Assessment of Medical Foundation Models for Survival Prediction with Whole Slide Images

Elena Spirina Menand
Unité de Génomique Fonctionnelle,
Institut de Cancérologie de l'Ouest,
Angers, 49055, France
elena.menand@ico.unicancer.fr

Abstract

Survival prediction with whole slide images (WSIs) can provide guidance for a better patient care and treatment selection but it is a challenging computer vision task with its particularities. Despite the great results showed by the recent survival analysis models with WSIs, the collection of the large annotated WSI datasets for survival analysis could be hindered by disease rareness or clinical trials constraints and be infeasible in the real-life medical practice. To overcome these limitations we propose to assess the performance of the digital pathology foundation models for prediction of survival outcomes on the small size ovarian cancer datasets. Our experimental results demonstrate that these models show promising results, their improved performance open the possibility to investigate the mechanisms of response to a particular therapy and in general could accelerate the adoption of machine learning models in medical practice.

1 Introduction

Whole Slide Imaging (WSI) technology has contributed to the growing availability of digital pathology datasets, opening new complex computer vision research opportunities. Modern progress in deep learning has shown impressive results in various clinical applications, especially with the recent advent of attention-based models [1].

However, WSI needs special approaches for supervised learning tasks such as survival prediction. In classical computer vision with natural images, the label is usually assigned to 256 x 256 pixels images, whereas WSIs can be very heterogeneous and as large as 150,000 x 150,000 pixels. Acquiring exhaustive localized annotations for WSIs is expensive and often infeasible, thus, the multiple instance learning (MIL) has been widely adopted for WSI based tasks. In this weakly supervised approach, each WSI is represented as a bag (set) containing tens of thousands of image patches (instances) and a label (outcome) is provided for the entire bag [2]. Hence, the goal in this methodology is to learn a model that predicts a bag label by aggregating the predictions of the instances.

Many MIL approaches adopt a two-stage schema for tractable representation learning of WSIs, in which: 1) instance-level feature representations are extracted from the WSI image patches, and then 2) global aggregation schemes are applied to the bag of instances to obtain a WSI-level representation for subsequent supervised tasks [3]. Traditionally, ImageNet pretrained neural networks, such as ResNet [4], are utilized to extract the representations of the WSI patches in bags. As for the aggregation scheme, the attention-based MIL pooling, proposed in [5], has shown great results in different natural image and digital pathology datasets. It was further extended to the cancer survival analysis task with WSIs in [6, 3, 7] demonstrating its interpretability by locating important patterns and features which contribute to accurate survival predictions.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Recently, the authors of [8] reported sparsity in attention, i.e, models tend to localize most of their attention to some prominent patterns in the image. While being beneficial in natural images, it is not optimal for WSIs with their complex phenotypes associated with diverse biological concepts. The unsupervised pretraining techniques, such as discriminative approaches based on contrastive learning (CL) have recently shown great promise to extract salient features, such as SimCLR [9, 10], MoCo [11, 12, 13], DINO [14, 15]. These approaches let the construction of the pretrained image encoders for computational pathology [16, 17, 18, 19], which could be considered as foundation models, as long as they are capable to deliver improved performance on various downstream tasks and require minimal task-specific customization.

In this work, we aim at assessing the performance of the publicly available foundation models for digital pathology. We propose to benchmark the features from these pretrained image encoders using different self-supervised learning (SSL) algorithms SimCLR, MoCo and DINO against the model trained with the task-specific manual annotation. We use the ImageNet pretrained ResNet model as the baseline for comparison. We evaluate the performance of these models while fine-tuning with the two small ovarian cancer datasets, for which the extensive retraining is infeasible. We contribute to the assessment of the performance of these models in the under-explored survival prediction task. Finally, we propose to analyze the overall survival (OS) along with the progression free survival (PFS), the outcome that was not analyzed by the attention-based survival models with WSIs till now.

2 Related work and proposed benchmark

It has been shown that high densities of tumor-infiltrating lymphocytes (TILs) correlate with favorable clinical outcome in multiple cancer types [20]. The recent study [21] demonstrated that patients with a higher density of TILs had a significantly prolonged overall survival (OS) and progression-free survival (PFS) in multiple ovarian cancer cohorts.

