2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782037

Contrastive Pre-Training and Multiple Instance Learning for Predicting Tumor Microsatellite Instability

Ronald Nap[1], Mohammed Aburidi[1], and Roummel Marcia[1]

***Abstract*— Accurate classification between tumor MicroSatellite Stability (MSS) and Instability (MSI) is crucial in gastrointestinal (GI) cancer prognosis and treatment. In this paper, we present a novel two-stage weakly supervised methodology, leveraging the synergy of Multiple Instance Learning (MIL) and a unique Contrastive Clustering Network (CCNet), aimed at enhancing MSI prediction in Whole Slide Image (WSI) analysis of GI cancers. In our framework, we utilize a contrastive learning-based feature extractor, coupled with MIL's efficient labeling. Our approach shows notable improvement in MSI classification, outperforming existing methods. Experiments using colorectal cancer and stomach adenocarcinoma datasets demonstrate the model's efficacy and generalizability, marking an advance in computational pathology and cancer diagnostics. Furthermore, we explored the efficacy of transfer learning using our model, examining the performance of pretrained feature extractors from ImageNet and STAD datasets. Our framework outperforms existing methods when pretrained on STAD and transferred to CRC data.**

***Clinical relevance*— The potential to enhance the accuracy of gastrointestinal cancer diagnosis and prognosis through advanced machine learning techniques. By leveraging transfer learning and weakly supervised frameworks clinicians can benefit from improved MSI prediction in histopathological images, aiding in personalized treatment strategies and patient outcomes.**

I. INTRODUCTION

In 2023, the United States reported 348,840 new cases of gastrointestinal (GI) cancers, comprising 17.8% of all cancer cases, with 172,010 deaths (28.2%) [1]. Within GI cancers, distinguishing between MicroSatellite Stability (MSS) and Instability (MSI) is crucial. Tumor stability involves preserving a cancer cell's genome, focusing on microsatellites—regions of repetitive DNA [2]. Microsatellite stable tumors (MSS) retain microsatellites with minimal alterations, while microsatellite unstable tumors (MSI) exhibit instability due to mutations affecting DNA repeats [3]. This differentiation is vital, impacting cancer prognosis and treatment efficacy, including immunotherapy. MSI-H tumors often respond better to immunotherapy due to numerous mutation-associated neoantigens [4].

Traditionally, the distinction between MSS and MSI has heavily relied on the expertise of highly-trained pathologists and the involvement of sophisticated techniques such as next-generation sequencing or polymerase chain reaction (PCR)-based methods [5]. These techniques require specialized equipment and expertise, making them less accessible in some settings.

Recent years have seen a surge in the application of supervised deep learning techniques [6]–[8] in the analysis of Whole Slide Images (WSIs) in cancer pathology. While these methods have been successful, they face significant hurdles, including the requirement for extensive labeled data. This demand for labeled data, especially at the patch level for WSIs, proves to be a major challenge. Patch-level annotations are particularly time-consuming and costly to obtain, making them infeasible in many scenarios. Additionally, the high resolution of WSIs leads to substantial memory requirements, which further complicates the development of accurate and efficient deep learning models. These challenges highlight the need for more efficient annotation methods and more advanced computational techniques to handle the large size and complexity of WSIs.

Weakly supervised learning is a machine learning paradigm where training data is labeled at a higher, less granular level compared to traditional supervised learning, making it a cost-effective and scalable approach [9]. In the context of studying WSIs, a notable instance of weakly supervised learning is Multiple Instance Learning (MIL) [10]. MIL is particularly advantageous for WSIs as it allows for the labeling of bags of instances, where the specific labels for individual instances within the bags are either unknown or ambiguous [11]–[13]. This is especially relevant in pathology, where WSIs exhibit a high degree of variability in tissue types and structures. MIL enables the model to learn from loosely labeled data, reducing the need for exhaustive and expensive annotations for each instance in large datasets of WSIs, making it a practical and efficient strategy for studying WSIs.

In this paper, we introduce an innovative two-stage weakly supervised framework designed to enhance the accuracy of MSI prediction in WSI analysis. Drawing inspiration from established techniques [14]–[18], our approach seamlessly integrates contrastive learning and Multiple Instance Learning (MIL) to forecast tumor stability at the patch level within whole slide histopathology images. The proposed method demonstrates significant advancements in MSI prediction. A key aspect of our contribution lies in the effective fusion of MIL with a feature extractor trained through a Contrastive Clustering-based method, termed Contrastive Clustering Network (CCNet). This integration enhances the accuracy of predicting MSI, marking a substantial stride in the field. Our findings indicate that our two-stage model outperforms ex-

*This research is partially supported by NSF Grant IIS 1741490 and DMS 1840265.

