# Comparing Natural Image Pre-training with Digital Pathology Foundation Models for Whole Slide Image-Based Survival Analysis

Kleanthis Marios Papadopoulos[a], Tania Stathaki[a], Panagiotis Barmpoutis[b], and Nazim Benzerdjeb[c, d]

[a]Department of Electrical and Electronic Engineering, Imperial College London, South Kensington Campus, Exhibition Road, London SW7 2AZ, United Kingdom
[b]Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, United Kingdom
[c]Department of Pathology, Institut de Pathologie Multisite, Groupement Hospitalier Sud, Hospices Civils de Lyon, Pierre-Bénite, France
[d]Université Claude Bernard Lyon 1, University of Lyon, Lyon, France

## ABSTRACT

Due to their high information density, Whole Slide Images (WSIs) are considered invaluable tools for survival analysis. Many existing Multiple Instance Learning (MIL) frameworks for this task rely on a ResNet50 backbone pre-trained on natural images. In this work, we leverage recently introduced histopathology foundation models, such as UNI and Hibou, to improve the predictive prowess of existing MIL frameworks. Additionally, we demonstrate that deploying an ensemble of digital pathology foundation models yields higher accuracy, although the gains diminish when applied to more complex MIL architectures.

**Keywords:** Medical Image Analysis, Multiple Instance Learning, Whole Slide Images, Digital Pathology, Survival Analysis, Model Ensembles, Foundation Models.

## 1. INTRODUCTION

Whole Slide Images (WSIs) are obtained from the digitization of histopathological glass slides and are the primary imaging modality in digital pathology.[1] Their ultra-high resolution allows multi-scale inspection, allowing users to zoom in for detailed examination of tissue features.[1] WSIs are typically prepared using Hematoxylin and Eosin (H&E) staining, which highlights nuclei in blue and cytoplasmic components in pink, enabling detailed morphological analysis.[2]

Digital pathology applies computational techniques to WSIs for tasks such as cancer classification, tumor subtyping, and survival analysis.[1] In clinical settings, survival analysis plays a critical role, as it estimates the time to significant clinical events, such as death or disease recurrence.

Traditional survival analysis techniques, including Cox proportional hazards and Kaplan-Meier models rely on handcrafted features.[3] In contrast, deep learning enables automated extraction of rich, unstructured image features. Early approaches, such as DeepConvSurv,[4] apply Convolutional Neural Networks (CNNs) to WSI patches but depended on manually annotated regions. Later methods, like the one presented in,[5] address this through random patch sampling and clustering, though they still require separate dimensionality reduction and lack end-to-end training.

Due to the enormous size of WSIs and the high resulting cost of obtaining pixel-level annotations, directly training neural networks on WSIs is not feasible.[1] Multiple Instance Learning (MIL) has emerged as a practical approach. Most MIL-based methods use ResNet50,[6] pre-trained on the ImageNet dataset,[7] as a feature extractor followed by a pooling mechanism.

Further author information: (Send correspondence to Kleanthis Marios Papadopoulos)
Kleanthis Marios Papadopoulos: E-mail: kp4718@ic.ac.uk

Progress in computer vision is recently driven by Self-Supervised Learning (SSL) on large-scale datasets. SSL operates by leveraging unlabeled data to learn robust visual representations through pretext tasks without relying on manual annotations.[8] SSL models are first pre-trained on these pretext tasks to capture general features and are then fine-tuned on downstream tasks, allowing them to effectively adapt to specific applications.[8] These pre-trained networks, often referred to as foundation models, have gained significant traction in digital pathology. Notably, models such as UNI[9] and Hibou[10] are developed through SSL-based pre-training on large collections of unlabeled WSIs, enabling them to capture pathology-specific visual representations.

In this work, we aim to provide a comprehensive evaluation of MIL-based survival analysis neural networks, focusing on their performance when leveraging ResNet50, UNI, and Hibou. We also propose a unique strategy of combining multiple feature extractors to boost the prognostic accuracy of end-to-end MIL architectures. While multiple works investigate the performance of foundation models on WSI classification[11,12] and cancer subtyping,[13] no other manuscript focuses specifically on survival analysis to the best of our knowledge.

## 2. RELATED WORK

### 2.1 Multiple Instance Learning

MIL is a weakly supervised learning paradigm which is widely adopted in workflows related to WSI analysis. In a typical binary classification problem, a sequence of images $X = \{X_1, X_2, ..., X_N\}$ are used to train a network that can predict a label $y_k \in \{0, 1\}$, for $k = \{1, 2, ..., N\}$. However, in the case of MIL, each image $X_k$ is a bag composed of a sequence of $M$ instances denoted by $X_k = \{x_{k1}, x_{k2}, ..., x_{kM}\}$ . A label is provided for the bag, but individual instances remain unlabeled. Mathematically, this can be expressed as:

$$y_k = \begin{cases} 1, & \text{if there exists any instance } x_{km} = 1 \text{ within the bag } X_k \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The standard MIL assumption is that individual instances are statistically independent from each other and that their ordering does not affect the MIL algorithm's operation. This problem formulation can be extended to multi-class classification and survival analysis. A key challenge in MIL is the weakly supervised nature of the problem, where predictions must be made at the bag level despite not having labels for individual instances within the bag. There exist two approaches for tackling MIL problems:

1. Instance-level approach: Each instance is assigned a score by a neural network and these scores are subsequently pooled.[14]
2. Embedding-level approach: Bag instances are converted into compact embeddings providing a matrix of embeddings for each bag. MIL pooling is subsequently used for the downstream task.[15]

The latter approach is often preferred as it avoids introducing further bias into the classification network.[15] Following feature extraction, a number of MIL aggregation mechanisms can be applied. Mean-pooling and max-pooling are the earliest aggregation techniques introduced.[16] Ilse et al. propose leveraging the attention mechanism, which forms the foundation of the Vision Transformer (ViT) architecture,[17] as a powerful instance aggregation method.[18] Subsequent architectures such as TransMIL[19] leverage the ViT architecture and others introduce approaches to alleviate the high computational load associated with training ViTs. Examples include S4MIL[20] and MambaMIL.[21]

### 2.2 Digital Pathology Foundation Models

SSL models are pre-trained on vast unlabeled datasets and subsequently fine-tuned on smaller, labeled, task-specific datasets.[8] This approach has seen great success in natural language processing and is now established in computer vision. Examples of pretext tasks include predicting an object's rotation angle or reconstructing missing image patches.[8]

The most popular backbone choices for pretext tasks include ResNet variants and ViTs.[8] Among SSL frameworks, Image BERT Pretraining with Online Tokenizer (iBOT) and Distillation with No Labels Version

2 (DINOv2) have become prominent for ViT pre-training because of their effectiveness in learning high-quality visual features. The iBOT framework employs a masked auto-encoder approach by training models to predict masked-out patches in images.[22] This helps the ViT backbone learn contextual relationships, a challenge typically faced by ViTs. DINOv2 utilizes a different learning strategy known as self-distillation.[23] In this approach, the model learns by comparing and aligning its representations across augmented views of the same image, enabling it to capture semantic features without reliance on labeled data.[23]

The success of the DINOv2 network in particular has motivated the release of several foundation models in the digital pathology domain. Recently released cases include the UNI model[9] by Mahmood Lab and the Hibou family of models[10] by HistAI. The UNI model was pre-trained on the Mass-100K dataset, which consists of over 100,000 Formalin-Fixed, Paraffin-Embedded (FFPE) Hematoxylin and Eosin (H&E) WSIs collected from Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH), as well as external slides from the GTEx consortium.[9] The pre-training process mimics the steps followed in the development of the DINOv2 framework.[9]

DINOv2 with Registers enhances the original DINOv2 by introducing additional tokens, otherwise known as registers, to mitigate artifacts observed in the original framework's attention maps.[24] The Hibou model family adopts this framework for model training. Similarly to UNI, HistAI's models were pre-trained on a private dataset of approximately 1 million images.[10] This collection of images includes non H&E and cytology slides.[10] The inclusion of these different types of WSIs aims to bolster dataset diversity and thus the model's ability to capture generalizable features.

## 3. METHODOLOGY

We tackle survival analysis as a weakly supervised slide-level task and we adopt the convention of fine-tuning a number of MIL frameworks following WSI feature extraction from backbone models. We utilize the Clustering-constrained Attention Multiple Instance Learning (CLAM) framework[25] to separate tissue content from background, divide WSIs into patches, and convert them into feature representations. CLAM employs median blurring, thresholding and morphological closing to obtain a segmented tissue mask. This segmented region is then divided into patches. The process is illustrated in Figure 1.

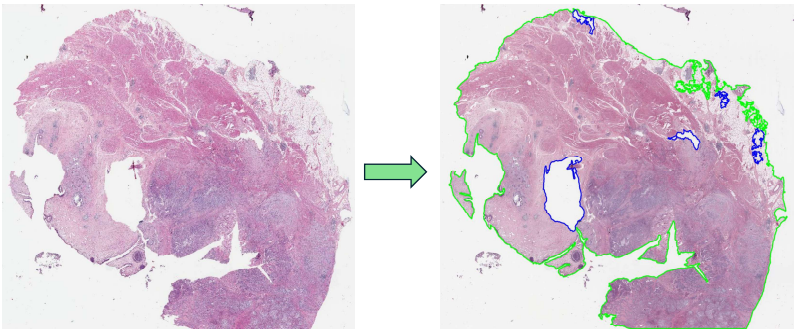


Figure 1: Illustration of the tissue segmentation process. The left panel shows the original whole slide image, while the right panel demonstrates the application of segmentation techniques, including thresholding, median blurring, and morphological operations, to identify and highlight the tissue regions (outlined in green). The boundaries of WSI regions with holes are drawn in blue.