In order to characterize TILs as a biomarker to guide future clinical research in precision oncology and immunotherapy, the study [22] proposed a deep learning pipeline for TIL detection and classification based on the diagnostic slides from The Cancer Genome Atlas (TCGA). It was limited to 13 different types of cancer, collected manually labeled individual patches and used the human-in-the-loop approach, i.e. in an iterative train-review-retrain process. This work was extended in [20, 23], the manual and computer-generated annotations were combined to produce the TIL maps across 23 different types of cancer. The authors trained VGG-16 [24], ResNet-34 [4] and Inception-V4 [25] networks to classify the patches as TIL-positive or TIL-negative. This improved framework resulted in better performance, which was attributed to the use of the state-of-the-art networks and larger and more diverse training datasets (TIL-Maps-23). The Inception-V4 obtained the best performance in the ovarian cancer dataset composed of 299 manually labeled test patches from TCGA ovarian cancer project (TCGA-OV).

The exhaustive task-specific manual annotation is often infeasible in digital pathology. To overcome this issue, self-supervised learning (SSL) could be a promising solution that relies only on unlabeled data to generate informative representations and can generalize well to various downstream tasks even with limited annotations.

The study [16] used 57 histopathology datasets, including 35 WSI datasets, to train a SimCLR model [9]. Most of the datasets used are stained with hematoxylin and eosin (H&E), come with 40x resolution and the majority are from TCGA and Clinical Proteomic Tumor Analysis Consortium (CPTAC). Their best trained model is based on the ResNet-18 architecture, trained for 1000 epochs, using 400 thousand images.

Another effort to learn universal feature representations more suitable for tasks in histopathology is the work [17]. This study proposes the strategy of semantically-relevant contrastive learning (SRCL), which compares relevance between instances to mine more positive pairs and introduces more visual diversity resulting in more informative semantic representations. This strategy is an extension of MoCo v3 methodology [13] but uses a convolutional neural network (CNN) and a multi-scale Swin Transformer architecture [26] as backbone. This hybrid architecture model (CTransPath) is pretrained on unlabeled histopathology images from TCGA and pathology AI platform (PAIP) [27], comprising 15 million unlabeled patches cropped from over 30 thousand WSIs.

More recently, the work [19] introduced UNI, a general-purpose self-supervised model for pathology. It was pretrained using over 100 thousand diagnostic H&E stained WSIs (more than 100 million images) across 20 major tissue types collected from Massachusetts General Hospital (MGH) and Brigham and Women's Hospital(BWH), as well as the Genotype-Tissue expression consortium [28]. In the pretraining stage, the authors used a self-supervised learning approach called DINO v2 [15].

The prognostic models from histology images have also demonstrated great promise [5] by integrating the MIL concepts, attention mechanisms and survival loss functions. The recent work of [3, 7] proposed a co-attention multimodal framework PORPOISE to jointly examine pathology WSIs and genomic features from 14 cancer types. Their work used the log likelihood function for a discrete survival model [29]. The overall solution is flexible by offering the possibility of training unimodal attention-bases MIL (AMIL) model that uses only WSIs.

The survival prediction with WSIs from TCGA-OV project has been approached by [30], the authors trained the SimCLR encoder to first extract features from 600 randomly selected non-background patches per WSI, second, they used the transformer encoder to integrate the extracted patch features and the corresponding patch positions to obtain the patient-level features with spatial information, third they trained the attention-based architecture for OS prediction using the negative Cox log partial likelihood [30]. The three presented blocks form the SeTranSurv model. We hypothesize that this work used not only the diagnostic TCGA-OV slides but also the tissue slides resulting in the dataset of 298 patients and 1481 WSIs. The code of this model is not publicly available, additionally the customized model architecture does not allow the fair comparison with other models, hence, we did not use this work in our benchmark but compare the SeTranSurv reported performance with our results.

Thus, we compared the pretrained TIL-Maps-23 [23], SimCLR by [16], CTransPath [17] and UNI [19], which use different CL algorithm models, in order to assess their ability to derive the universal histopathology features for the subsequent OS and PFS prediction using the PORPOISE model [7]. The proposed benchmark demonstrated as well the possibility to combine these recent deep learning models as building blocks to construct more sophisticated architectures for the datasets with limited annotation and size.