[1]Ronald Nap, Mohammed Aburidi, and Roummel Marcia are with the Department of Applied Mathematics, University of California Merced, Merced, California, USA.

(Corresponding author: Mohammed Aburidi, email: maburidi@ucmerced.edu)

isting MIL classification techniques, and the incorporation of a contrastive clustering-based feature extractor significantly enhances feature extraction, improving overall performance beyond previous network designs. We also capitalize on pre-trained knowledge via transfer learning, allowing for superior performance even with minimal labeled data. Our results indicate good performance when models pretrained on Stomach Adenocarcinoma (STAD) are transferred to Colorectal Cancer (CRC).

II. Method

In this section, we present our framework, comprising two distinct phases. In the initial phase, we put forth a contrastive clustering model to extract latent feature vectors from patches in an unsupervised manner. Moving to the second phase, the extracted feature vectors are aggregated into bags and subsequently input into a classifier. This classifier, in turn, predicts the tumor microsatellite stability.

A. Contrastive Learning with CCNet

We introduce CCNet, a self-supervised learning model designed for feature extraction and overall improvements to model prediction. CCNet is used to significantly enhance the differentiation between similar data embeddings while simultaneously reducing the similarity among dissimilar ones in hopes of generating higher-quality feature representation. Unlike standard contrastive learning algorithms [19], CCNet utilizes a dual-headed contrastive architecture, which integrates both instance-level and cluster-level contrastive heads. This unique structure enhances the model's ability to extract intricate and comprehensive data representations, providing a more nuanced understanding of the data.

At the core of CCNet, similar to other contrastive learning methods [19], [20], is the Pair Construction Backbone (PCB). PCB generates pairs of data points through diverse augmentations, such as cropping, rotating, or color adjustments. This process is designed to maintain the core features of data points in the feature space despite significant visual transformations. The feature extraction module of CCNet then processes these pairs, developing robust features invariant to the applied augmentations. This methodology enables the model to recognize and interpret a wide array of data representations.

CCNet utilizes two types of contrastive heads: one at the instance level and another at the cluster level. The former enhances the model's capacity to recognize separate data instances. Let $\mathbf{Z}^a, \mathbf{Z}^b$ be the features extracted from the first head (the instance representations) for both views (a and b). Given a pair (i, j), they are contrasted using the following contrastive loss function:

$$\mathcal{L}^a_{I,i} = -\log \frac{\exp\left(\frac{s(\mathbf{z}^a_i, \mathbf{z}^b_i)}{\tau_I}\right)}{\sum_{j=1}^{N}\left\{\exp\left(\frac{s(\mathbf{z}^a_i, \mathbf{z}^a_j)}{\tau_I}\right) + \exp\left(\frac{s(\mathbf{z}^a_i, \mathbf{z}^b_i)}{\tau_I}\right)\right\}}, \quad (1)$$

where $s(\cdot,\cdot)$ is the pair-wise cosine distance, and $\mathbf{z}^a_i$ and $\mathbf{z}^b_i$ are two corresponding rows from the feature matrices $\mathbf{Z}^a$ and $\mathbf{Z}^b$, respectively. Here, τ_I is the instance-level temperature parameter [21] that is used to control the "softness" of this loss function.

Similarly, the second head outputs cluster-level representation and its also contrasted using cluster-level representation loss, which is given by the following formula

$$\mathcal{L}^a_{C,i} = -\log \frac{\exp\left(\frac{s(\mathbf{p}^a_i, \mathbf{p}^b_i)}{\tau_c}\right)}{\sum_{j=1}^{K}\left\{\exp\left(\frac{s(\mathbf{p}^a_i, \mathbf{p}^a_i)}{\tau_c}\right) + \exp\left(\frac{s(\mathbf{p}^a_i, \mathbf{p}^b_i)}{\tau_c}\right)\right\}}, \quad (2)$$

where $\mathbf{p}^a_i$ and $\mathbf{p}^b_i$ are two corresponding columns from the probability matrices $\mathbf{P}^a$ and $\mathbf{P}^b$, respectively. The matrices $\mathbf{P}^a$ and $\mathbf{P}^b$ are the soft labels, or the output probability matrices of the cluster contrastive head that are corresponding to the two views of the images. Here τ_c is the cluster-level temperature parameter. To include every possible positive pair across the dataset, the instance-level contrastive loss, and the cluster-level contrastive loss are as follows:

$$\mathcal{L}_I = \tfrac{1}{2N}\sum_{i=1}^{N}(\mathcal{L}^a_{I,i} + \mathcal{L}^b_{I,i})$$

$$\mathcal{L}_C = \tfrac{1}{2K}\sum_{i=1}^{K}(\mathcal{L}^a_{C,i} + \mathcal{L}^b_{C,i}) - S(\mathbf{P}),$$

where $S(\mathbf{P}) = -\sum_{i=1}^{K}[\mathbf{p}^a_i \log \mathbf{p}^a_i + \mathbf{p}^b_i \log \mathbf{p}^b_i]$ is the entropy of cluster assignment probabilities added to prevent assigning all instances within the mini-batch to the same cluster [22]. The functions $\mathcal{L}^b_{I,i}$ and $\mathcal{L}^b_{C,i}$ are defined similarly as in (1) and (2), respectively. By incorporating both levels, CCNet can identify and perceive larger patterns and connections present in the dataset.

$$L = \mathcal{L}_I + \mathcal{L}_C. \quad (3)$$

The presence of a cluster head ensures that CCNet can effectively capture details in a balanced manner, significantly improving the training phase. As depicted in Equation (1), the combination of $\mathcal{L}_I$, the instance-level contrastive loss, and $\mathcal{L}_C$ representing the cluster-level contrastive loss allows for the optimization of specific details through the instance-level while capturing comprehensive data trends via the cluster-level.

B. Multiple Instance Learning

MIL is a variation of weakly supervised learning particularly suited for scenarios where there is uncertainty in labeling individual data points, but labels are available at a group or bag level. In the context of WSI analysis, MIL demonstrates remarkable effectiveness due to the complexities and expansiveness of these images. By treating each WSI as a collection of instances, MIL makes diagnoses based on the combined characteristics of these patches, eliminating the necessity for annotations of every single pixel. By attaching labels to aggregations of tiles instead, MIL improves upon conventional methods [23] that attempt to extend slide-level labels to smaller sections or tiles, which could lead to