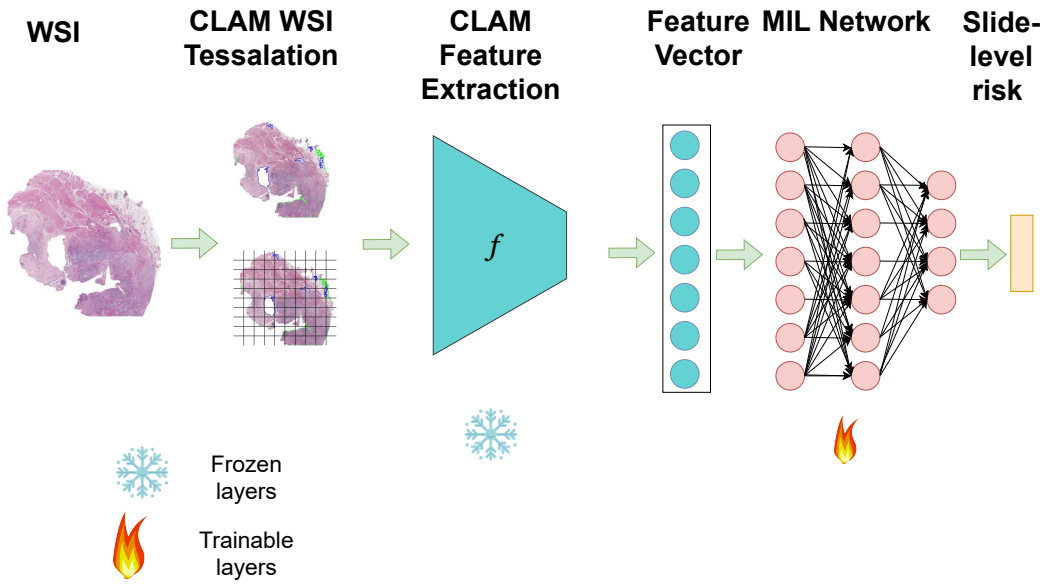



Figure 2: Overview of single backbone training strategy

We extract patches at 10× magnification and we perform feature extraction using ResNet50, UNI, and Hibou-Base. Subsequently, we train several downstream MIL frameworks, including MeanMIL, MaxMIL, Attention-based Multiple Instance Learning (ABMIL),[18] and TransMIL.[19] We summarize our approach in Figure 2.

Additionally, we explore foundation model ensembles based on tile embeddings through feature concatenation. A feature extractor $f_i$ transforms a given WSI $X_k$ into a feature matrix $\mathbf{F}_k \in \mathbb{R}^{n_k \times d}$, where $n_k$ denotes the number of segmented patches for the $k^{th}$ WSI and $d$ denotes the embedding space dimension. We therefore obtain a feature representation $\mathbf{G}$ mathematically defined as:

$$\mathbf{G} = \begin{bmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \vdots \\ \mathbf{F}_N \end{bmatrix}, \tag{2}$$

where $\mathbf{F}_i$ for $i = 1, \ldots, N$ denotes the $i^{th}$ feature matrix and $N$ is the total number of feature extractors under consideration. Following feature extraction, a trainable downstream MIL mechanism $m(.)$ converts the feature map into a slide-level risk prediction $\hat{h}$:

$$\hat{h} = m(\mathbf{G}) \tag{3}$$

We provide a graphical summary of the ensemble approach in Figure 3.

Figure 2 and Figure 3 depict MIL-based neural network architectures for slide-level risk prediction. In both, the CLAM network segments tissue, tessellates WSIs into patches, and encodes these into feature vectors. In Figure 2, a single feature extractor is used (i.e. $N = 1$), whereas in Figure 3, embeddings from multiple extractors are vertically concatenated into a unified feature vector, which is then passed to the downstream MIL network for patient-level risk prediction.

## 4. EXPERIMENTS & RESULTS

### 4.1 Datasets

We use WSIs from the publicly available TCGA database, and specifically the Bladder Urothelial Carcinoma (BLCA), the Lung Adenocarcinoma (LUAD), and the Breast Invasive Carcinoma (BRCA) cohorts. We select