3 Experimental results

3.1 Datasets & evaluation metrics

To assess the performance of the digital pathology foundation models, we used two ovarian cancer datasets with high-resolution WSIs (20x). They are TCGA ovarian cancer dataset (TCGA-OV) and Ovarian Bevacizumab Response (OBR) [31, 32]. TCGA-OV is composed of the H&E stained diagnostic slides from TCGA and is available at `https://portal.gdc.cancer.gov`. The matching overall survival (OS) and progression free survival (PFS) and censorship statuses are published within TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) [33].

OBR is a dataset of H&E stained WSIs for classification of bevacizumab treatment effectiveness of ovarian cancer. The WSIs, as well as the matching clinical data (OS, PFS and censorship statuses) are available at `https://www.cancerimagingarchive.net/collection/ovarian-bevacizumab-response`. The number of cases, WSIs and censored cases is presented in the Table 1.

For each cancer dataset, we trained the PORPOISE model [7], AMIL network with WSI only input in a 5-fold cross-validation. The 5-fold split was done using R package MTLR [34] with the OS and PFS times and censorship stratification in order to have similar distributions of survival times and censorship in training and test sets.

We report the cross-validated concordance index (C-index) to measure the predictive performance of correctly ranking the predicted patient risk scores with respect to OS and PFS. C-index is a standard evaluation metric in survival analysis, ranging from 0 to 1, with a bigger C-index corresponding to a better model.

Relying solely on the C-index may not fully capture the model performance, thus, we include as well the cross-validated Integrated Brier Score (IBS). It is an extension of Brier Score (BS) over an

Table 1: The numbers of WSIs, patients and censoring % in each dataset.

Model/ Dataset	TCGA-OV	OBR
Number of cases	106	74
Number of WSIs	107	276
Number of censored OS	33	53
Number of censored PFS	35	28

interval of time, where BS is the mean squared error of the probability estimates. For this metric, smaller values signify better performance.

Finally, in order to plot the Kaplan-Meier curves, we aggregated the risk predictions from the test folds and plotted them against their survival times. We use the log rank test to measure if the difference of two survival distributions is statistically significant. For TCGA-OV datasets, the high-risk and low-risk groups are defined by the median of the risk predictions. As for OBR dataset, the predicted risks were first aggregated to the mean per patient, then the median value of the mean risks served to define the high-risk and low-risk groups.

3.2 Implementation details

For each WSI, automated segmentation of tissue was performed using the public tool for WSI analysis CLAM [35]. Subsequently, image patches of size 256x256 and 299x299 were extracted at the 20x level from all tissue regions identified. Following patch generation, the feature vectors were extracted using the following models:

- ResNet-50 model pretrained on ImageNet was used as an encoder to convert each 256x256 patch into a 1024-dimensional feature vector.
- TIL-Maps-23 Inception-V4 model [23] was used as an encoder to convert each 299x299 patch into a 1536-dimensional feature vector.
- SimCLR model pretrained on the histopathology images [16] was used as an encoder to convert 256x256 patches, first resized to 224x224, to 512-dimensional feature vector.
- CTransPath model [17] was used as an encoder to convert 256x256 patches, first resized to 224x224, to 768-dimensional feature vector.
- UNI model [19] was used as an encoder to convert 256x256 patches, first resized to 224x224, to 1024-dimensional feature vector.

We used the PORPOISE [7] hyperparameters suggested by the authors, except for: the $alpha\_surv$ (serves to weigh the uncensored patients) set at 0.5 and max_epochs (the maximum number of epochs to train) set at 40 in our experiments.

The tissue segmentation and patch extraction as well as the survival model training can be run on the GPU equipped desktop computer. We used GeForce RTX 2080 Super GPU with 8Gb of RAM, the tissue segmentation and patch extraction durations were less than 48 hours for each dataset and PORPOISE survival model training took less than 24 hours for each dataset/outcome.

3.3 Performance comparison

The obtained results are presented in Table 2 for C-index, Table 3 for IBS and Figures 1, 2 for Kaplan-Meier curves. In the Tables 2 and 3 the models' performance resulting in the significant difference of the survival distributions of the high versus low risk stratification is annotated with "(*)" and the best mean value is reported in bold.