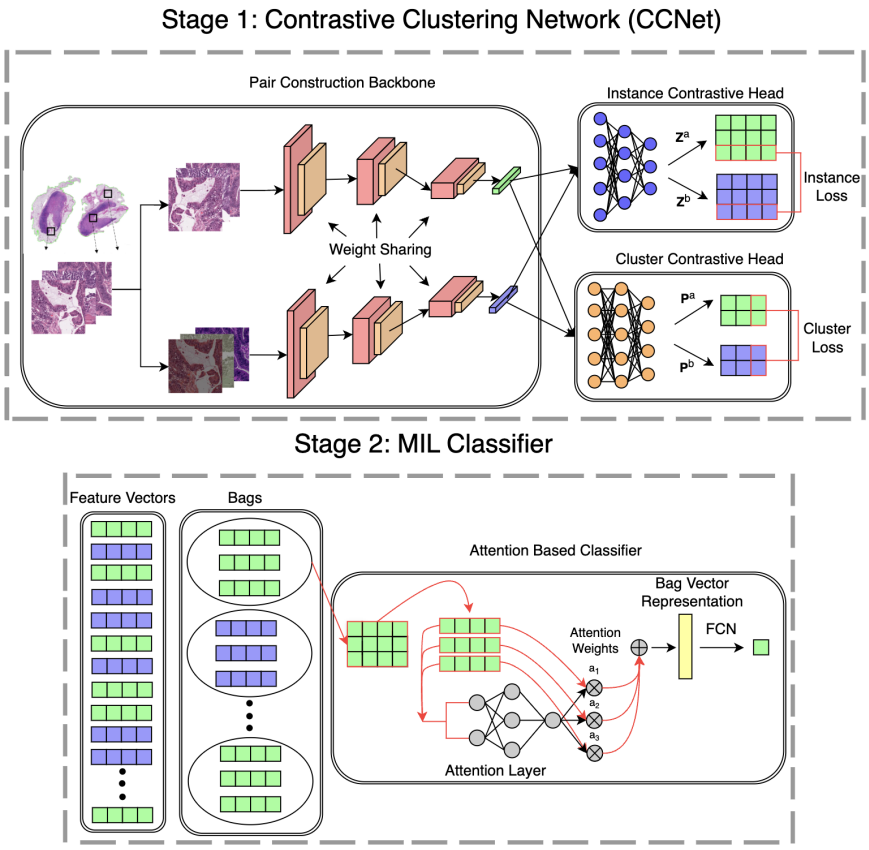


Fig. 1. Stage 1 of our framework features CCNet, starting with data pair creation via a PCB. These pairs undergo feature extraction through a ResNet18 shared encoder. Next, two Multi-Layer Perceptrons (MLPs) project features into row and column spaces for contrastive learning. Post-training, the cluster head is removed, and feature vectors are extracted from patches. Stage 2 involves the MIL Classifiers architectures which constructs bags from the extracted feature vectors from Stage 1. Each instance within these bags is assigned an attention weight determined through a two-layer MLP. Subsequently, these weights are normalized and a bag representation is formed by calculating a weighted average of each instance. Finally, this aggregated bag representation is used in the final classification decision

inaccuracies since these subregions might not share identical properties.

Our methodology in MIL for WSI analysis differs from conventional practices that typically generate bags based on specific contextual information, like spatial proximity or morphological similarities. Instead, we opt for a strategy of random bag creation. This approach is designed to capture a wider array of features within each WSI. By integrating patches from various regions of the slide into each bag without restricting to similarities or proximity, we ensure a diverse composition in each bag. This randomness in selection is key to encompassing a wide spectrum of tissue types. Overall, this will allow the ability to strengthen the robustness of our model by exposing it to a diverse range of instances and reduce the likelihood of overfitting to particular patterns that may recur across slides.

DeepMIL and VarMIL are 2 state-of-the-art MIL classifiers designed to effectively handle weakly labeled data [16], [24]. DeepMIL utilizes a trainable attention mechanism to concentrate on the most informative instances within each bag. This process involves an MLP attention network equipped with parameters $\mathbf{W}$, $\mathbf{V}$, and $\mathbf{U}$ to allocate a weight a_k for each embedded instance $\mathbf{z}_k$. The attention weights a_k are given by,

$$a_k = \frac{\exp\{\mathbf{W}^\top(\tanh(\mathbf{V}\mathbf{z}_k^\top) \odot \text{sigm}(\mathbf{U}\mathbf{z}_k^\top))\}}{\sum_{j=1}^{N}\exp\{\mathbf{W}^\top(\tanh(\mathbf{V}\mathbf{z}_j^\top) \odot \text{sigm}(\mathbf{U}\mathbf{z}_j^\top))\}} \tag{4}$$

where $\text{sigm}(\cdot)$ is the sigmoid function, and $\odot$ is the element wise product, N is the number of the embedded instance vectors in a bag. In DeepMIL, the overall bag representation z is computed by the weighted mean μ of these instance embeddings as follows

$$\mu = \frac{1}{N}\sum_{k=1}^{N} a_k \mathbf{z}_k \tag{5}$$

VarMIL extends this approach by also considering the variance of instances within each bag. The overall bag representation z in VarMIL is computed by concatenating the weighted mean μ and the square root of the weighted variance σ^2:

$$z = [\mu, \sigma^2] \tag{6}$$

where

$$\sigma^2 = \sqrt{\frac{1}{N}\sum_{k=1}^{N} a_k(\mathbf{z}_k - \mu)^2} \tag{7}$$

After obtaining the bag representation, a linear layer (followed by a softmax activation and argmax) is applied to determine whether a bag is positive or negative (e.g., MSI), based on the collective features.

C. Our Model

In our study, we propose a novel methodology that combines a two-stage process of weakly supervised learning for robust feature extraction through Contrastive Clustering, and classification using MIL with DeepMIL and VarMIL. The core architecture of our model is illustrated in Figure (1). In the training phase, a contrastive clustering network undergoes training on small image patches extracted from WSIs. Once trained, we use CCNet to encode patches from a specific WSI, producing embedded latent vector representations. These representations form the basis for constructing a bag corresponding to that WSI. Each bag is assembled by randomly selecting a set of embedded vectors associated with the WSI, which are then combined and utilized as input for stage 2. Consequently, multiple bags can be constructed from the same WSI. These bags are then utilized for classifier training purposes.