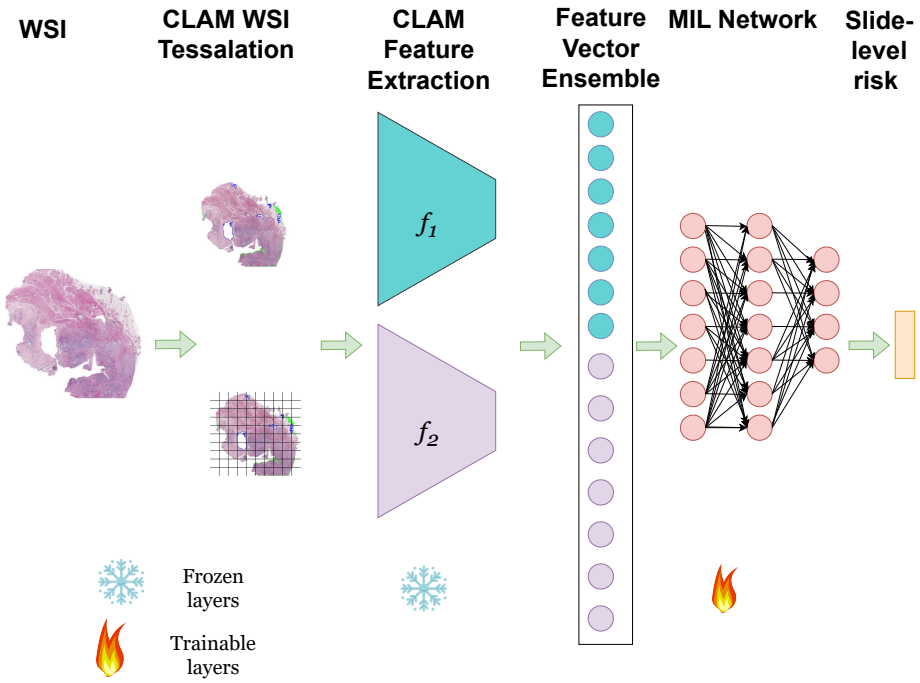



Figure 3: Overview of feature extractor ensemble training strategy

FFPE H&E diagnostic slides for each cohort and we follow the convention of using one WSI per patient by selecting slides having "DX1" in their TCGA identifier. According to the TCGA documentation, the "DX" label denotes a diagnostic slide. A large number of patients have multiple "DX" slides but not all. This selection process yields $N = 373$, $N = 443$ and $N = 1061$ WSIs for BLCA, LUAD, and BRCA respectively. Survival analysis datasets comprise both censored and uncensored observations. Uncensored observations correspond to patients with known survival times, while censored observations pertain to patients for whom survival time is not recorded, often because the patient was alive at the time of analysis. Figure 4 illustrates the distribution the distribution of censored and uncensored patients across each dataset, with the proportion of censored patients to the total number of patients being approximately 45% for the BLCA, 65% for the LUAD and 86% for the BRCA dataset respectively.

### 4.2 Implementation Details

Table 1 lists the dimension of the feature vector produced by each backbone along with the total number of parameters used during the pre-training phase. The same information for all model ensembles is provided in Table A.1 in Appendix A. We use the Negative Log Likelihood (NLL) as a bag-level loss function to train the aforementioned MIL networks and the concordance index to assess their prognostic accuracy. The number of trainable parameters for every MIL method is provided in Table B.1 in Appendix B. We divide the TCGA cohorts into train and validation sets and we report the results after applying $K$-fold cross validation ($K = 5$).

Our codebase is inspired by the MambaMIL framework,[21] and we build upon its publicly available implementation. The project is implemented in PyTorch, and our code repository can be found at `https://github.com/MarioPaps/WSI-Survival-Analysis-ImageNet-vs-Digital-Pathology-Backbones.git`. For training, we use the Adam optimizer and we set the batch size equal to 1 because the number of patches extracted for each WSI varies. We therefore perform backpropagation using gradient accumulation with the number of steps set to 32. Furthermore, we leverage early stopping and regularization to combat overfitting.

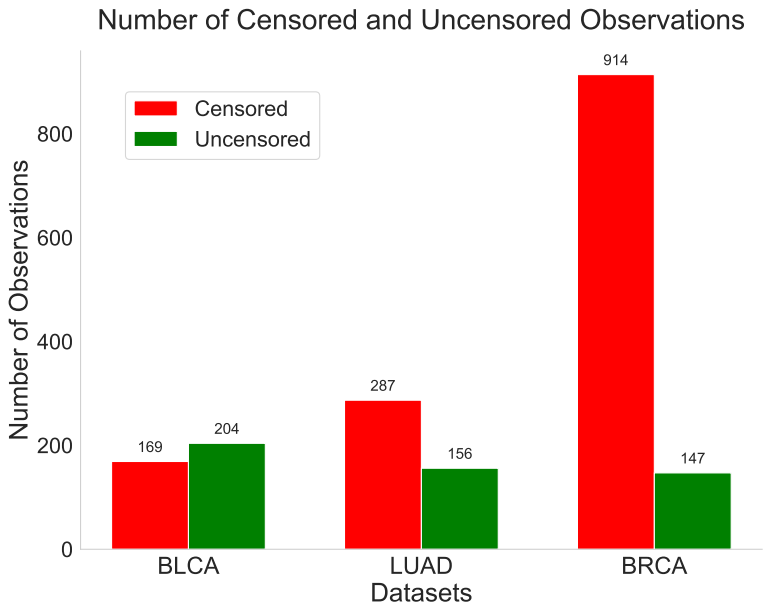



Figure 4: Distribution of censored and uncensored patients in each TCGA dataset

Table 1: Embedding dimensions of feature extractors used in the study

| **Feature Extractor** | **Embedding Dimension** | **Number of Parameters** |
|---|---|---|
| ResNet50 | 1024 | $25,557,032 \approx 25.6M$ |
| UNI | 1024 | $303,350,784 \approx 303M$ |
| Hibou-Base | 768 | $8,5741,056 \approx 8.6M$ |

### 4.3 Model Training & Evaluation

We approach survival analysis as an ordinal regression problem in this manuscript. The neural networks are trained to predict the correct temporal order of death occurrences by leveraging the ordinal structure of survival data. In real-world settings, patient survival data is frequently censored, which means the exact time of an event, such as death or disease recurrence, is not observed for every patient.[26] We therefore define the censoring indicator variable $\delta_i$ for patient $i$ as follows:

$$\delta_i = \begin{cases} 1, \text{ patient experienced death event by end of study (uncensored observation)} \\ 0, \text{ patient did not experience death event by end of study (censored observation)} \end{cases} \tag{4}$$

We denote by $t_i$ the OS time of patient $i$ and the associated risk by $r_i$. A pair of patients $(i, j)$ where the $j^{th}$ patient has a longer survival time than the $i^{th}$ $(t_j > t_i)$ patient is considered concordant if $r_i > r_j$ and assuming that patient $i$ has experienced death event by the end of the observation period. The concordance index is thus mathematically defined in Equation 5 as follows, in accordance with[27] :

$$\text{concordance index} = \frac{1}{N} \sum_{i \in \{1 \ldots N | \delta_i = 1\}} \sum_{t_j > t_i} I\left[r_i > r_j\right], \tag{5}$$

where $N$ denotes the number of usable pairs and $I(.)$ denotes the indicator function defined in Equation 6 as :

$$I(X) = \begin{cases} 1, \texttt{ condition } X \texttt{ is satisfied} \\ 0, \texttt{ condition } X \texttt{ is not satisfied} \end{cases} \tag{6}$$

The concordance index can be more simply viewed as the ratio of the number of concordant pairs divided by the total number of pairs that can be evaluated.