In general, the features obtained by the self-supervised CL pretraining are more representative for histopathology survival prediction than ImageNet features. We observed as well that all the self-supervised pretrained models in this benchmark outperformed the model pretrained using CL technique with only the TCGA-OV dataset [30], where the reported average C-index of OS prediction was 0.692. We note that this comparison would be fairer if the same WSI dataset and survival loss function were used in both settings.

Table 2: Study results assessing C-index performance of different representation extraction techniques across 2 ovarian cancer datasets and 2 different outcomes.

Model/ Dataset, outcome	TCGA-OV, OS	TCGA-OV, PFS	OBR, OS	OBR, PFS
ImageNet ResNet-50	0.559±0.137	0.573±0.113	0.616±0.039	0.581±0.101
TIL-Maps-23 [23]	0.436±0.089	0.454±0.106	0.643±0.119 (*)	0.710±0.083 (*)
SimCLR [16]	**0.788±0.057 (*)**	0.706±0.091 (*)	0.541±0.068	0.699±0.095
CTransPath [17]	0.756±0.073 (*)	**0.758±0.063 (*)**	0.562±0.063	0.705±0.170
UNI [19]	0.785±0.094 (*)	0.687±0.124 (*)	**0.710±0.103 (*)**	**0.728±0.134**

Table 3: Study results assessing IBS performance of different representation extraction techniques across 2 ovarian cancer datasets and 2 different outcomes.

Model/ Dataset, outcome	TCGA-OV, OS	TCGA-OV, PFS	OBR, OS	OBR, PFS
ImageNet ResNet-50	**0.213±0.010**	0.219±0.029	0.333±0.118	0.227±0.062
TIL-Maps-23 [23]	0.222±0.012	0.238±0.039	**0.248±0.081 (*)**	**0.209±0.061 (*)**
SimCLR [16]	0.255±0.050 (*)	0.223±0.025 (*)	0.301±0.081	0.219±0.056
CTransPath [17]	0.217±0.011 (*)	**0.208±0.034 (*)**	0.354±0.102	0.214±0.042
UNI [19]	0.238±0.030 (*)	0.233±0.065 (*)	0.251±0.069 (*)	0.217±0.092

The studied foundation models in this work result in a better than the average OS C-index of 0.625 obtained by the PORPOISE WSI only AMIL model on TCGA-UCEC dataset. This Uterine Corpus Endometrial Carcinoma dataset from TCGA with 538 cases is another gynecological malignancy dataset but five time larger than the TCGA-OV dataset.

SimCLR model [16] and CTransPath model [17] obtained the best C-index on TCGA-OV dataset in OS and PFS prediction task respectively. The results also show that CTransPath [17] outperformed SimCLR model [16] in 3 out of 4 test settings in terms of C-index and IBS, we hypothesize that this is most probably due to the bigger size of the pretraining dataset than the choice of the CL algorithm or SSL augmentation techniques.

UNI model [19] achieves the best performance on the OBR dataset in PFS prediction task. Given that SimCLR model [16] and CTransPath [17] were trained on TCGA slides, their performance may be optimistically biased by the data leakage. Hence, we think that UNI model [19] is more robust and performs uniformly across the previously unseen datasets or in so-called out of distribution setting. As the UNI model used larger and more diverse datasets for pretraining, these results corroborate as well the hypothesis that using more unlabeled images in pretraining improves the downstream task performance.

Interestingly as well, the TIL-Maps-23 model trained with TCGA slides to detect the TIL-positive and TIL-negative patches results in the 2nd best C-index and the best IBS on the OBR dataset in PFS and OS prediction tasks. On the other hand, it did not perform well on the TCGA-OV dataset. Given that UNI and CTransPath models showed good results in TIL classification task on TCGA-OV as reported by [19], and that the study of [23] reported a relatively small mean TIL area in TCGA-OV dataset, we hypothesize that SSL pretrained models use other patterns for OS and PFS prediction on TCGA-OV dataset. Nevertheless, our results suggest that TILs could be a plausible hypothesis for developing the biomarkers of bevacizumab response prediction in ovarian cancer.

4 Conclusion

In this work we presented the performance assessment of the digital pathology foundation models. Our benchmark compared the generalization capabilities of the self-supervised pretrained image encoders SimCLR [16], CTransPath [17] and UNI [19] against the model TIL-Maps-23 [23] trained with manually annotated task-specific data. We compared the performance of the models on the under-explored task of survival prediction (OS and PFS) while using the relatively small ovarian cancer datasets.