To prevent overfitting, our MIL classifiers adopt a Smooth Support Vector Machine (SVM) [25] loss complemented by Kullback–Leibler (KL-divergence) [26] regularization. The smooth SVM loss is formulated to accommodate the unique structure of data represented in terms of bags and instances and defined as follows. First, let A_j represent the j-th bag in MIL with associated label y_j. Next, let $f(A_j)$ be a function applied to the entire bag A_j that generates predictions or scores for each instance within the bag. Denote these instance predictions by ζ_i. We define $\xi_i = \max(0, 1-\zeta_i \times y_j)$, which captures the margin violation for each instance, promoting a margin of at least 1 between the logits ζ_i and their correspodning labels y_j. We define the smooth SVM loss for bag j by

$$l(y_j, f(A_j), \delta) = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2\delta}\xi_i^2 & \text{if } \xi_i \leq \delta \\ \frac{1}{N}\sum_{i=1}^{N} \xi_i - \frac{\delta}{2} & \text{otherwise} \end{cases},$$

where the smoothness parameter $\delta > 0$ ensures that the loss function remains differentiable. The total smooth SVM loss function is given by

$$\text{Smooth SVM Loss} = \frac{1}{M}\sum_{j=1}^{M} l(y_j, f(A_j), \delta) \tag{8}$$

where M denotes the total number of bags.

The KL Divergence component acts as a regularizer by encouraging the attention distribution A_{ji} within each bag to approximate a uniform distribution U_{ji}. This ensures that the model does not disproportionately focus on a few instances but rather considers the entire bag, aligning with the MIL paradigm. The KL Divergence is given by

$$\text{KL Divergence} = \frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N} A_{ji}\log\left(\frac{A_{ji}}{U_{ji}}\right) \tag{9}$$

III. EXPERIMENTS

Datasets: The evaluation of our model was conducted using two intricate image datasets: the Colorectal Cancer (CRC) dataset and the Stomach Adenocarcinoma (STAD) dataset, both obtained from the TCGA cohort [27]. The CRC dataset was utilized for comparative analysis [23], [24], whereas the STAD dataset was employed to externally validate our model's efficacy. We partitioned the training data into k-folds for cross-validation and reserved the testing split for final validation. Detailed descriptions of these datasets are provided in Table I.

TABLE I
SUMMARY OF CRC AND STAD DATASET FOR ENCODER AND CLASSIFIER TRAINING/TESTING

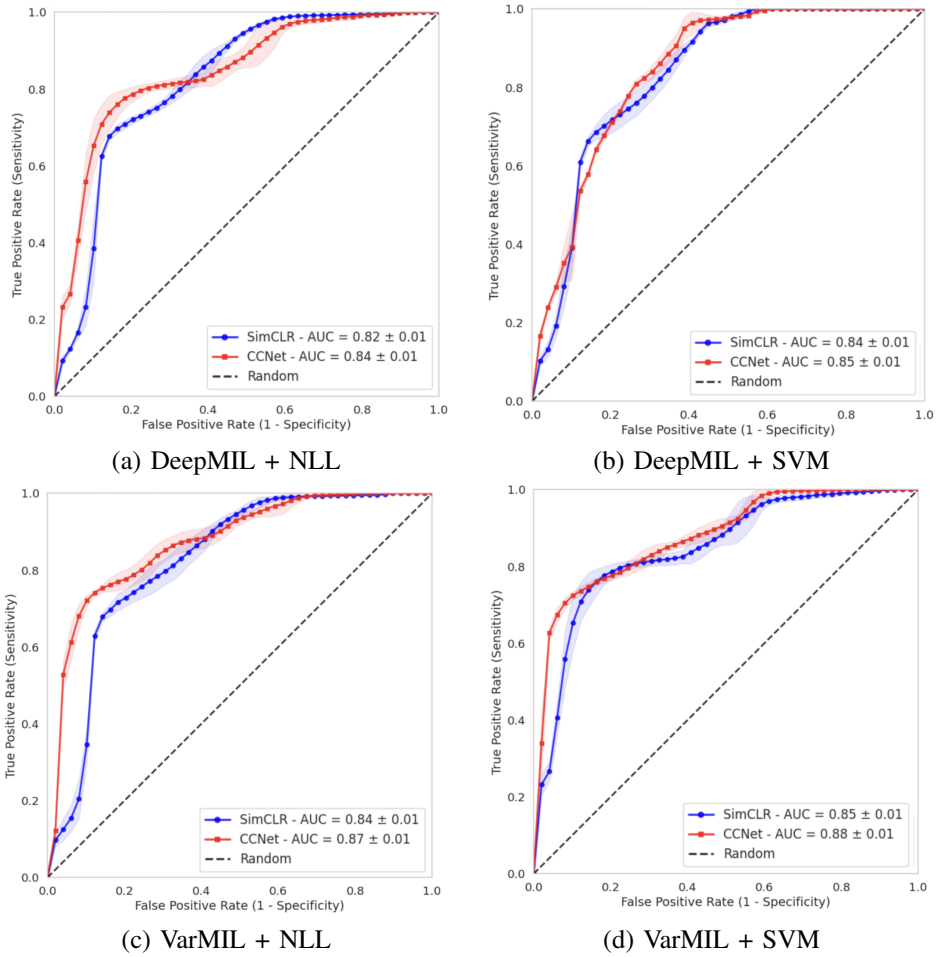
Dataset	Label	# of WSIs		# of Patches		# of Bags	
		Train	Test	Train	Test	Train	Test
CRC	MSI	39	26	46,704	29,335	1850	1122
	MSS	221	74	46,704	70,569	1757	2787
STAD	MSI	35	25	50,285	27,904	N/A	N/A
	MSS	150	74	50,285	90,104	N/A	N/A

