	0.9298 (±0.0010)
Densenet-121	0.9599 (±0.0011)	0.9512 (±0.0013)	0.9512 (±0.0013)	0.9314 (±0.0011)
ResNeXt-101	0.9605 (±0.0010)	0.9548 (±0.0009)	0.9243 (±0.0013)	0.9393 (±0.0014)

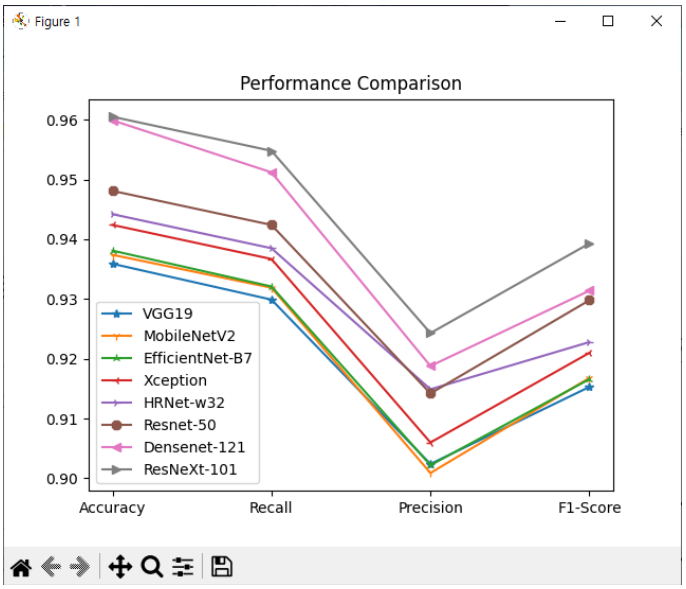

Figure 10. Evaluation results on various network architectures (Metastatic Tissue dataset)

In parallel with the initial experiment, I replicated the study on the Metastatic Tissue dataset to assess the applicability of the proposed approach to a distinct dataset. As depicted in Figures 5 and 7, the pixel distributions of the two datasets exhibit notable differences. The results, outlined in Table 2 and illustrated in Figure 10, reveal that all four evaluation metrics consistently achieve outstanding performance, surpassing the 90% threshold. This compelling outcome demonstrates that the proposed approach effectively addresses and mitigates the dataset bias issue encountered by previous methods, thereby significantly enhancing accuracy.

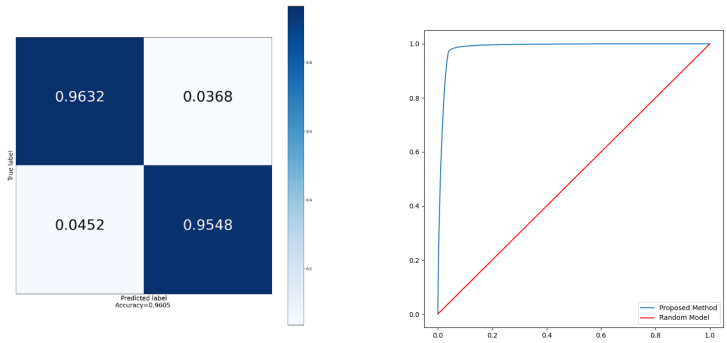

Figure 11. Confusion matrix and ROC curve (Metastatic Tissue dataset)

Figure 11 displays the confusion matrix and the ROC curve for the methods implemented on the Metastatic Tissue dataset. Similar to the observations in Figure 9, the results show the stability and consistency achieved by the proposed approach.

Ablation Study

To assess the effectiveness of the proposed approach, an ablation study was conducted. The experiment involved comparing the performance of two sets of convolutional neural network architectures. The first group

comprises ablation models trained without the proposed approach, relying solely on a supervised manner. The second group includes models trained using the proposed approach.