We follow the approach of dividing the continuous time axis into $Q$ discrete time bins defined by the intervals $[0, \beta_1), [\beta_1, \beta_2), ..., [\beta_{Q-1}, \beta_Q]$, where $0 = \beta_0 < \beta_1 < \beta_2 < \cdots < \beta_Q$. The neural network layer is thus equipped with $Q$ distinct neurons and predicts the discrete hazard. We utilize $Q = 4$ bins in this study. The discrete hazard is otherwise referred to as risk and is defined as the conditional probability of a death event being recorded at time $t$, assuming the patient has survived until time $t$. We note that time $t$ corresponds to a specific time bin in this formulation. The discrete hazard is mathematically expressed in Equation 7 as:

$$h(t|X) = P(T = t|T \geq t, X), \tag{7}$$

where $X$ denotes the covariates or input features provided to the model and $T$ denotes the time-to-event variable. The hazard function is constrained to the range $[0, 1]$ by applying a suitable activation function, typically a sigmoid function. The survival function in Equation 8 is obtained from the discrete hazard as follows:

$$S(t|X) = P(T > t|X) = \Pi_{s=0}^{s=t}\big(1 - h(s|X)\big) \tag{8}$$

The probability of survival at each time step $s$ is given by $\big(1-h(s|X)\big)$, where we use $h(s|X)$ to denote the hazard probability. Therefore, the cumulative probability up to time $t$ is obtained by multiplying the probabilities at each prior time step $s$, as indicated in Equation 8.

We adopt the NLL function as the loss function for our experiments. Because of censoring, the NLL objective function is composed of two components: a censored loss $\mathcal{L}_{\text{censored}}$ and an uncensored loss term $\mathcal{L}_{\text{uncensored}}$.

$$\mathcal{L}_{\text{uncensored}} = -\sum_{i=1}^{N_u} \delta_i \big[\log h(t_i|X_i) + \log S(t_i|X_i)\big] \tag{9}$$

$$\mathcal{L}_{\text{censored}} = -\sum_{i=1}^{N_c} \big((1 - \delta_i) \log S(t_i|X_i)\big), \tag{10}$$

where $i$ is the indexing variable, $\delta_i$ denotes the censoring status of the $i$-th patient, and $t_i$ represents the corresponding time bin. We denote the number of uncensored and censored observations by $N_u$ and $N_c$ respectively and it follows that the total number of patients is given by $N_u + N_c$. We therefore obtain the complete NLL loss function in Equation 11 as a sum of the two components :

$$\mathcal{L} = \mathcal{L}_{\text{uncensored}} + \mathcal{L}_{\text{censored}} \tag{11}$$

### 4.4 Results

We report the 5−fold validation results by listing the concordance index and the standard deviation for every TCGA cohort in Table 2. We also list the concordance index values achieved by two model ensembles, ResNet50 with UNI and UNI with Hibou-Base, in Table 3. We use bold font in Table 2 to highlight the entries which correspond to the highest concordance index for each MIL network when using a specific feature extractor. Similarly, bold font in Table 3 indicates the ensemble of feature extractors that achieves the highest prognostic accuracy.

Figure 5 presents the Kaplan-Meier survival curves obtained from each feature extractor after training with the ABMIL[18] network . A neural network is considered effective in distinguishing between low-risk and high-risk patients if the p-value from the log-rank test satisfies the condition $p < 0.05$. As shown in Figures 5c and 5d, the histopathological feature extractors satisfy this condition. However, the p-value for the ResNet50 network in Figure 5b slightly exceeds 0.05, which suggests that the ResNet50 network extracted less robust features for survival estimation.