Our results confirm the usefulness of the general purpose foundation models for digital pathology with the out of distribution domain data of diverse tissue types for pretraining. We advocate as well

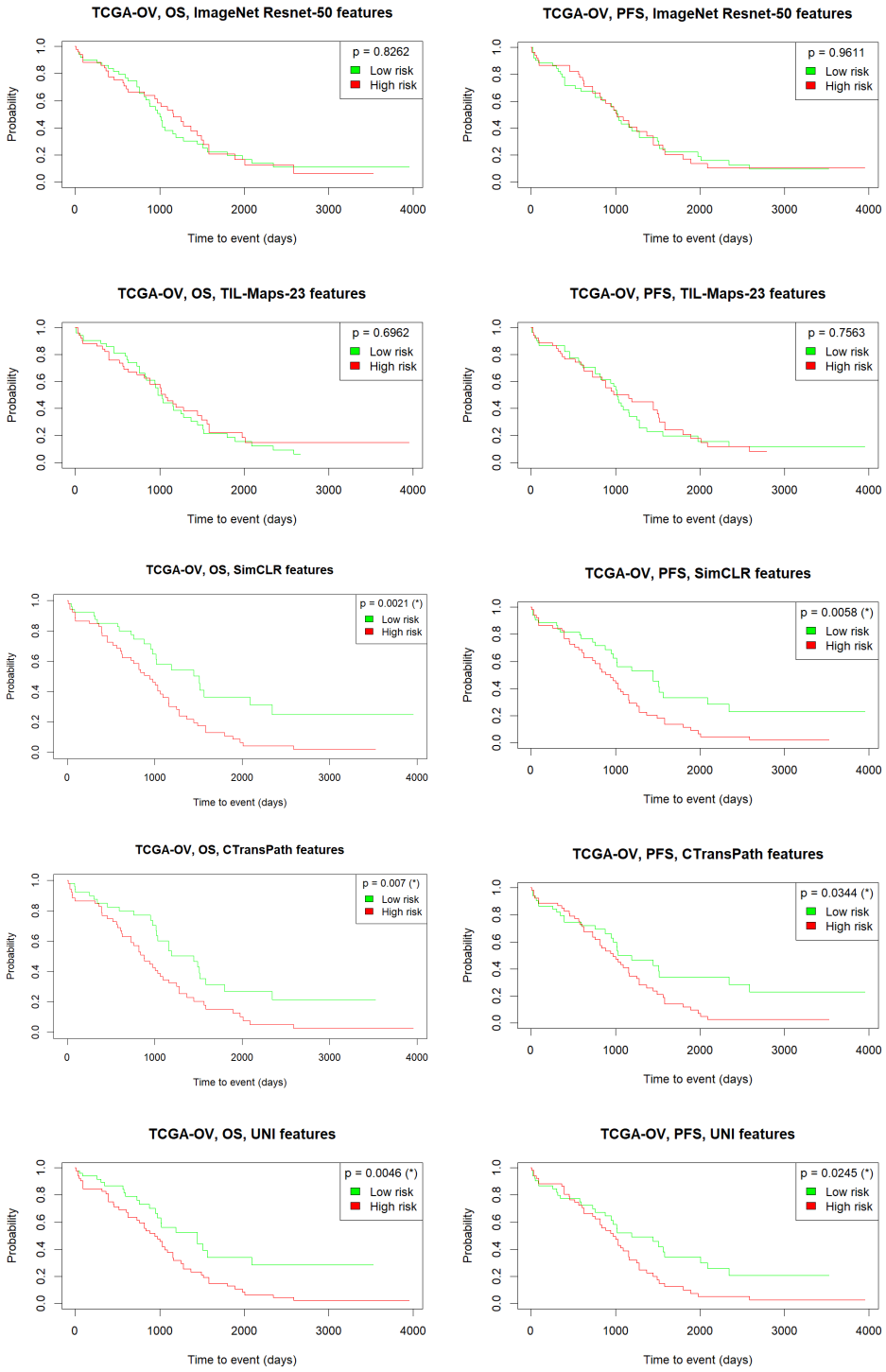

Figure 1: Kaplan-Meier curves of the high-risk and low-risk TCGA-OV patients. The groups were defined by the median of the risk predictions of PORPOISE/AMIL model trained with the benchmarked extracted features. Log rank test was used for statistical significance in survival distributions between high-risk and low-risk groups (*$p < 0.05$).