CCNet Implementation: Our model employs ResNet18 [28] as the encoder backbone architecture. To optimize both the projection heads and the backbone network concurrently, we utilize the Adam optimizer [29], setting the learning rate at 0.0003. The dimension of the latent vector is fixed at 128, and the temperature parameters are set at 0.5. ReLU activation was used in between the two layers. Softmax activation was used in the cluster-level contrastive projection head to produce soft labels. We utilize a batch size of 256, which spans 100 epochs, starting from scratch. The training is carried out on UC Merced's Pinnacles Cluster, utilizing two units of 2x NVIDIA Tesla A100 PCIe v4 40GB HBM2 GPU.

Data Augmentations: Following established methods [17], [19], we incorporate a variety of augmentation techniques, including random cropping, color jittering, grayscale conversion, and horizontal flipping.

MIL Classifier Implementation: The bag size is set to 25. Early stopping is incorporated into the training process. A dropout rate of 0.5 is applied to both attention and classification layers and a batch size of 2 is employed. Model optimization is done using the Adam optimizer with a learning rate of 0.0001 and a weight decay of 0.0001.

Evaluation Metrics: For each cross-validation fold, we select the best model based on validation loss and assess testing performance on the model that achieved the lowest validation loss. We then compute the mean and standard deviation of both AUROC and F1 scores.



(a) DeepMIL + NLL (b) DeepMIL + SVM

(c) VarMIL + NLL (d) VarMIL + SVM

Fig. 2. Comparative receiver operating characteristic (ROC) curves. The plots present the true positive rate (on the y axis) against the false positive rate (on the x axis) for SimCLR (blue) and CCNet (red) feature extractors when combined with VarMIL and DeepMIL using both NLL and SVM loss functions. The corresponding Area Under the Curve (AUC) values are presented.

IV. RESULTS

A. Comparative study

We performed a comparative analysis of the performance of CCNet against SimCLR, which is recognized as a well-established contrastive learning framework [19] within a MIL framework on the TCGA-CRC dataset. CCNet, combined with VarMIL and SVM loss, outperformed SimCLR, achieving an AUROC of 0.88 ± 0.01 and an F1 score of 0.84 ± 0.01, as shown in Figure (2) and Figure (3). In contrast, SimCLR's best performance, using VarMIL with SVM loss, yielded an AUROC of 0.85 ± 0.01 and an F1 score of 0.80 ± 0.01. We employed DeepMIL which revealed consistent superiority of CCNet in both AUROC and F1 metrics. Our findings also highlight a notable improvement in performance of utilizing SVM loss compared to the standard negative log-likelihood loss (NLL).

In comparison to the results reported in previous literature [23], [24], we observed similar AUROC values but obtained higher F1 scores. Given the nature of an unbalanced testing set, we don't find this significant. The low standard deviation across various configurations suggests robustness and consistency in our models. We believe this is attributed to rigorous regularization strategies implemented in our classifiers, which helped in maintaining generalizability across various folds.

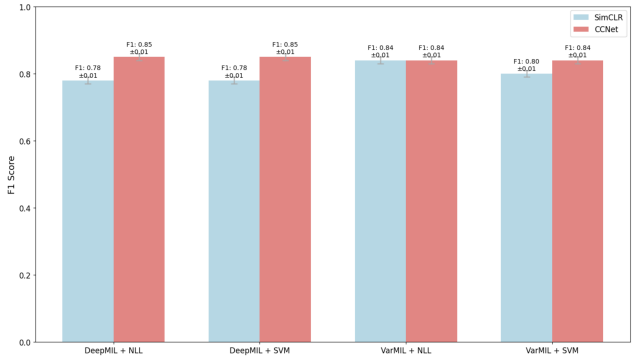


Fig. 3. Comparative F1 Scores. This chart presents the F1 scores of CCNet against the F1 scores for SimCLR when combined with VarMIL and DeepMIL using both NLL and SVM loss functions.