Table 2: Concordance index values for different MIL methods across TCGA cohorts

| MIL Method | BLCA | LUAD | BRCA | Average |
|---|---|---|---|---|
| **ResNet50** | | | | |
| MeanMIL[16] | 0.536 ± 0.072 | 0.580 ± 0.021 | 0.567 ± 0.061 | 0.561 ± 0.063 |
| MaxMIL[16] | **0.543 ± 0.044** | 0.596 ± 0.044 | 0.570 ± 0.049 | 0.570 ± 0.056 |
| ABMIL[18] | 0.570 ± 0.069 | 0.594 ± 0.031 | 0.565 ± 0.040 | 0.576 ± 0.061 |
| TransMIL[19] | 0.528 ± 0.049 | 0.597 ± 0.063 | 0.591 ± 0.053 | 0.560 ± 0.051 |
| **UNI** | | | | |
| MeanMIL[16] | 0.570 ± 0.05 | 0.603 ± 0.030 | **0.597 ± 0.085** | **0.590 ± 0.073** |
| MaxMIL[16] | 0.530 ± 0.042 | **0.625 ± 0.076** | **0.610 ± 0.092** | **0.588 ± 0.089** |
| ABMIL[18] | **0.601 ± 0.046** | **0.605 ± 0.061** | 0.616 ± 0.066 | **0.607 ± 0.071** |
| TransMIL[19] | **0.565 ± 0.014** | 0.602 ± 0.047 | **0.628 ± 0.053** | **0.598 ± 0.051** |
| **Hibou-Base** | | | | |
| MeanMIL[16] | **0.572 ± 0.036** | **0.614 ± 0.073** | 0.575 ± 0.089 | 0.587 ± 0.085 |
| MaxMIL[16] | 0.518 ± 0.031 | 0.605 ± 0.073 | 0.550 ± 0.080 | 0.558 ± 0.080 |
| ABMIL[18] | 0.570 ± 0.043 | 0.565 ± 0.080 | **0.617 ± 0.085** | 0.584 ± 0.088 |
| TransMIL[19] | 0.542 ± 0.053 | **0.611 ± 0.034** | 0.557 ± 0.045 | 0.570 ± 0.055 |

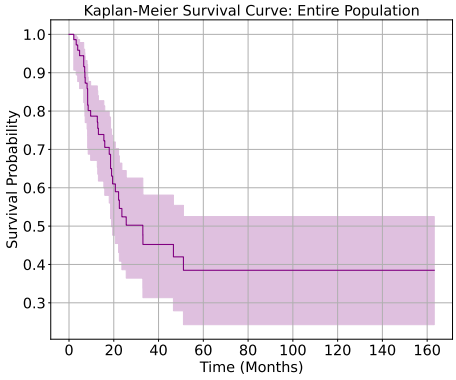
Table 3: Concordance index values for feature ensembles of MIL methods across TCGA cohorts

| MIL Method | BLCA | LUAD | BRCA | Average |
|---|---|---|---|---|
| **ResNet50+UNI** | | | | |
| MeanMIL[16] | 0.591 ± 0.028 | 0.586 ± 0.064 | 0.581 ± 0.080 | 0.586 ± 0.075 |
| MaxMIL[16] | 0.527 ± 0.050 | 0.603 ± 0.049 | 0.585 ± 0.089 | 0.572 ± 0.080 |
| ABMIL[18] | 0.596 ± 0.034 | 0.600 ± 0.083 | 0.606 ± 0.063 | 0.601 ± 0.078 |
| TransMIL[19] | **0.580 ± 0.033** | 0.589 ± 0.050 | 0.600 ± 0.078 | 0.590 ± 0.070 |
| **UNI+Hibou-Base** | | | | |
| MeanMIL[16] | **0.615 ± 0.020** | 0.601 ± 0.059 | 0.634 ± 0.068 | 0.617 ± 0.065 |
| MaxMIL[16] | **0.544 ± 0.050** | **0.615 ± 0.032** | **0.616 ± 0.069** | **0.592 ± 0.064** |
| ABMIL[18] | 0.587 ± 0.026 | **0.623 ± 0.063** | 0.610 ± 0.008 | 0.607 ± 0.049 |
| TransMIL[19] | 0.551 ± 0.025 | 0.578 ± 0.060 | 0.608 ± 0.058 | 0.579 ± 0.062 |
| **ResNet50+Hibou-Base** | | | | |
| MeanMIL[16] | 0.603 ± 0.020 | **0.605 ± 0.058** | **0.656 ± 0.080** | **0.621 ± 0.071** |
| MaxMIL[16] | 0.530 ± 0.034 | 0.592 ± 0.046 | 0.563 ± 0.064 | 0.562 ± 0.061 |
| ABMIL[18] | 0.579 ± 0.034 | 0.604 ± 0.089 | **0.655 ± 0.078** | 0.613 ± 0.087 |
| TransMIL[19] | 0.561 ± 0.021 | **0.599 ± 0.050** | **0.643 ± 0.057** | 0.601 ± 0.056 |
| **ResNet50 + UNI + Hibou-Base** | | | | |
| MeanMIL[16] | 0.603 ± 0.022 | 0.595 ± 0.079 | 0.650 ± 0.074 | 0.616 ± 0.078 |
| MaxMIL[16] | 0.525 ± 0.044 | 0.603 ± 0.049 | 0.570 ± 0.063 | 0.566 ± 0.064 |
| ABMIL[18] | **0.598 ± 0.022** | 0.604 ± 0.096 | **0.644 ± 0.058** | **0.615 ± 0.081** |
| TransMIL[19] | 0.563 ± 0.036 | 0.590 ± 0.040 | 0.609 ± 0.053 | **0.616 ± 0.078** |

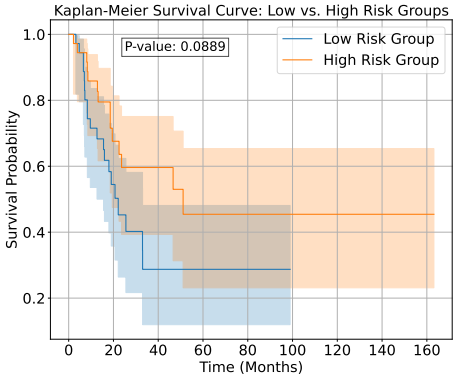
## 5. CONCLUSION

The results in Table 2 indicate that both histopathological feature extractors can consistently enhance the predictive prowess of the MIL networks used in this study. In most cases, the UNI backbone achieves a larger improvement over ResNet50 compared to Hibou-Base.