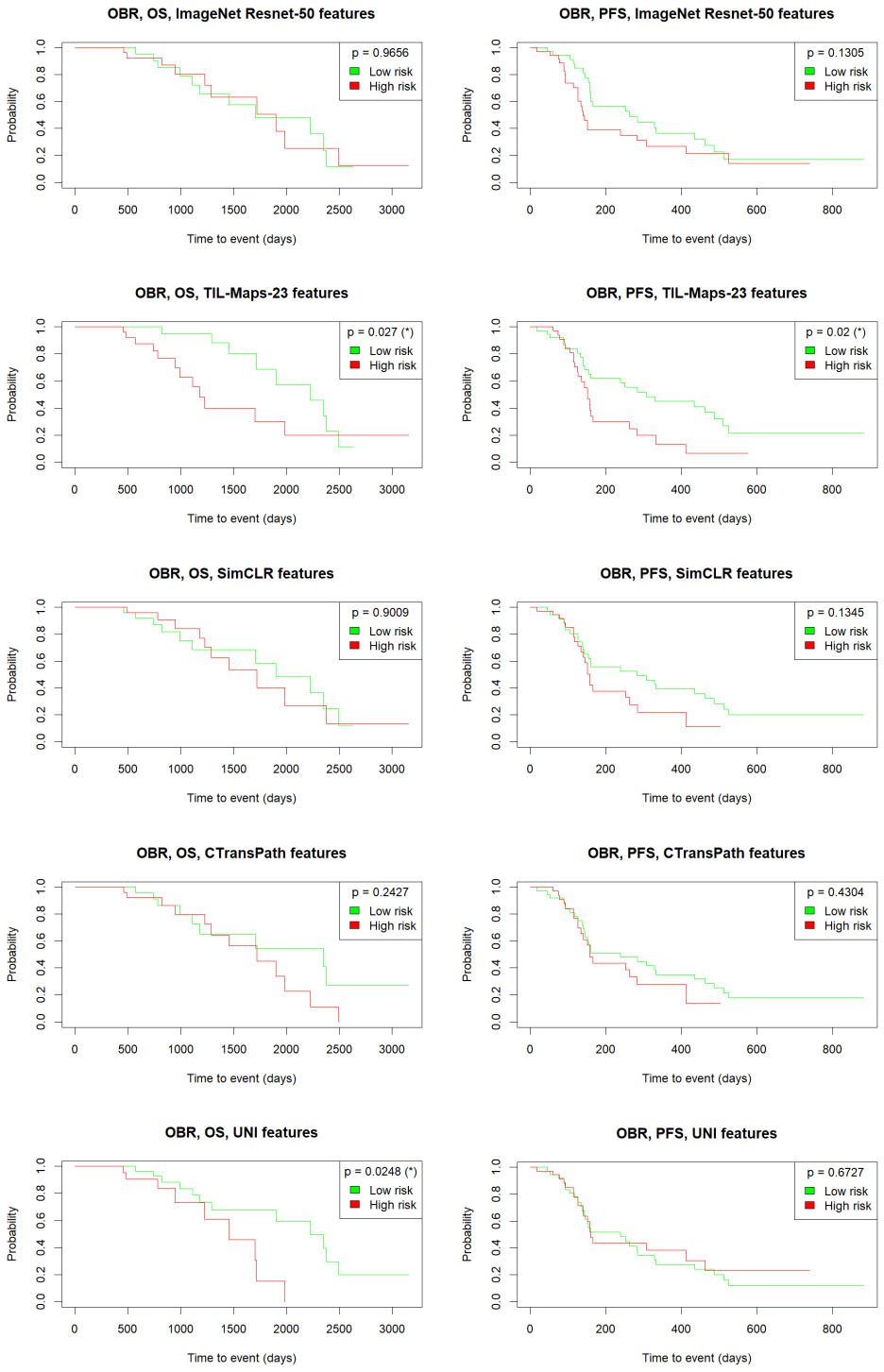

Figure 2: Kaplan-Meier curves of the high-risk and low-risk OBR patients. The risk predictions of PORPOISE/AMIL model trained with the benchmarked extracted features were used to first calculate the mean risk per patient, then the two groups were defined by the median of the mean risks. Log rank test was used for statistical significance in survival distributions between high-risk and low-risk groups (*$p < 0.05$).

that the presented models can be considered building blocks for the construction of more sophisticated architectures without the need of extensive training and large annotated datasets and could accelerate the development lifecycle of machine learning models for medical imaging.