Table 3. Ablation study results (Breast Cancer dataset)

	Accuracy (ablation model)	**Accuracy (proposed method)**
VGG19	0.7802 (±0.0008)	0.8068 (±0.0011)
MobileNetV2	0.7834 (±0.0009)	0.8087 (±0.0014)
EfficientNet-B7	0.7895 (±0.0014)	0.8157 (±0.0013)
Xception	0.8267 (±0.0011)	0.8524 (±0.0011)
HRNet-w32	0.8480 (±0.0009)	0.8787 (±0.0012)
Resnet-50	0.8528 (±0.0013)	0.8842 (±0.0008)
Densenet-121	0.8810 (±0.0010)	0.9148 (±0.0011)
ResNeXt-101	0.8872 (±0.0007)	0.9224 (±0.0013)

Table 4. Ablation study results (Metastatic Tissue dataset)

	Accuracy (ablation model)	**Accuracy (proposed method)**
VGG19	0.9118 (±0.0011)	0.9359 (±0.0014)
MobileNetV2	0.9151 (±0.0014)	0.9374 (±0.0013)
EfficientNet-B7	0.9146 (±0.0009)	0.9381 (±0.0007)
Xception	0.9129 (±0.0012)	0.9424 (±0.0006)
HRNet-w32	0.9174 (±0.0008)	0.9442 (±0.0006)
Resnet-50	0.9129 (±0.0013)	0.9481 (±0.0012)
Densenet-121	0.9253 (±0.0016)	0.9599 (±0.0011)
ResNeXt-101	0.9287 (±0.0007)	0.9605 (±0.0010)

Table 3 provides a summary of the evaluation results for the Breast Cancer dataset, while Table 4 presents the results for the Metastatic Tissue dataset. Remarkably, the evaluation results of the proposed method consistently outperform the ablation model across all convolutional neural network architectures by a significant margin. This outcome demonstrates that the proposed approach effectively addresses the dataset bias issue inherent in supervised training methods. The superiority of these results can be attributed to the proposed contrastive learning-based representation learning which aims to train the model to extract latent patterns of cancer cells.

Conclusion

This research has aimed to revolutionize the landscape of pathology image analysis, specifically targeting Whole Slide Images (WSIs), by introducing a novel contrastive learning-based multi-task system. The experimental results demonstrated the efficacy of the proposed approach in enhancing the accuracy and reliability of whole slide image classification. The proposed system, leveraging contrastive learning techniques, has consistently outperformed existing state-of-the-art machine learning models, as demonstrated in the evaluation on both Breast Cancer and Metastatic Tissue datasets. Notably, the proposed system exhibited stability and consistency across diverse datasets, addressing the bias issues encountered by previous methods

and achieving remarkable performance metrics above 90%. This research contributes to the field in several key ways. First, it addresses the time-consuming and labor-intensive nature of traditional pathology processes by automating whole slide image classification. Second, it tackles the bias problem inherent in machine learning models which ensures their adaptability to real-world scenarios. Third, the proposed system provides a robust and consistent approach to pathology image analysis.

Acknowledgments

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

References

Bansal, S. (2023, Jan 25). "*Classification in machine learning: Types and methodologies*": Analytixlabs https://www.analytixlabs.co.in/blog/classification-in-machine-learning/

Fran, C. (2017). Deep learning with depth wise separable convolutions. In IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.48550/arXiv.1610.02357

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

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708). https://doi.org/10.48550/arXiv.1608.06993

Kaggle. (2020, Apr 20). "*Metastatic Tissue Classification - PatchCamelyon*": Kaggle https://www.kaggle.com/datasets/andrewmvd/metastatic-tissue-classification-patchcamelyon

Kuntz, S., Krieghoff-Henning, E., Kather, J. N., Jutzi, T., Höhn, J., Kiehl, L., ... & Brinker, T. J. (2021). Gastrointestinal cancer classification and prognostication from histology using deep learning: Systematic review. European Journal of Cancer, 155, 200-215.

Mooney P. (2017, Dec 19). "*Breast Histopathology Images*": Kaggle https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images

Nicolas, K. (2019, Feb 7). "*Histological images for MSI vs. MSS classification in gastrointestinal cancer, FFPE samples*": Zenodo https://zenodo.org/records/2530835

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520). https://doi.org/10.48550/arXiv.1801.04381

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

Sun, C., Xu, A., Liu, D., Xiong, Z., Zhao, F., & Ding, W. (2019). Deep learning-based classification of liver cancer histopathology images using only global labels. IEEE journal of biomedical and health informatics, 24(6), 1643-1651.

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

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
Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).
https://doi.org/10.48550/arXiv.1611.05431

Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In Proceedings of the first workshop on evaluation and comparison of NLP systems (pp. 79-91).

Yang, H., Chen, L., Cheng, Z., Yang, M., Wang, J., Lin, C., ... & Li, W. (2021). Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study. BMC medicine, 19, 1-14.