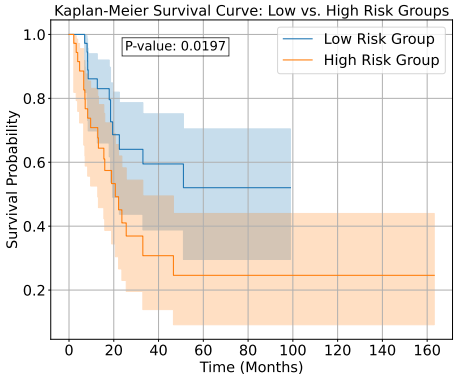
The reported concordance index values in Table 3 demonstrate that a significant performance improvement can be obtained by utilizing an ensemble of feature extractors. This improvement is particularly pronounced with the two simpler MIL architectures, which are MeanMIL[16] and MaxMIL.[16] Despite the ensemble of ResNet50 and UNI having a larger embedding dimension, the combinations of Hibou-Base with ResNet50 and UNI deliver more frequent improvements in outcome prediction accuracy. Furthermore, while combining two histopathological



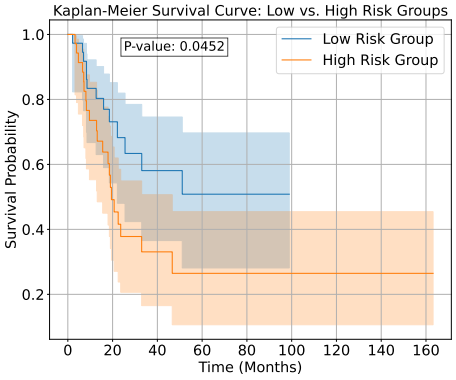

(a) Ground-truth Kaplan-Meier survival curve for the BLCA cohort based on patient data




(b) Kaplan-Meier survival curve produced by ABMIL[18] using ResNet50 as the feature extractor.




(c) Kaplan-Meier survival curve produced by ABMIL[18] using UNI as the feature extractor.




(d) Kaplan-Meier survival curve produced by ABMIL[18] using Hibou-Base as the feature extractor.

Figure 5: Kaplan-Meier survival curves generated by ABMIL for the BLCA dataset using different feature extractors. Sub-figures (a-d) show survival curves for the BLCA cohort (ground truth), ResNet50, UNI, and Hibou-Base feature extractors, respectively. Low-risk and high-risk patient groups are represented by blue and orange curves. Purple color is used for the Kaplan-Meier estimate (ground truth) in Figure 5a.

feature extractors typically results in a higher concordance index, mixing a feature extractor pre-trained on ImageNet with one pre-trained on WSIs outperforms the dual histopathological extractor approach in some cases. Lastly, the ensemble of all three feature extractors does not appear to yield a notable performance improvement in any MIL architecture.

This study can be expanded by incorporating datasets from non-TCGA-affiliated sites, enabling a more comprehensive evaluation of model generalizability across diverse patient cohorts and scanning protocols. Future work includes incorporating feature extractors pre-trained using different SSL approaches, besides self-distillation employed in DINOv2, to further assess the effect of different strategies on prognostic accuracy.

## APPENDIX A. FEATURE EXTRACTOR ENSEMBLE EMBEDDINGS

Table A.1 provides an overview of the embedding dimensions for different ensemble combinations of feature extractors used in this study.

Table A.1: Embedding Dimensions of the Ensembles

| Ensemble Combination | Embedding Dimension |
|---|---|
| ResNet50 + UNI | 2048 |
| Hibou-Base + ResNet50 | 1792 |
| Hibou-Base + UNI | 1792 |
| ResNet50 + UNI + Hibou-Base | 2816 |

## APPENDIX B. MIL MODEL PARAMETERS

Table B.1 presents the number of trainable parameters for MeanMIL,[16] MaxMIL,[16] ABMIL,[18] and TransMIL,[19] assuming an input embedding dimension of 1024.

Table B.1: Trainable parameter counts for MIL networks with an input embedding dimension of 1024.

| MIL Network | Number of Parameters |
|---|---|
| MeanMIL[16] | $526,852 \approx 527K$ |
| MaxMIL[16] | $526,852 \approx 527K$ |
| ABMIL[18] | $592,645 \approx 593K$ |
| TransMIL[19] | $2,673,172 \approx 2.7M$ |

## APPENDIX C. HYPERPARAMETER CHOICES

Table C.1 summarizes the hyperparameter choices that are consistently applied across all model setups.

Table C.1: Hyperparameter options common across all model configurations

| Datasets | BLCA | LUAD | BRCA |
|---|---|---|---|
| Number of WSIs | 373 | 443 | 1061 |
| Batch size | 1 | 1 | 1 |
| Gradient accumulation steps | 32 | 32 | 32 |
| Optimizer | Adam | Adam | Adam |
| Bag weight | 0.7 | 0.7 | 0.7 |
| Learning rate | $2e-4$ | $1e-4$ | $5e-5$ |
| Epochs | 200 | 200 | 200 |
| Early stopping | Enabled | Enabled | Enabled |
| Earliest stop epoch | 40 | 40 | 40 |
| L1 regularization | $1e-4$ | $1e-4$ | $1e-4$ |
| Patience epochs | 10 | 5 | 10 |
| Dropout rate | 0.25 | 0.25 | 0.25 |
| Weight decay | $1e-3$ | $5e-4$ | $5e-4$ |

## ACKNOWLEDGMENTS

We acknowledge computational resources and support provided by the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232).