As a potential clinical application, these combined models with the improved performance can help to guide patient stratification with the existing molecular subtyping. Another potential application of the further development of such models is to gain more insights into the mechanisms underlying the response or recurrence under particular treatment regimen. As a future work, we plan to thoroughly analyze the studied foundation models predictions in OS and PFS setting by exploiting the attention mechanism of the PORPOISE model to recognize the significant patterns that contribute to survival prediction.

Besides the following limitations could be considered as well. Most MIL methods neglect the spatial relationship among patches, the integration of the patch spatial information within the WSI could be a promising direction of future research. We did not search to optimize the PORPOISE model hyperparameters, our main goal was to compare the generalization capabilities of the studied foundation models. Finally, the OBR dataset containing multiple WSIs per patient, we have not used any strategy to aggregate the patient-level predictions. In addition, this dataset contains various histologic subtypes of ovarian cancer, while the TCGA-OV dataset is relatively homogeneous and is composed of high grade serous ovarian carcinoma subtype. This fact could explain the variance observed in the results on the OBR dataset.

Acknowledgments and Disclosure of Funding

Not Applicable

References

[1] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. June 3, 2021.

[2] Marc-André Carbonneau et al. "Multiple Instance Learning: A Survey of Problem Characteristics and Applications". In: *Pattern Recognition* 77 (May 2018), pp. 329–353. ISSN: 00313203. DOI: 10.1016/j.patcog.2017.10.009.

[3] Richard J. Chen et al. "Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, Oct. 2021, pp. 3995–4005. ISBN: 978-1-66542-812-5. DOI: 10.1109/ICCV48922.2021.00398.

[4] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 10, 2015.

[5] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. *Attention-based Deep Multiple Instance Learning*. June 28, 2018.

[6] Jiawen Yao et al. "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks". In: *Medical Image Analysis* 65 (Oct. 2020), p. 101789. ISSN: 13618415. DOI: 10.1016/j.media.2020.101789.

[7] Richard J. Chen et al. "Pan-cancer integrative histology-genomic analysis via multimodal deep learning". In: *Cancer Cell* 40.8 (Aug. 2022), 865–878.e6. ISSN: 15356108. DOI: 10.1016/j.ccell.2022.07.004.

[8] Saarthak Kapse et al. "Attention De-sparsification Matters: Inducing diversity in digital pathology representation learning". In: *Medical Image Analysis* 93 (Apr. 2024), p. 103070. ISSN: 13618415. DOI: 10.1016/j.media.2023.103070.

[9] Ting Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations*. June 30, 2020.

[10] Ting Chen et al. *Big Self-Supervised Models are Strong Semi-Supervised Learners*. Oct. 25, 2020.

[11] Kaiming He et al. *Momentum Contrast for Unsupervised Visual Representation Learning*. Mar. 23, 2020.

[12] Xinlei Chen et al. *Improved Baselines with Momentum Contrastive Learning*. Mar. 9, 2020.

[13] Xinlei Chen, Saining Xie, and Kaiming He. *An Empirical Study of Training Self-Supervised Vision Transformers*. Aug. 16, 2021.

[14] Mathilde Caron et al. *Emerging Properties in Self-Supervised Vision Transformers*. May 24, 2021.

[15] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. Feb. 2, 2024.

[16] Ozan Ciga, Tony Xu, and Anne Louise Martel. "Self supervised contrastive learning for digital histopathology". In: *Machine Learning with Applications* 7 (Mar. 2022), p. 100198. ISSN: 26668270. DOI: 10.1016/j.mlwa.2021.100198.

[17] Xiyue Wang et al. "Transformer-based unsupervised contrastive learning for histopathological image classification". In: *Medical Image Analysis* 81 (Oct. 2022), p. 102559. ISSN: 13618415. DOI: 10.1016/j.media.2022.102559.