B. In- and out-of-domain transfer learning

In the domain of medical imaging, the scarcity of labeled data presents significant challenges for training machine learning models. To mitigate this, we employed transfer

learning, allowing us to capitalize on the knowledge acquired from larger and more diverse datasets. In addition to our initial findings, we delved deeper into the potential of our model in the context of transfer learning, as detailed in Table II. We first investigate the use of an out-of-domain pretrained feature extractor, specifically utilizing ResNet18 pretrained on ImageNet [30]. Following this, we assessed the performance of SimCLR and CCNet, both pretrained on the STAD dataset, these pretrained feature extractors are then transferred and used to generate feature vectors used in the second stage on the CRC data. Our findings indicate that the ImageNet-pretrained model yielded suboptimal results. In contrast, CCNet demonstrated superior performance compared to SimCLR across model configurations. This comparative analysis suggests that CCNet's feature extraction capabilities are more robust and generalizable than SimCLR's, resulting in a level of performance that rivals that achieved through direct pretraining on the CRC dataset.

TABLE II

IN AND OUT-OF-DOMAIN TRANSFER LEARNING RESULTS.

Extractor	Classifier	Loss	AUROC	F1
ResNet18	DeepMIL	NLL	0.58 ± 0.01	0.67 ± 0.01
		SVM	0.58 ± 0.01	0.69 ± 0.01
	VarMIL	NLL	0.59 ± 0.01	0.69 ± 0.01
		SVM	0.59 ± 0.01	0.69 ± 0.01
SimCLR	DeepMIL	NLL	0.81 ± 0.01	0.82 ± 0.01
		SVM	0.82 ± 0.01	0.81 ± 0.01
	VarMIL	NLL	0.78 ± 0.01	0.82 ± 0.01
		SVM	0.81 ± 0.01	0.81 ± 0.01
CCNet	DeepMIL	NLL	0.81 ± 0.01	0.81 ± 0.01
		SVM	0.83 ± 0.01	0.82 ± 0.01
	VarMIL	NLL	0.81 ± 0.01	0.83 ± 0.01
		SVM	0.83 ± 0.01	0.81 ± 0.01

V. CONCLUSIONS

In this paper, we propose a novel contrastive and MIL-based framework for predicting tumor microsatellite instability, a challenging task that relies on the expertise of highly-trained pathologists and the involvement of sophisticated techniques such as next-generation sequencing. Our proposed framework demonstrates superior performance, evidenced by higher AUROC and F1 scores across multiple configurations, compared to the state-of-the-art method designed for this task, namely SimCLR. A key finding from our study is the superior efficacy of SVM loss over NLL loss, which further enhances the performance of our models. These results underscore the potential of our proposed method in advancing the field of medical image analysis, particularly in the accurate classification of tumor instability. Future studies, including further validations in independent cohorts, are crucial to fully establish the advantages of CCNet and its contribution to improving diagnostic accuracy.

REFERENCES

[1] R. L. Siegel et al., "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, p. 4, 2023. doi: 10.3322/caac.21763.

[2] K. Li, H. Luo, L. Huang, H. Luo, and X. Zhu, "Microsatellite instability: a review of what the oncologist should know," *Cancer Cell International*, vol. 20, no. 16, 2020. doi: 10.1186/s12935-019-1091-8.

[3] R. Gupta et al., "The impact of microsatellite stability status in colorectal cancer," *Curr Probl Cancer*, vol. 42, no. 6, pp. 548-559, Nov. 2018. doi: 10.1016/j.currproblcancer.2018.06.010.

[4] I. H. Sahin, M. Akce, O. Alese, W. Shaib, G. B. Lesinski, B. El-Rayes, and C. Wu, "Immune checkpoint inhibitors for the treatment of MSI-H/MMR-D colorectal cancer and a perspective on resistance mechanisms," *British Journal of Cancer*, vol. 121, no. 10, pp. 809–818, 2019. doi: 10.1038/s41416-019-0599-y.

[5] F. Dedeurwaerdere, K. B. M. Claes, J. Van Dorpe, I. Rottiers, J. Van der Meulen, J. Breyne, K. Swaerts, and G. Martens, "Comparison of microsatellite instability detection by immunohistochemistry and molecular techniques in colorectal and endometrial cancer," *Scientific Reports*, vol. 11, 2021. doi: 10.1038/s41598-021-91974-x.