## REFERENCES


[1] Van der Laak, J., Litjens, G., and Ciompi, F., "Deep learning in histopathology: the path to the clinic," *Nature Medicine* **27**(5), 775–784 (2021).

[2] Fuchs, T. J. and Buhmann, J. M., "Computational pathology: Challenges and promises for tissue analysis," *Computerized Medical Imaging and Graphics* **35**(7), 515–530 (2011). Whole Slide Image Process.

[3] Qaiser, T., Lee, C.-Y., Vandenberghe, M., Yeh, J., Gavrielides, M., Hipp, J., Scott, M., and Reischl, J., "Usability of deep learning and he images predict disease outcome-emerging tool to optimize clinical trials," *npj Precision Oncology* **6**, 37 (06 2022).

[4] Zhu, X., Yao, J., and Huang, J., "Deep convolutional neural network for survival analysis with pathological images," in [*2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*], 544–547 (December 2016).

[5] Zhu, X., Yao, J., Zhu, F., and Huang, J., "Wsisa: Making survival prediction from whole slide histopathological images," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 6855–6863 (July 2017).

[6] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[7] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE Conference on Computer Vision and Pattern Recognition*], 248–255 (2009).

[8] Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D., "Using self-supervised learning can improve model robustness and uncertainty," *Advances in neural information processing systems* **32** (2019).

[9] Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., and Mahmood, F., ""towards a general-purpose foundation model for computational pathology"," *Nature Medicine* **30**, 850–862 (mar 2024).

[10] Nechaev, D., Pchelnikov, A., and Ivanova, E., "Hibou: A family of foundational vision transformers for pathology," (2024).

[11] Lenz, T., Neidlinger, P., Ligero, M., Wölflein, G., van Treeck, M., and Kather, J. N., "Unsupervised foundation model-agnostic slide-level representation learning," (2024).

[12] Wölflein, G., Ferber, D., Meneghetti, A. R., Nahhas, O. S. M. E., Truhn, D., Carrero, Z. I., Harrison, D. J., Arandjelović, O., and Kather, J. N., "Benchmarking pathology feature extractors for whole slide image classification," (2024).

[13] Meseguer, P., del Amor, R., Colomer, A., and Naranjo, V., "Foundation models for slide-level cancer subtyping in digital pathology," (2024).

[14] Liu, G., Wu, J., and Zhou, Z.-H., "Key instance detection in multi-instance learning," in [*Asian conference on machine learning*], 253–268, PMLR (2012).

[15] Wang, X., Yan, Y., Tang, P., Bai, X., and Liu, W., "Revisiting multiple instance neural networks," *Pattern Recognition* **74**, 15–24 (2018).

[16] Oner, M. U., Kye-Jet, J. M. S., Lee, H. K., and Sung, W.-K., "Studying the effect of mil pooling filters on mil tasks," (2020).

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale," (2021).

[18] Ilse, M., Tomczak, J., and Welling, M., "Attention-based deep multiple instance learning," in [*Proceedings of the 35th International Conference on Machine Learning*], Dy, J. and Krause, A., eds., *Proceedings of Machine Learning Research* **80**, 2127–2136, PMLR (10–15 Jul 2018).

[19] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., and zhang, y., "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," in [*Advances in Neural Information Processing Systems*], Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., eds., **34**, 2136–2147, Curran Associates, Inc. (2021).

[20] Fillioux, L., Boyd, J., Vakalopoulou, M., Cournède, P.-h., and Christodoulidis, S., "Structured state space models for multiple instance learning in digital pathology," in [*Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*], Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., and Taylor, R., eds., 594–604, Springer Nature Switzerland, Cham (2023).

[21] "Yang, S., Wang, Y., and Chen, H., ""mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology"," in [*"Medical Image Computing and Computer Assisted Intervention – MICCAI 2024"*], 296–306, Springer Nature Switzerland, Cham (2024).

[22] Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T., "ibot: Image bert pre-training with online tokenizer," (2022).

[23] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P., "Dinov2: Learning robust visual features without supervision," (2024).

[24] Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P., "Vision transformers need registers," in [*The Twelfth International Conference on Learning Representations*], (2024).

[25] Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F., ""data-efficient and weakly supervised computational pathology on whole-slide images"," *Nature Biomedical Engineering* **5**(6), 555–570 (2021).

[26] Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., Girard, N., Elemento, O., Nicholson, A. G., Blay, J.-Y., Galateau-Sallé, F., Wainrib, G., and Clozel, T., "Deep learning-based classification of mesothelioma improves prediction of patient outcome," *Nature Medicine* **25**(10), 1519–1525 (2019).

[27] Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., and Raykar, V. C., "On ranking in survival analysis: Bounds on the concordance index," in [*Advances in Neural Information Processing Systems*], Platt, J., Koller, D., Singer, Y., and Roweis, S., eds., **20**, Curran Associates, Inc. (2007).