[18] Shekoofeh Azizi et al. "Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging". In: *Nature Biomedical Engineering* 7.6 (June 8, 2023), pp. 756–779. ISSN: 2157-846X. DOI: 10.1038/s41551-023-01049-7.

[19] Richard J. Chen et al. "Towards a general-purpose foundation model for computational pathology". In: *Nature Medicine* 30.3 (Mar. 2024), pp. 850–862. ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-024-02857-3.

[20] Shahira Abousamra et al. *Learning from Thresholds: Fully Automated Classification of Tumor Infiltrating Lymphocytes for Multiple Cancer Types*. July 8, 2019.

[21] Kohei Hamada et al. "A Deep Learning–Based Assessment Pipeline for Intraepithelial and Stromal Tumor-Infiltrating Lymphocytes in High-Grade Serous Ovarian Carcinoma". In: *The American Journal of Pathology* (Mar. 2024), S000294402400110X. ISSN: 00029440. DOI: 10.1016/j.ajpath.2024.02.016.

[22] Joel Saltz et al. "Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images". In: *Cell Reports* 23.1 (Apr. 2018), 181–193.e7. ISSN: 22111247. DOI: 10.1016/j.celrep.2018.03.086.

[23] Shahira Abousamra et al. "Deep Learning-Based Mapping of Tumor Infiltrating Lymphocytes in Whole Slide Images of 23 Types of Cancer". In: *Frontiers in Oncology* 11 (Feb. 16, 2022), p. 806603. ISSN: 2234-943X. DOI: 10.3389/fonc.2021.806603.

[24] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Apr. 10, 2015.

[25] Christian Szegedy et al. *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. Aug. 23, 2016.

[26] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. Aug. 17, 2021.

[27] Yoo Jung Kim et al. "PAIP 2019: Liver cancer segmentation challenge". In: *Medical Image Analysis* 67 (Jan. 2021), p. 101854. ISSN: 13618415. DOI: 10.1016/j.media.2020.101854.

[28] The GTEx Consortium et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348.6235 (May 8, 2015), pp. 648–660. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1262110.

[29] Shekoufeh Gorgi Zadeh and Matthias Schmid. "Bias in Cross-Entropy-Based Training of Deep Survival Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.9 (Sept. 1, 2021), pp. 3126–3137. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2020.2979450.

[30] Ziwang Huang et al. "Integration of Patch Features Through Self-supervised Learning and Transformer for Survival Analysis on Whole Slide Images". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by Marleen De Bruijne et al. Vol. 12908. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 561–570. ISBN: 978-3-030-87236-6 978-3-030-87237-3. DOI: 10.1007/978-3-030-87237-3_54.

[31] Ching-Wei Wang et al. *A dataset of histopathological whole slide images for classification of Treatment effectiveness to ovarian cancer (Ovarian Bevacizumab Response)*. Version 2. 2021. DOI: 10.7937/TCIA.985G-EY35.

[32] Ching-Wei Wang et al. “Weakly supervised deep learning for prediction of treatment effectiveness on ovarian cancer from histopathology images”. In: *Computerized Medical Imaging and Graphics* 99 (July 2022), p. 102093. ISSN: 08956111. DOI: 10.1016/j.compmedimag.2022.102093.

[33] Jianfang Liu et al. “An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics”. In: *Cell* 173.2 (Apr. 2018), 400–416.e11. ISSN: 00928674. DOI: 10.1016/j.cell.2018.02.052.

[34] Chun-Nam Yu et al. “Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors”. In: *Advances in Neural Information Processing Systems 24* (2011), p. 10.

[35] Ming Y. Lu et al. “Data-efficient and weakly supervised computational pathology on whole-slide images”. In: *Nature Biomedical Engineering* 5.6 (Mar. 1, 2021), pp. 555–570. ISSN: 2157-846X. DOI: 10.1038/s41551-020-00682-w.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We did our best to be as accurate and concise as possible.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the Conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our work includes only empirical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the implementation details to reproduce the obtained results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The models and the datasets used in our work are all publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the data splits and our hyperparameters values, for the other hyperparameters we refer to the cited references where they can be found.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the 5-fold mean value $\pm$ standard deviation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the type of GPU used as well as the time of execution in the implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We did our best to follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential possible application of our way to integrate the studied models.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any particular data or model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited all the used datasets and models and respected the license and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.