[6] J. van der Laak, G. Litjens, and F. Ciompi, "Deep learning in histopathology: the path to the clinic," *Nature Medicine*, vol. 27, pp. 775–784, 14 May 2021. doi: 10.1038/s41591-021-01343-4.

[7] N. Coudray et al., "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, pp. 1559–1567, 2018. doi: 10.1038/s41591-018-0177-5.

[8] M. Aburidi and R. Marcia, 'Adversarial Attack and Training for Graph Convolutional Networks using Focal Loss-Projected Momentum', in 2024 3rd IEEE International Conference on Computing and Machine Intelligence (ICMI), 2024.

[9] A. Echle et al., "Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning," *Gastroenterology*, vol. 159, pp. 1406–1416, 2020. doi: 10.1053/j.gastro.2020.06.021.

[10] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018. Elsevier.

[11] Z. Shao et al., "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[12] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, "Neural Image Compression for Gigapixel Histopathology Image Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 567–578, Feb. 2021. doi: 10.1109/TPAMI.2019.2936841.

[13] W. Aswolinskiy, D. Tellez, G. Raya, L. van der Woude, M. Looijen-Salamon, J. van der Laak, K. Grunberg, and F. Ciompi, "Neural image compression for non-small cell lung cancer subtype classification in H&E stained whole-slide images," in *Medical Imaging 2021: Digital Pathology*, J. E. Tomaszewski and A. D. Ward, Eds. SPIE Conference Series, vol. 11603, Feb. 2021, p. 1160304. doi: 10.1117/12.2581943.

[14] M. Aburidi and R. Marcia, 'CLOT: Contrastive Learning-Driven and Optimal Transport-Based Training for Simultaneous Clustering', in 2023 IEEE International Conference on Image Processing (ICIP), 2023, pp. 1515–1519.

[15] M. Aburidi and R. Marcia, 'Optimal Transport and Contrastive-Based Clustering for Annotation-Free Tissue Analysis in Histopathology Images', in 2023 International Conference on Machine Learning and Applications (ICMLA), 2023, pp. 301–307.

[16] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," 2018, arXiv:1802.04712.

[17] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive Clustering," 2020, arXiv:2009.09687.

[18] M. Aburidi and R. Marcia, 'Wasserstein Distance-Based Graph Kernel for Enhancing Drug Safety and Efficacy Prediction *', in 2024 IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC), 2024, pp. 113–119.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *CoRR*, vol. abs/2002.05709, 2020. [Online]. Available: https://arxiv.org/abs/2002.05709.

[20] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," 2019, arXiv:1807.03748.

[21] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination," *CoRR*, vol. abs/1805.01978, 2018. [Online]. Available: http://arxiv.org/abs/1805.01978.

[22] W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama, "Learning Discrete Representations via Information Maximizing Self-Augmented Training," *arXiv preprint arXiv:1702.08720*, 2017.

[23] J. N. Kather et al., "Deep learning can predict microsatellite instability

directly from histology in gastrointestinal cancer," *Nature Medicine*, vol. 25, pp. 1054–1056, 2019. doi: 10.1038/s41591-019-0462-y.

[24] Y. Schirris, E. Gavves, I. Nederlof, H. M. Horlings, and J. Teuwen, "DeepSMILE: Contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer," *Medical Image Analysis*, vol. 79, Jul. 2022, p. 102464. doi: 10.1016/j.media.2022.102464.

[25] L. Berrada, A. Zisserman, and M. P. Kumar, "Smooth Loss Functions for Deep Top-k Classification," *CoRR*, vol. abs/1802.07595, 2018. [Online]. Available: http://arxiv.org/abs/1802.07595.

[26] R. Mondal, P. Dey, G. Sharma, and T. Pal, "Regularizing Multilayer Perceptron for Generalization Using KL-Divergence," in *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, pp. 1-6, 2020. doi: 10.1109/ICCSEA49143.2020.9132891.

[27] J. N. Kather, "Histological images for MSI vs. MSS classification in gastrointestinal cancer, FFPE samples," Feb. 2019, Zenodo. doi: 10.5281/zenodo.2530835.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016. doi: 10.1109/CVPR.2016.90.

[29] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2017, arXiv:1412.6980.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, 'Imagenet: A large-scale hierarchical image database', in 2009 IEEE conference on computer vision and pattern recognition, 2009, pp. 248–255.